

The next phase in heritage language studies: Methodological considerations and advancements

Edited by

Fatih Bayram, Maki Kubota
and Sergio Miguel Pereira Soares

Published in

Frontiers in Psychology
Frontiers in Communication



FRONTIERS EBOOK COPYRIGHT STATEMENT

The copyright in the text of individual articles in this ebook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this ebook is the property of Frontiers.

Each article within this ebook, and the ebook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this ebook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or ebook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-8325-4693-2
DOI 10.3389/978-2-8325-4693-2

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

The next phase in heritage language studies: Methodological considerations and advancements

Topic editors

Fatih Bayram — UiT The Arctic University of Norway, Norway

Maki Kubota — UiT The Arctic University of Norway, Norway

Sergio Miguel Pereira Soares — Max Planck Institute for Psycholinguistics, Netherlands

Citation

Bayram, F., Kubota, M., Pereira Soares, S. M., eds. (2024). *The next phase in heritage language studies: Methodological considerations and advancements*.

Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-4693-2

Table of contents

05	Editorial: The next phase in heritage language studies: methodological considerations and advancements Fatih Bayram, Maki Kubota and Sergio Miguel Pereira Soares
09	Predictors of language proficiency and cultural identification in heritage bilinguals Sayuri Hayakawa, Ashley Chung-Fat-Yim and Viorica Marian
33	Eyetracking evidence for heritage speakers' access to abstract syntactic agreement features in real-time processing Zuzanna Fuchs
50	Assessing language background and experiences among heritage bilinguals Alessandra Macbeth, Natsuki Atagi, Jessica L. Montag, Michelle R. Bruni and Christine Chiarello
67	Effects of markedness in gender processing in Italian as a heritage language: A speed accuracy tradeoff Grazia Di Pisa, Maki Kubota, Jason Rothman and Theodoros Marinis
83	Processing pro-drop features in heritage Turkish Serkan Uygun
98	Multiple wh-interrogatives in child heritage Romanian: On-line comprehension and production Anamaria Bentea and Theodoros Marinis
116	(In)frequently asked questions: On types of frequency and their role(s) in heritage language variability Silvia Perez-Cortes and David Giancaspro
131	Moving away from deficiency models: Gradiency in bilingual speech categorization Ethan Kutlu, Samantha Chiu and Bob McMurray
147	Requests and apologies in two languages among bilingual speakers: A comparison of heritage English speakers and English- and Hebrew-dominant bilinguals Sagit Bar On and Natalia Meir
166	Factors predicting secondary school language course enrollment and performance among U.S. heritage speakers of Spanish My V. H. Nguyen, Ellen J. Serafini, Jennifer Leeman and Adam Winsler
182	Production, acceptability, and online comprehension of Spanish differential object marking by heritage speakers and L2 learners Begoña Arechabaleta Regulez and Silvina Montrul

- 201 **Adjective position in the code-switched speech of Spanish and Papiamentu heritage speakers in the Netherlands: Individual differences and methodological considerations**
Brechtje van Osch, M. Carmen Parafita Couto, Ivo Boers and Bo Sterken
- 217 **Using self-paced reading in research with heritage speakers: a role for reading skill in the online processing of Spanish verb argument specifications**
Jill Jegerski and Gregory D. Keating
- 228 **Documenting heritage language experience using questionnaires**
Aleksandra Tomić, Yulia Rodina, Fatih Bayram and Cécile De Cat
- 246 **Practice beats age: co-activation shapes heritage speakers' lexical access more than age of onset**
Nuria Sagarra and Joseph V. Casillas
- 264 **Morphological transparency and markedness matter in heritage speaker gender processing: an EEG study**
Alicia Luque, Eleonora Rossi, Maki Kubota, Megan Nakamura, César Rosales, Cristina López-Rojas, Yulia Rodina and Jason Rothman
- 280 **Parental emotionality and power relations in heritage language maintenance: experiences of Chinese and African immigrant families in Australia**
Yining Wang, Vera Williams Tetteh and Sithembinkosi Dube
- 294 **Use of the first-acquired language modulates pupil size in the processing of island constraint violations**
Gita Martohardjono, Michael A. Johns, Pamela Franciotti, Daniela Castillo, Ilaria Porru and Cass Lowry
- 315 **Turkish-German heritage speakers' predictive use of case: webcam-based vs. in-lab eye-tracking**
Onur Özsoy, Büsra Çiçek, Zeynep Özal, Natalia Gagarina and Irina A. Sekerina
- 332 **Light-weights placed right: post-field constituents in heritage German**
Wintai Tsehay
- 345 **Insights into the time course of evidentiality processing in Turkish heritage speakers using a self-paced reading task**
Suzan D. Tokaç-Scheffer, Seçkin Arslan and Lyndsey Nickels
- 364 **The role of external factors on the reactivation of the heritage language of Turkish-German returnees**
Elena Antonova-Unlu and Fatih Bayram



OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Fatih Bayram
✉ fatih.bayram@uit.no

RECEIVED 27 February 2024
ACCEPTED 04 March 2024
PUBLISHED 19 March 2024

CITATION
Bayram F, Kubota M and Pereira Soares SM
(2024) Editorial: The next phase in heritage
language studies: methodological
considerations and advancements.
Front. Psychol. 15:1392474.
doi: 10.3389/fpsyg.2024.1392474

COPYRIGHT
© 2024 Bayram, Kubota and Pereira Soares.
This is an open-access article distributed
under the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited,
in accordance with accepted academic
practice. No use, distribution or reproduction
is permitted which does not comply with
these terms.

Editorial: The next phase in heritage language studies: methodological considerations and advancements

Fatih Bayram^{1*}, Maki Kubota¹ and Sergio Miguel Pereira Soares²

¹UiT The Arctic University of Norway, Tromsø, Norway, ²Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

KEYWORDS

heritage language speakers, bilingualism, online methods, language processing, language maintenance

Editorial on the Research Topic

The next phase in heritage language studies: methodological considerations and advancements

Over the past three decades, research on heritage language (HL) bilingualism has undergone significant advancement revealing the intricate dynamics of linguistic competencies among heritage speakers (HSs). As a unique subgroup within the bilingual community, these individuals typically acquire their native language(s) in environments where it is not the dominant language, often due to migration, where HLs may be spoken at home but not formally taught or reinforced in dominant societal/educational settings (e.g., Rothman, 2009; Montrul, 2016; Polinsky, 2018). Despite being native speakers of their home language(s), HSs exhibit vast outcomes variation of linguistic competence/performance compared to other bilinguals and monolingual peers (see Kupisch and Rothman, 2018). This variability has prompted researchers to explore methodologies that capture the nuances of HS linguistic knowledge and processing. This line of investigations has delved into how HSs maintain, adapt or even lose competence in their native language over time, and also explored the sociolinguistic and experiential factors that shape such observations. Traditionally, these studies were rooted in adjacent fields such as L1 acquisition and adult L2 acquisition, predominantly employing behavioral methodologies to understand HS performance. While informative, these approaches often overlooked the methodological complexities inherent in studying HS linguistic realities, which can dynamically shift across the lifespan (Bayram et al., 2021).

Recent advancements, however, have marked a paradigm shift in HL bilingualism research, with a focus on methodological innovations aimed at more accurately capturing the linguistic competencies of HSs (Bayram et al., 2021). This movement unfolds on three main fronts. Firstly, there is a departure from traditional HSs vs. non-HSs comparisons, as researchers now explore comparisons among different HS groups, seeking to comprehend HL grammars in their own right. This shift allows for a more nuanced understanding of the variations underlying HL competence. Secondly, studies have delved into the multidimensional relationship between HSs' sociolinguistic networks and their linguistic competence, acknowledging the role of individual differences within HS groups. This approach recognizes that linguistic competence is not solely shaped by exposure to the

HL but is also affected by the sociolinguistic environments in which HSs are placed in. Finally, the adoption of novel (for the field) online/processing methodologies, such as eye-tracking and electroencephalography/event-related potentials (EEG/ERPs), represents another frontier. These innovative techniques provide insights into automatic language processing, offering a more granular understanding of the underlying cognitive mechanisms at play in HL competence. By employing these advanced methods, researchers aim to circumvent confounding variables that can be more challenging to tease apart in more traditional methodologies and capture a more accurate representation of the interplay between linguistic competence and processing in HSs (e.g., [Pereira Soares, 2022](#)).

By leveraging on all these innovations, the studies within this Research Topic aimed to chart the multifaceted landscape of HL bilingualism, the underlying mental systems of HL grammatical outcomes, processing, and maintenance within the context of diverse linguistic and socio-cultural environments. Drawing from a range of innovative methodologies and approaches, including offline experimental studies, psycho-/neurolinguistic studies employing online methods, and corpus analyses, the articles in this Research Topic span over a rich array of inquiry. By exploring the influence of linguistic exposure, proficiency levels, language attitudes, and socio-cultural contexts on HL competence/performance, these articles provide valuable insights into mechanisms underlying heritage language development. More importantly, they collectively contribute to the evolving landscape of HL bilingualism research, thus bridging the current state-of-the-art with future directions in HL studies.

In the three following sub-chapters, we present a comprehensive exploration of HL bilingualism, highlighting the methodological intricacies and theoretical implications that shape the current understanding of this complex linguistic phenomenon.

The first group of studies focus on assessing individual experiences and HL competence/performance via employment of detailed questionnaires and/or other background measures. [Tomić et al.'s](#) validation of the Heritage Language Experience (HeLex) questionnaire provides a comprehensive assessment tool for documenting heritage bilingualism, highlighting the importance of methodological choices in assessing language background and proficiency levels. They proposed a comprehensive online questionnaire for documenting heritage bilingualism, validated against an extended version of an already existing questionnaire, revealing important distributional patterns in their data. In a similar vein, [Perez-Cortes and Giancaspro's](#) exploration of frequency effects in HL acquisition underscores the complexity of linguistic development among bilingual individuals, emphasizing the need for comprehensive (subjective) assessments of language exposure and proficiency. Similarly, [Macbeth et al.'s](#) study on bilingual language experiences underscores the importance of employing diverse assessment methods to capture the intricacies of real-world language use among HSs. They examined bilinguals' language experiences using self-report questionnaires and audio recordings, revealing significant predictors of real-world language use via self-reported language use and age of English acquisition. [van Osch et al.](#) examined adjective-noun word order in code-switching among Spanish and Papiamentu HSs in the Netherlands.

They found that both linguistic (e.g., matrix, type of insertion) and non-linguistic (e.g., age, exposure, use) aspects influence how HSs navigate code-switching, and that children may require more time or exposure for adult-like norms. Focusing on the linearization of constituents at the right sentence periphery in German, [Tsehaye's](#) study analyzed spoken and written productions from English-German HSs and monolingually-raised speakers of German in different registers. Their findings offer insights into the impact of language contact and exposure on syntactic variation, contributing to our understanding of language change and adaptation. Assessing a different syntactic domain, [Arechabaleta Regulez and Montrul's](#) analysis of differential object marking (DOM) among Spanish HSs and L2 learners also found that type of task and type of sentence each have an effect on speakers' use of DOM, together with experiential factors such as language experience and practices. Finally, [Kutlu et al.'s](#) research on speech perception among bilingual communities introduces a novel approach to examining categorical perception, challenging existing theories and highlighting the need for a more precise understanding of speech categorization. They reexamined the theory of categorical perception in speech, introducing the Visual Analog Scaling task to enable a more precise examination of speech categorization in diverse bilingual communities, specifically HSs who often show gradient speech perception across different contexts.

The focus of the next cohort of studies is understanding the impact of socio-economic, cultural, and educational factors on the multifaceted and diverse nature of HLs. Firstly, [Nguyen et al.'s](#) study draws attention to the socio-economic factors influencing language course enrollment and performance among HSs, shedding light on disparities in educational access and outcomes within bilingual communities, e.g., by highlighting the impact of disability status, poverty, and prior academic performance. By examining and emphasizing language proficiency and cultural identity among heritage speakers, [Hayakawa et al.'s](#) work uncovered the predictors of language proficiency, vocabulary, and cultural identification in different groups of HSs, highlighting the importance of accounting for individual language history (such as overall HL exposure, HL experience in informal and formal contexts). The next two studies draw attention to diverse aspects of immigration influence on HLs. [Wang et al.'s](#) cross-sectional exploration of emotional experiences within Chinese and African immigrant families underscores the significance of language emotions in shaping family language policies and language ideologies, providing valuable insights into the socio-emotional dimensions of HL maintenance. [Antonova-Unlu and Bayram's](#) investigation into HL performance among Turkish-German returnees (into Turkey) sheds light on the challenges and opportunities faced by individuals reintegrating into their HL community, highlighting the role of external factors (the length of residence, the age at return to the homeland, and the frequency of HL use in the migration context) in language proficiency, maintenance and (re-)activation of their HL. Finally, [Bar On and Meir's](#) investigation into speech act pragmatics among HSs sheds light on the cross-cultural and cross-linguistic differences in request and apology realizations. They compared English (HL)-Hebrew adult speakers in Israel with Hebrew-dominant and English-dominant speakers. They found

distinct hybrid strategies in requests and apologies among HSs, showing cross-cultural and cross-linguistic differences in their pragmatic competencies.

The third and last set of studies employed a diverse array of psycho-/neurolinguistic methods to understand how HL processing unfolds in the minds of HSs. Uygun examined the real-time sentence processing of plural-marked and unmarked verbs in sentences with overt and null subjects using self-paced reading task (SPRT) among Turkish HSs. Their results show both qualitative and quantitative differences in processing strategies between Turkish HSs and Turkish non-HSs, suggesting that Turkish HSs do indeed have the syntactic structure but may need more time to integrate this information during real-time processing. Tokaç-Scheffer et al. also used a SPRT among Turkish HSs to examine their processing of evidentiality, i.e., the linguistic marking of information source. Their findings reveal quantitative differences between HSs and non-HSs in the sense that HSs were generally slower and less accurate than non-HSs in both reading times and acceptability judgements, but both groups showed similar patterns regarding reading times on evidential-marked verb forms that matched or mismatched to the information source. The studies by Uygun and Tokaç-Scheffer et al. collectively demonstrate that when tested in both online and offline modes, HSs consistently show quantitative differences in an online paradigm, suggesting that HSs have difficulties in dealing with cognitive load that comes with real-time processing of linguistic structures. Indeed, Di Pisa et al.'s investigation of grammatical gender variability in Italian HSs show converging evidence showing that, only in an SPR paradigm, HSs show greater sensitivity to markedness (agreement violations realized on feminine adjectives) compared to non-HSs, while both groups make use of markedness information in offline grammaticality judgement task. Jegerski and Keating's study on Spanish verb argument specifications adds to the findings of other studies in this Research Topic employing SPRT, by demonstrating that lower self-ratings for reading skill in Spanish and slower average reading speed correlated to a larger spillover effect of transitivity among HSs. Their study underlines the role that general reading skills play when testing morphosyntactic processing among HSs using an online processing paradigm such as SPRT. Bentea and Marinis extends aforementioned studies using SPRT to child bilingualism, examining online comprehension and production of multiple interrogatives in Romanian-English HS children. In contrast to the findings in the adult HS literature, they found no differences in online comprehension between HS children and monolingual children, but rather significant differences emerged in production, in which HS children produced less complex wh-movement structures. Together, the studies in this Research Topic employing SPRT reveal the importance of utilizing both online and offline measures to gauge on *what* HSs know and *how* they use that knowledge in real-time linguistic processing.

While self-paced reading task is an accessible, resource-efficient method that can be used to reveal how HSs process grammatical information in real-time, the following five studies take advantage of even more granular methodologies such as eye-tracking or EEG/ERP to examine linguistic processing in HSs. Özsoy et al. addressed the predictive use of case-marking in Turkish HSs and monolinguals, using both in-lab and web-cam based eye tracking. While both groups used case-marking to predict the

upcoming noun with in-lab eye tracking experiments, they were only able to replicate these results using web-based eye tracking with monolinguals, but not with HSs due to the greater variability in data collection environment. Similarly, but in a lab-based eye-tracking setup, Fuchs reports on Polish HSs' use of grammatical gender cues. Unlike Spanish, where gender cues are frequent in definite articles, Polish cues appear on optional and infrequent adjectives. The results show that HSs can use gender on inflected adjectives to fixate on the target noun faster when the cue uniquely identifies it. This supports a grammatical account rather than probabilistic account of the facilitative use of grammatical gender, indicating that HSs access abstract syntactic information in real time to aid word recognition. Sagarra and Casillas add to the previous two eye-tracking studies by investigated factors (e.g., AoA, language proficiency and use) affecting Spanish stress-tense suffix associations among adult Spanish-English HSs, English-Spanish L2 learners, and Spanish monolinguals. Results showed that all groups were fixating on target verbs, with monolinguals displaying more fixations. Higher proficiency increased fixations in HSs and L2 learners, while increased use affected only HSs. The study highlighted HSs' reliance on lexical competitors and phonotactic frequency over token frequency or AoA. Altogether, the eye-tracking studies of Özsoy et al., Fuchs, and Sagarra and Casillas nicely showcase the importance of investigating HL from distinct linguistic domains (and language combinations) to complement each other and further expand our understanding of linguistic online processing in HL bilinguals.

The last two studies employed online methods that have only recently been used in psycholinguistic studies of bilingualism. Martohardjono et al. looked at pupillary responses to syntactic island constructions in two groups of Spanish/English bilinguals (HSs and late bilinguals). The findings offer insights into individual variation in language processing among HSs and late bilinguals, emphasizing the importance of considering usage patterns and exposure levels in assessing language competence. In the only neurolinguistic (EEG/ERP) study of this Research Topic, Luque et al. explored grammatical gender knowledge and processing among HSs and highlight the complex interplay between linguistic representations and processing mechanisms. More precisely, they showed that HSs' bilingual experience modulated some aspects of morphosyntactic processing (expressed as P600 and biphasic N400 effects), corroborating similar findings observed in the late L2 learners' literature (e.g., Alemán Bañón et al., 2018; Grey, 2023). These results highlight the necessity to further include brain methods in HL bilingualism in order to better understand what underlies HSs competence and processing outcomes.

Together, these studies provide a comprehensive overview of diverse heritage language linguistic phenomena, socio-cultural and (individual) processing/mechanistic aspects within HS communities, shedding new light on the multifaceted nature of bilingual language development and maintenance.

Conclusion

This Research Topic offers an expansive overview of the intricate landscape of HL acquisition, processing, and maintenance. Through a diverse spectrum of empirical studies and theoretical

explorations, the contributions within this volume have brought to light the dynamics underlying the development and usage of HLs across the lifespan. They further highlight the complexity and richness that underlies HL bilingualism, emphasizing the intricate interplay between linguistic, (neuro)cognitive and socio-cultural factors in shaping HL acquisition. The findings presented in this Research Topic serve as a steppingstone for future research and pedagogical innovations, advancing our understanding of HL phenomena and their implications for linguistic theory, language education, and societal multilingualism. Moving forward, it is essential that we embrace the complexities and uniqueness within HL bilingualism, aim for more precise and inclusive methodologies, acknowledging the diverse experiences and trajectories of HL speakers worldwide.

Author contributions

FB: Conceptualization, Writing – original draft, Writing – review & editing. MK: Conceptualization, Writing – review & editing. SMPS: Conceptualization, Writing – review & editing.

References

- Alemán Bañón, J., Fiorentino, R., and Gabriele, A. (2018). Using event-related potentials to track morphosyntactic development in second language learners: the processing of number and gender agreement in Spanish. *PLoS ONE* 13:e0200791. doi: 10.1371/journal.pone.0200791
- Bayram, F., Rothman, J., Di Pisa, G., and Slabakova, R. (2021). “Current trends and emerging methodologies in charting heritage language bilingual grammars,” in *The Cambridge Handbook of Heritage Language and Linguistics*, eds S. Montrul, and M. Polinsky (Cambridge, MA: Cambridge University Press). doi: 10.1017/9781108766340.025
- Grey, S. (2023). Variability in native and nonnative language: an ERP study of semantic and grammar processing. *Stud. Second Lang. Acquis.* 45, 137–166. doi: 10.1017/S0272263122000055
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9781139030502
- Pereira Soares, S. M. (2022). *Examining Effects of Early (Heritage) Bilingualism for Later Multilingual Acquisition and Neurocognition* [Dissertation]. Konstanz: University of Konstanz.
- Polinsky, M. (2018). *Heritage Languages and their Speakers*. Cambridge, MA: Cambridge University Press. doi: 10.1017/9781107252349
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for
Psycholinguistics, Netherlands

REVIEWED BY

Cristina Maria Flores,
University of Minho, Portugal
Katrin Monika Schmitz,
University of Wuppertal, Germany

*CORRESPONDENCE

Sayuri Hayakawa
sayuri.hayakawa@northwestern.edu

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Communication

RECEIVED 15 July 2022

ACCEPTED 11 August 2022

PUBLISHED 31 August 2022

CITATION

Hayakawa S, Chung-Fat-Yim A and
Marian V (2022) Predictors of language
proficiency and cultural identification
in heritage bilinguals.
Front. Commun. 7:994709.
doi: 10.3389/fcomm.2022.994709

COPYRIGHT

© 2022 Hayakawa, Chung-Fat-Yim
and Marian. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Predictors of language proficiency and cultural identification in heritage bilinguals

Sayuri Hayakawa ^{1,2*}, Ashley Chung-Fat-Yim ¹ and
Viorica Marian ¹

¹Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL,
United States, ²Department of Psychology, Oklahoma State University, Stillwater, OK, United States

According to the 2020 U.S. Census Bureau, more than 66 million residents over the age of 5 in the United States speak a language other than English at home. Some bilinguals become dominant in the majority language that is spoken in the community as opposed to their native “heritage” language acquired at home. The objective of the current study was to uncover the predictors of language proficiency and cultural identification in different groups of heritage speakers. In our sample, heritage speakers acquired their heritage language first and English second and rated their proficiency in their heritage language lower than in English. We found that English proficiency was most reliably predicted by the duration of heritage language immersion, while heritage language proficiency was most reliably predicted by contexts of acquisition and exposure to both languages. Higher heritage language proficiency was associated with greater heritage language experience through friends and reading, less English experience through family, and later age of English acquisition. The trade-off between heritage language and English language experience was more pronounced for non-Spanish than Spanish heritage speakers. Finally, despite higher proficiency in English, cultural identification was higher with the heritage language, and was predicted by heritage language receptive proficiency and heritage language experience through family and reading. We conclude that self-reported proficiency and cultural identification differ depending on heritage speakers’ native languages, as well as how the heritage language and majority language are acquired and used. Our findings highlight the importance of taking individual language history into consideration when combining different groups of heritage speakers.

KEYWORDS

vocabulary knowledge, cultural identification, proficiency, native language, heritage speakers, bilingualism

Introduction

A growing percentage of the U.S. population speaks a language other than English at home. From 23.06 million in 1980 (Zeigler and Camarota, 2019) to 66.09 million in 2020 (U.S. Census Bureau, 2020), the number of people over the age of 5 who speak a non-English language at home has nearly tripled. These non-English home languages are often referred to as heritage languages and carry familial, cultural, and historical significance. Heritage bilinguals tend to feel strong personal connections to their heritage culture. However, as a result of acquiring the majority language at an early age and being formally educated in the majority language, heritage bilinguals generally prefer using the language of the community as opposed to their home language(s) (Valdés, 2000; Scontras et al., 2015).

Heritage bilinguals vary greatly in the age of second language acquisition and heritage language proficiency. While some heritage bilinguals immigrate to the host country with their parents and acquire the majority language in early childhood at school, others are born in the host country to foreign-born parents and acquire both languages simultaneously. Furthermore, while some heritage bilinguals have native-like proficiency in both languages, others show better linguistic command in the majority language than home language. Some can communicate fluently in both languages but are unable to read and write in the heritage language, and others have some understanding of the heritage language but have limited expressive skills (Montrul, 2005). Thus, heritage bilinguals are qualitatively distinct from second-language learners and native monolingual speakers (see Montrul, 2011 for review). Given that heritage speakers exist along a continuum of linguistic abilities and experiences, the present study aims to capture the linguistic predictors associated with self-reported measures of proficiency and cultural identification in different groups of heritage bilinguals.

Language proficiency

There are several factors impacting heritage language proficiency, including language exposure (Gathercole and Thomas, 2009; Hoff et al., 2012; Thomas et al., 2014; Gollan et al., 2015; Jia and Paradis, 2015; Unsworth, 2016; Hovsepian, 2018; Makarova et al., 2019; Giguere and Hoff, 2020; Tao et al., 2021; Vorobyeva and Bel, 2021) and frequency of use (Hakuta and D'Andrea, 1992; Bedore et al., 2012; Albirini, 2014; Chen et al., 2018; Schmid and Yilmaz, 2018; Daskalaki et al., 2019; Otwinowska et al., 2021) in and outside of the home. Access to a heritage language community that extends beyond the home context positively predicts heritage language vocabulary and lexical retrieval (Albirini, 2014; Gollan et al., 2015; Schmid and Yilmaz, 2018; Tao et al., 2021), morphosyntax (Kupisch and Rothman, 2018; Rodina et al., 2020; Torregrossa et al.,

2022), and pronunciation (Au and Romo, 1997; de Leeuw et al., 2010; Stoehr et al., 2017; Karayayla and Schmid, 2019; McCarthy and de Leeuw, 2022). Being surrounded by native speakers of the heritage language affords opportunities to listen and practice the language in various settings and discuss a wide variety of topics. Furthermore, heritage bilinguals exist along a continuum of linguistic abilities. In terms of reading and writing, heritage bilinguals are more likely to be literate in the majority language by virtue of being educated in that language. If heritage bilinguals do become literate in the heritage language, their reading skills tend to be better than their writing skills (Polinsky, 2015). Therefore, home and socio-linguistic contexts play important roles in the development of heritage language proficiency, and heritage bilinguals often exhibit variable degrees of fluency depending on the type of linguistic ability under examination (e.g., listening, speaking, reading, and writing).

In addition to the frequency of heritage language use, the age of second language acquisition and duration of immersion have been found to predict heritage language proficiency. The later a child becomes exposed to the majority language, the more likely they are to attain and retain competency in their heritage language (Polinsky and Kagan, 2007; Albirini, 2014; Jia and Paradis, 2015; Montrul, 2016; Gharibi and Boers, 2017; Meir et al., 2017; Armon-Lotem et al., 2021; Meir and Janssen, 2021). Studies have shown that sequential bilinguals often have greater proficiency in their heritage language than simultaneous bilinguals (e.g., Jia and Aaronson, 2003; Carreira and Kagan, 2011). For instance, children who acquire the majority language simultaneously or soon after the heritage language (e.g., before the age of 3) often score lower on tests of HL vocabulary (Gharibi and Boers, 2017; Armon-Lotem et al., 2021) and morphosyntax (Albirini, 2014; Jia and Paradis, 2015; Meir et al., 2017; Meir and Janssen, 2021) relative to children who spent more time learning the heritage language before acquiring the majority language. Age of acquisition predicts language aptitude and preference even among bilinguals who acquire the majority language later in adolescence. In a longitudinal study, Jia and Aaronson (2003) evaluated the changes in language preferences and Chinese proficiency among native Chinese-speaking children and adolescents who immigrated to the United States. Participants who immigrated to the United States at an early age (before the age of 9) switched their language preference from Chinese to English and became more proficient in English than Chinese within the first year. Those who immigrated to the United States at a later age (between 10 and 16 years of age) maintained their preference for Chinese across all 3 years and continued to use Chinese with their parents and siblings. Altogether, such findings demonstrate that both age of second language acquisition and duration of immersion influence heritage language proficiency.

To determine which factors promote heritage language proficiency, Gollan et al. (2015) tested Chinese-English and Spanish-English heritage bilingual adults on the Multilingual

Naming Test (MINT; Gollan et al., 2012), which is an *objective* measure of language proficiency. For Chinese-English bilinguals, higher heritage language proficiency was associated with exposure to a greater number of heritage speakers during childhood. For Spanish-English bilinguals, higher heritage language proficiency was instead associated with less English use. The authors proposed that the differences between groups may stem from cross-cultural variations in the interpretation of the questionnaire items. Across all participants, proficiency in the heritage language was uniquely predicted by the number of heritage language speakers encountered during childhood, the primary caregiver's level of English proficiency, and the participants' age of English acquisition. For Persian-English bilingual children, parents' attitude toward the heritage language was the strongest predictor of heritage language proficiency (as measured by a verbal fluency task and auditory picture-word matching test) in simultaneous bilinguals, whereas the age at emigration was the strongest predictor of heritage language proficiency in sequential heritage bilinguals (Gharibi and Boers, 2017). These findings demonstrate that individual variation within and across different groups of heritage bilinguals influences heritage language proficiency. The present study thus compares Spanish-English heritage speakers to other groups of heritage speakers (i.e., non-Spanish) living in the United States on heritage language and English proficiency ratings.

Cultural identification

Language serves as a bridge for creating a sense of belonging to an ethnic group in children (Yu, 2015; Arredondo et al., 2016), adolescents (Phinney et al., 2001; Oh and Fuligni, 2010), and adults (Noels et al., 1996; Cho, 2000; Chen et al., 2008; Gathbonton and Trofimovich, 2008; Yu, 2015). Across all ages, greater proficiency in a heritage language is associated with stronger ethnic identity and affiliation with the ethnic group. However, heritage speakers vary in their cultural affiliation toward their heritage language and majority language. For example, individuals from minority groups sometimes report conflicting identities, in which they want to preserve the cultural values associated with their heritage language, but also want to fit in with the culture associated with the majority language (Phinney, 1990). On open-ended questions from the Multigroup Ethic Identity Measure and Ethnic Identity Scale, Arredondo et al. (2016) found that Spanish-English heritage bilingual children reported feeling a sense of pride for being able to speak Spanish, showed an appreciation for cultural diversity, enjoyed communicating exclusively with friends and family in a "secret" language, and expressed positivity toward helping their parents learn English and in turn, learning Spanish from their parents. In the same study, some of the children described Spanish as confusing or too difficult at times. Furthermore, heritage bilinguals are more likely to assimilate

to the customs and practices of the host culture with each successive generation compared to the last (Felix-Ortiz et al., 1994). Hence, among heritage bilinguals, factors related to heritage language proficiency and migration, such as age of second language acquisition and duration of immersion, may predict cultural identification.

In a large heterogeneous sample of adult bilinguals varying in language and cultural backgrounds, Schroeder et al. (2017) identified the linguistic factors that predict cultural identification. Increased first language (L1) exposure through media, higher L1 proficiency, fewer years immersed in a second language (L2) family context, but more years immersed in an L2 school/work context led to increased first-language cultural affiliation. In contrast, increased immersion in an L2 school/work context, lower L2 perceived accent, and earlier L2 age of acquisition was associated with increased cultural identification with the second-language culture. These findings demonstrate that factors related to the second language influence both first-language and second-language cultural affiliation, whereas factors associated with the first language only influence first-language cultural identification. These effects also differed by age of L2 acquisition and whether the language was learned in a formal or informal context. Schroeder et al. argued that through language, bilinguals can access their culture by interacting with members of the same cultural group, actively participating in various cultural activities, and engaging in media from that culture (through TV, radio, and books). To our knowledge, no study to date has taken a similar approach in identifying linguistic predictors of cultural identification among different groups of heritage bilinguals.

With over 40.5 million people over the age of 5 speaking Spanish at home, Spanish is by far the most spoken non-English language in the United States (U.S. Census Bureau, 2020). Therefore, Spanish heritage bilinguals may have more opportunities to use and practice with native speakers and engage in cultural activities with members of the same cultural group than other non-English heritage bilinguals. For example, in the city of Chicago, Latinos are the second largest ethnic group at 29.7%, whereas Asians and other cultural groups make up around 12.7% of the city's population (The Economist, 2017). Second, the one-to-one mapping between the Spanish language and Latino culture is less clear, as multiple cultural groups from various countries speak Spanish. In contrast, the mapping between language and culture for other languages is more consistent (e.g., Korean with Korea). Hence, there is the possibility that the linguistic and cultural experiences of Spanish heritage bilinguals are more diverse and less homogeneous compared to non-Spanish heritage bilinguals. Third, studies have shown that the motivation for maintaining the heritage language differs between Spanish-English heritage learners and non-Spanish heritage learners. Hur et al. (2021) examined the expectations and attitudes toward heritage language courses. While Spanish-English heritage learners perceived their classes

TABLE 1 Linguistic profiles of Spanish and non-Spanish heritage speakers.

Measure	Heritage group	Heritage language	English	HL vs. English
<i>Proficiency (0 = None to 10 = Perfect)</i>				
Speaking	Spanish	8.08 (1.10)	9.56 (0.66)	***
	Non-Spanish	7.72 (1.60)	9.43 (0.79)	***
Reading	Spanish	7.59 (1.52)	9.65 (0.60)	***
	Non-Spanish	6.04 (2.89)	9.44 (0.86)	***
Understanding	Spanish	8.76 (1.05)	9.68 (0.63)	***
	Non-Spanish	8.43 (1.27)	9.52 (0.84)	***
<i>Age of Acquisition</i>				
Overall Acquisition	Spanish	1.09 (1.14)	4.93 (1.89)	***
	Non-Spanish	0.48 (0.67)	4.95 (2.13)	***
Reading Acquisition	Spanish	5.90 (2.54)	6.02 (1.77)	
	Non-Spanish	4.97 (2.85)	6.04 (2.18)	*
<i>Context of Acquisition (0 = Not a Contributor to 10 = Most Important Contributor)</i>				
Family	Spanish	9.38 (1.75)	4.04 (3.35)	***
	Non-Spanish	9.46 (0.95)	3.41 (3.20)	***
Friends	Spanish	5.15 (3.20)	8.66 (1.82)	***
	Non-Spanish	5.56 (3.15)	9.04 (1.72)	***
Individual (Language Tapes/Self instruction)	Spanish	1.37 (2.18)	3.34 (3.91)	***
	Non-Spanish	1.69 (2.25)	2.80 (3.36)	*
TV	Spanish	5.95 (2.98)	8.15 (1.79)	***
	Non-Spanish	5.76 (2.96)	6.43 (2.48)	
Radio/Music	Spanish	5.72 (3.23)	6.95 (3.07)	**
	Non-Spanish	2.09 (2.61)	3.98 (3.32)	***
Reading	Spanish	5.72 (2.88)	8.86 (1.46)	***
	Non-Spanish	4.91 (2.99)	8.37 (2.56)	***
<i>Context of Exposure (0 = Never to 10 = Always)</i>				
Family	Spanish	8.93 (2.1)	4.40 (3.29)	***
	Non-Spanish	8.48 (2.3)	3.39 (3.04)	***
Friends	Spanish	3.81 (2.90)	8.81 (2.00)	***
	Non-Spanish	3.94 (2.92)	8.98 (1.93)	***
Individual (Language-Lab/Self-instruction)	Spanish	2.01 (2.88)	2.92 (3.76)	*
	Non-Spanish	0.72 (1.37)	1.65 (3.11)	**
TV	Spanish	4.18 (3.02)	8.70 (1.77)	***
	Non-Spanish	3.65 (3.27)	7.17 (3.01)	***
Radio/Music	Spanish	5.30 (3.38)	7.93 (2.13)	***
	Non-Spanish	3.94 (3.19)	7.48 (2.56)	***
Reading	Spanish	3.40 (2.37)	8.73 (1.76)	***
	Non-Spanish	2.17 (2.20)	8.44 (2.82)	***
<i>Immersion (years)</i>				
Family	Spanish	23.57 (8.34)	15.86 (11.38)	***
	Non-Spanish	20.85 (6.77)	9.40 (11.17)	***
Country	Spanish	7.53 (8.21)	21.74 (6.47)	***
	Non-Spanish	6.18 (5.35)	17.28 (6.56)	***
School/Work	Spanish	7.92 (7.99)	18.05 (6.27)	***
	Non-Spanish	5.78 (5.85)	15.69 (5.60)	***

Values in parentheses represent standard deviations. Boxed values indicate significant differences between Spanish and non-Spanish heritage speakers (across rows; $p < 0.05$). Asterisks represent significant differences between the Heritage Language and English for each group (across columns). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

TABLE 2 Effects of heritage language experience on self-reported heritage language and English proficiency.

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
(Intercept)	8.77	0.08	109	116.19	<0.001	***
Language	1.60	0.07	327	21.67	<0.001	***
Heritage group	−0.46	0.16	109	−2.91	0.004	**
Measure	−0.04	0.07	327	−0.57	0.566	
AoA	−0.10	0.08	109	−1.25	0.215	
Immersion	−0.05	0.07	109	−0.65	0.519	
Family	−0.07	0.07	109	−0.97	0.337	
Friends	0.10	0.08	109	1.28	0.205	
Media	0.00	0.08	109	0.02	0.983	
Reading	0.35	0.08	109	4.33	<0.001	***
Individual	−0.42	0.09	109	−4.86	<0.001	***
Language:Heritage	0.38	0.15	327	2.47	0.014	*
Language:Measure	0.30	0.15	327	2.02	0.044	*
Heritage:Measure	−0.31	0.15	327	−2.02	0.044	*
Language:AoA	0.10	0.08	327	1.24	0.217	
Heritage:AoA	0.02	0.18	109	0.09	0.930	
Measure:AoA	−0.07	0.08	327	−0.83	0.410	
Language:Immersion	−0.22	0.07	327	−3.05	0.002	**
Heritage:Immersion	−0.24	0.15	109	−1.56	0.121	
Measure:Immersion	0.04	0.07	327	0.54	0.592	
Language:Family	−0.06	0.07	327	−0.82	0.413	
Heritage:Family	0.03	0.14	109	0.19	0.853	
Measure:Family	−0.11	0.07	327	−1.66	0.098	
Language:Friends	−0.17	0.08	327	−2.29	0.023	*
Heritage:Friends	−0.35	0.16	109	−2.18	0.031	*
Measure:Friends	0.01	0.08	327	0.20	0.844	
Language:Media	−0.05	0.08	327	−0.62	0.533	
Heritage:Media	0.23	0.16	109	1.44	0.153	
Measure:Media	0.01	0.08	327	0.07	0.945	
Language:Reading	−0.50	0.08	327	−6.31	<0.001	***
Heritage:Reading	0.70	0.17	109	4.22	<0.001	***
Measure:Reading	0.20	0.08	327	2.55	0.011	*
Language:Individual	0.30	0.09	327	3.51	0.001	**
Heritage:Individual	−0.57	0.18	109	−3.16	0.002	**
Measure:Individual	−0.06	0.09	327	−0.69	0.488	
Language:Heritage:Measure	0.52	0.31	327	1.69	0.091	
Language:Heritage:AoA	0.25	0.17	327	1.42	0.157	
Language:Measure:AoA	0.12	0.16	327	0.74	0.460	
Heritage:Measure:AoA	−0.09	0.17	327	−0.53	0.597	
Language:Heritage:Immersion	0.00	0.15	327	0.01	0.994	
Language:Measure:Immersion	−0.14	0.14	327	−0.97	0.333	
Heritage:Measure:Immersion	0.10	0.15	327	0.66	0.511	
Language:Heritage:Family	0.05	0.13	327	0.41	0.680	
Language:Measure:Family	0.23	0.14	327	1.72	0.087	
Heritage:Measure:Family	−0.05	0.13	327	−0.41	0.684	

(Continued)

TABLE 2 (Continued)

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
Language:Heritage:Friends	0.04	0.16	327	0.23	0.815	
Language:Measure:Friends	−0.03	0.15	327	−0.20	0.843	
Heritage:Measure:Friends	0.12	0.16	327	0.76	0.445	
Language:Heritage:Media	−0.25	0.16	327	−1.59	0.112	
Language:Measure:Media	0.04	0.15	327	0.29	0.773	
Heritage:Measure:Media	−0.09	0.16	327	−0.56	0.578	
Language:Heritage:Reading	−0.90	0.16	327	−5.52	<0.001	***
Language:Measure:Reading	−0.38	0.16	327	−2.40	0.017	*
Heritage:Measure:Reading	0.23	0.16	327	1.44	0.150	
Language:Heritage:Individual	0.56	0.18	327	3.15	0.002	**
Language:Measure:Individual	0.10	0.17	327	0.59	0.554	
Heritage:Measure:Individual	−0.04	0.18	327	−0.23	0.817	
Language:Heritage:Measure:AoA	0.54	0.35	327	1.55	0.123	
Language:Heritage:Measure:Immersion	−0.09	0.30	327	−0.31	0.755	
Language:Heritage:Measure:Family	−0.05	0.27	327	−0.20	0.844	
Language:Heritage:Measure:Friends	−0.08	0.31	327	−0.24	0.807	
Language:Heritage:Measure:Media	0.23	0.32	327	0.74	0.462	
Language:Heritage:Measure:Reading	−0.47	0.33	327	−1.45	0.147	
Language:Heritage:Measure:Individual	−0.36	0.36	327	−1.02	0.310	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

as a necessary tool for professional success, Korean-English heritage learners used their classes as a way to reconnect with their Korean culture and other members of their heritage language community. For these reasons, we distinguish between Spanish and non-Spanish heritage bilinguals to examine how predictors of proficiency and cultural identification are moderated by native language background.

The present study

The goal of the present study was to identify the predictors of self-reported language proficiency and cultural identification in different groups of heritage speakers. Specifically, we examined how age of acquisition, duration of immersion, and contexts of acquisition and exposure (i.e., through friends, family, media, reading, and language tapes and self-instruction) influenced self-reported measures of proficiency and cultural identification in the heritage language and in English among Spanish heritage bilinguals and non-Spanish heritage bilinguals. Considering heritage bilinguals typically have better comprehension than oral skills in their heritage language (Polinsky, 2015), we separated expressive (speaking) from receptive (understanding and reading) proficiency in our analyses. Based on past research, we hypothesized that heritage language proficiency and cultural identification will be predicted by heritage language usage in informal contexts, such as in the home through family

and in the community through friends (Gollan et al., 2015; Jia and Paradis, 2015; Montrul, 2016), as well as the age of English acquisition and length of immersion in an English-speaking country (Montrul, 2008; Gathercole and Thomas, 2009; Vorobyeva and Bel, 2021). In addition, we hypothesized that English proficiency and cultural identification will be predicted by *both* heritage language and English usage in informal contexts (i.e., home, friends) *and* more formal individual contexts (i.e., language tapes, language labs, and self-instruction). Altogether, the present study provides a deeper understanding of the interactivity between language and culture in heritage bilinguals.

Materials and methods

Participants

Participants included 133 heritage speakers who acquired a non-English native language first and English second, and who rated English as more proficient than their native language on the *Language Experience and Proficiency Questionnaire* (LEAP-Q; Marian et al., 2007). Data were compiled from previous studies conducted in our lab between 2011 and 2022 (i.e., secondary data analysis; Bartolotti et al., 2011; Chabal et al., 2015, 2022; Freeman et al., 2016, 2022; Shook and Marian, 2016; Chen et al., 2017; Marian et al., 2018, 2021; Hayakawa et al., 2020). Participants' mean age at the time of testing was 23.97 ($SD = 6.24$), and 67% were female. Seventy-nine

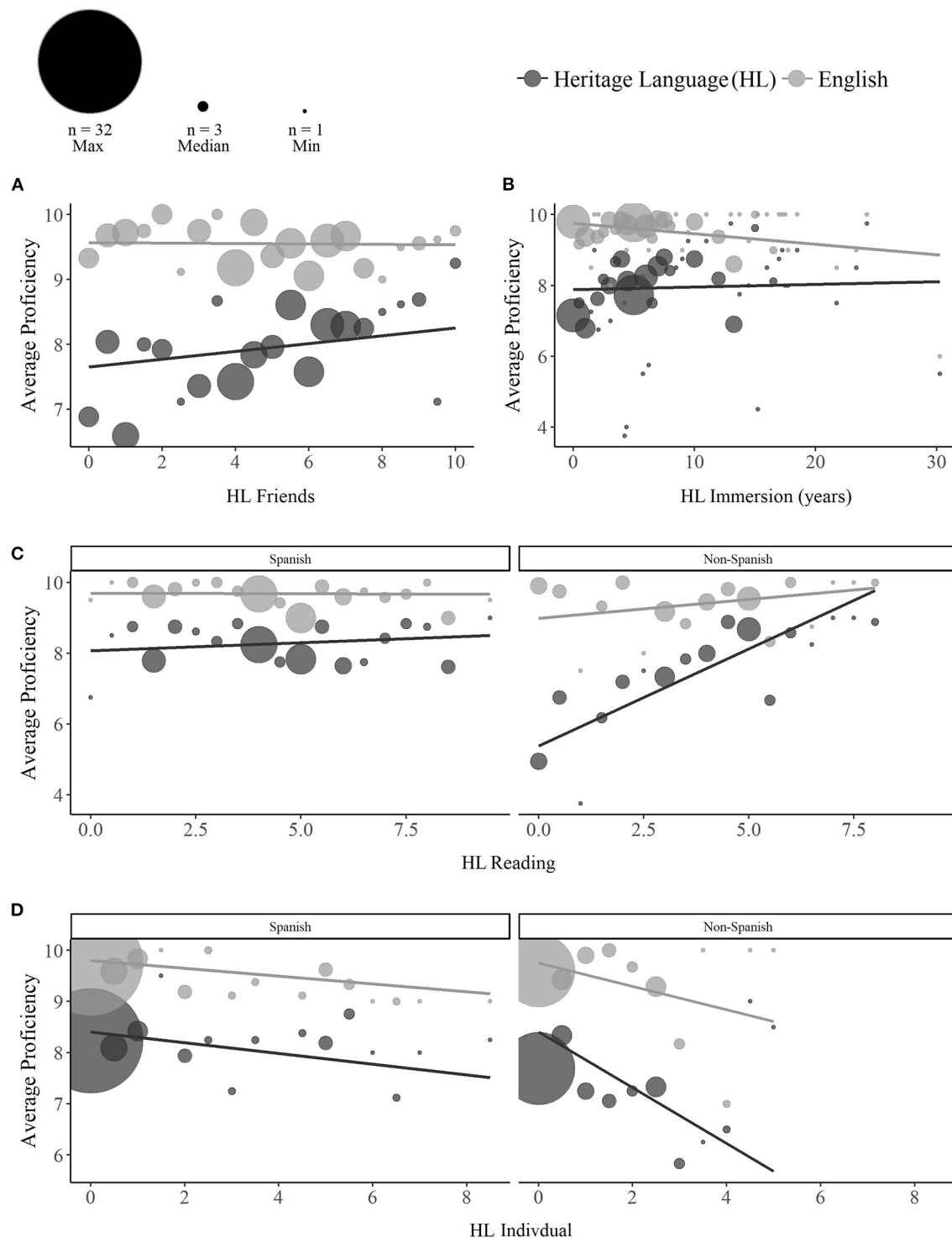


FIGURE 1

Effects of heritage language (HL) experience on self-reported HL (dark gray) and English (light gray) proficiency. Across both groups, HL proficiency increased with greater HL acquisition and exposure through friends (A), while English proficiency decreased with greater HL immersion duration (B). HL proficiency increased with greater HL reading acquisition and exposure for non-Spanish, but not Spanish bilinguals (C). Overall proficiency decreased with greater HL individual acquisition and exposure, which was particularly the case for HL proficiency among non-Spanish bilinguals (D). Dot sizes reflect the number of participants contributing to each aggregated value (max = 32, median = 3, min = 1).

TABLE 3 Effects of English experience on self-reported heritage language and English proficiency.

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
(Intercept)	8.70	0.09	110	99.15	<0.001	***
Language	1.71	0.08	330	20.53	<0.001	***
Heritage group	−0.42	0.18	110	−2.38	0.019	*
Measure	−0.10	0.08	330	−1.18	0.237	
AoA	0.14	0.09	110	1.53	0.129	
Immersion	0.06	0.11	110	0.51	0.612	
Family	−0.19	0.10	110	−2.01	0.047	*
Friends	0.01	0.10	110	0.10	0.917	
Media	0.03	0.11	110	0.24	0.807	
Reading	0.06	0.12	110	0.49	0.624	
Individual	−0.12	0.09	110	−1.32	0.191	
Language:Heritage	0.87	0.17	330	5.16	<0.001	***
Language:Measure	0.34	0.17	330	2.05	0.041	*
Heritage:Measure	−0.40	0.17	330	−2.35	0.019	*
Language:AoA	−0.35	0.09	330	−3.99	<0.001	***
Heritage:AoA	0.40	0.18	110	2.17	0.032	*
Measure:AoA	0.05	0.09	330	0.61	0.541	
Language:Immersion	0.14	0.11	330	1.30	0.193	
Heritage:Immersion	0.59	0.24	110	2.47	0.015	*
Measure:Immersion	−0.06	0.11	330	−0.57	0.570	
Language:Family	0.23	0.09	330	2.53	0.012	*
Heritage:Family	−0.22	0.20	110	−1.08	0.283	
Measure:Family	0.11	0.09	330	1.22	0.225	
Language:Friends	−0.17	0.09	330	−1.82	0.069	~
Heritage:Friends	0.21	0.20	110	1.09	0.277	
Measure:Friends	0.00	0.09	330	0.03	0.979	
Language:Media	0.13	0.11	330	1.19	0.235	
Heritage:Media	−0.24	0.22	110	−1.11	0.271	
Measure:Media	−0.07	0.11	330	−0.68	0.499	
Language:Reading	0.27	0.11	330	2.41	0.016	*
Heritage:Reading	−0.03	0.22	110	−0.16	0.874	
Measure:Reading	0.10	0.11	330	0.89	0.373	
Language:Individual	0.08	0.09	330	0.93	0.353	
Heritage:Individual	0.13	0.19	110	0.67	0.507	
Measure:Individual	−0.06	0.09	330	−0.76	0.449	
Language:Heritage:Measure	0.65	0.34	330	1.93	0.054	~
Language:Heritage:AoA	−0.35	0.18	330	−2.00	0.046	*
Language:Measure:AoA	−0.02	0.18	330	−0.12	0.906	
Heritage:Measure:AoA	0.02	0.18	330	0.12	0.907	
Language:Heritage:Immersion	−0.36	0.23	330	−1.57	0.118	
Language:Measure:Immersion	0.14	0.22	330	0.65	0.515	
Heritage:Measure:Immersion	−0.23	0.23	330	−1.01	0.313	
Language:Heritage:Family	0.42	0.19	330	2.19	0.029	*
Language:Measure:Family	−0.07	0.18	330	−0.39	0.697	

(Continued)

TABLE 3 (Continued)

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
Heritage:Measure:Family	0.00	0.19	330	0.00	0.996
Language:Heritage:Friends	−0.19	0.19	330	−0.99	0.323
Language:Measure:Friends	−0.07	0.18	330	−0.36	0.721
Heritage:Measure:Friends	0.25	0.19	330	1.36	0.173
Language:Heritage:Media	0.20	0.21	330	0.94	0.350
Language:Measure:Media	0.11	0.21	330	0.50	0.619
Heritage:Measure:Media	−0.08	0.21	330	−0.40	0.692
Language:Heritage:Reading	−0.10	0.21	330	−0.48	0.632
Language:Measure:Reading	−0.08	0.22	330	−0.38	0.704
Heritage:Measure:Reading	−0.16	0.21	330	−0.78	0.434
Language:Heritage:Individual	−0.11	0.18	330	−0.61	0.539
Language:Measure:Individual	−0.05	0.17	330	−0.29	0.772
Heritage:Measure:Individual	0.16	0.18	330	0.89	0.375
Language:Heritage:Measure: AoA	−0.19	0.35	330	−0.54	0.592
Language:Heritage:Measure:Immersion	0.09	0.46	330	0.20	0.845
Language:Heritage:Measure:Family	0.00	0.38	330	−0.01	0.994
Language:Heritage:Measure:Friends	−0.46	0.37	330	−1.23	0.221
Language:Heritage:Measure:Media	0.33	0.42	330	0.79	0.430
Language:Heritage:Measure:Reading	0.10	0.42	330	0.24	0.807
Language:Heritage:Measure:Individual	−0.50	0.36	330	−1.38	0.169

~ $p < 0.08$, * $p < 0.05$, *** $p < 0.001$.

participants had Spanish as their heritage language, while the remaining 54 participants had a non-Spanish language as their heritage language. The non-Spanish languages all utilized a different script than English and included Korean ($n = 31$), Chinese ($n = 16$), Thai ($n = 4$), Hebrew, Russian, and Tamil ($n = 1$ each). Spanish and non-Spanish heritage speakers did not significantly differ in age ($M = 24.68$ and 22.93 , respectively), gender (67.1 and 67.6% female), or years of education ($M = 14.66$ and 15.21), $ps > 0.137$. Non-Spanish bilinguals knew marginally more languages ($M = 2.43$) than Spanish bilinguals ($M = 2.22$), $p = 0.067$. Refer to Table 1 for the linguistic profiles of each group of heritage bilinguals (Spanish and non-Spanish bilinguals), including self-reported heritage language (HL) and English proficiency, ages of HL and English acquisition, and contexts of HL and English acquisition and exposure. Participants had no history of a language or learning disability and had normal or corrected-to-normal vision.

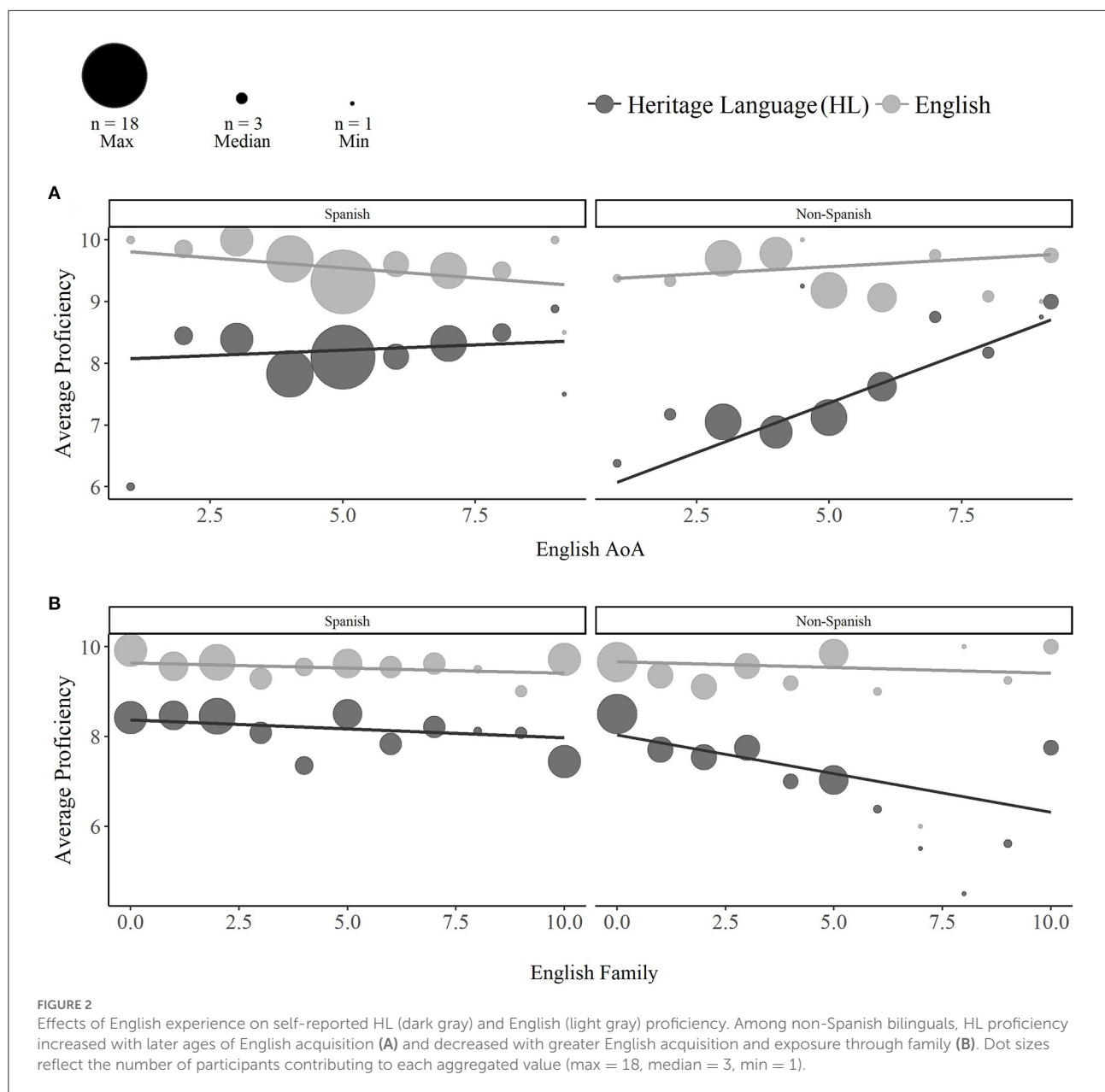
Materials

Language experience and proficiency questionnaire

The *Language Experience and Proficiency Questionnaire* (LEAP-Q; Marian et al., 2007) was used to acquire each participant's linguistic profile. Participants were asked to list the languages they know in order of dominance as well as

acquisition. Information about each language's (1) acquisition, (2) proficiency, and (3) exposure were obtained. For age of acquisition, participants provided the ages at which they began acquiring, became fluent, began reading, and became fluent at reading each language. For proficiency, participants rated their proficiency in each language in terms of speaking, understanding, and reading on a scale from 0 (None) to 10 (Perfect). For manner of acquisition, participants rated the extent to which various factors contributed to learning each language on a scale from 0 (Not a Contributor) to 10 (Most Important Contributor). These factors included friends, family, reading, language tapes/self-instruction, watching TV, and listening to radio/music. For language exposure, participants rated the extent to which they were currently exposed to each language in various contexts, including friends, family, watching TV, listening to radio/music, reading, and language lab/self-instruction on a scale from 0 (Never) to 10 (Always).

Cultural identification information was obtained by asking participants to list the cultures they identified with and rate the extent to which they identified with each culture on a scale from 0 (No Identification) to 10 (Complete Identification). Cultural identification with the HL and English was determined based on ratings given to cultures associated with each language (e.g., "Korea" for cultural identification with Korean as a HL, "USA" for cultural identification with English). If more than one culture associated with a language was listed (e.g., Culture 1: "Latino"



and Culture 2: “Mexican” for Spanish), we selected the rating for the culture that was ranked highest. In addition to linguistic and cultural information, demographic information such as age, gender, years of formal education, highest level of education, year of migration to the United States (if applicable), and any history of vision, hearing, language, or learning disabilities were provided by each participant. Although some participants were fluent in a third language, we did not analyze the third language information due to the small number of participants who were fluent in a third language.

Procedure

All studies included in the secondary analysis were reviewed and approved by Northwestern University’s Institutional Review Board. In all studies, the *Language Experience and Proficiency Questionnaire* (Marian et al., 2007) was administered toward the end of the testing session. Participants provided informed consent prior to the start of the experiment and were debriefed at the end of the experiment.

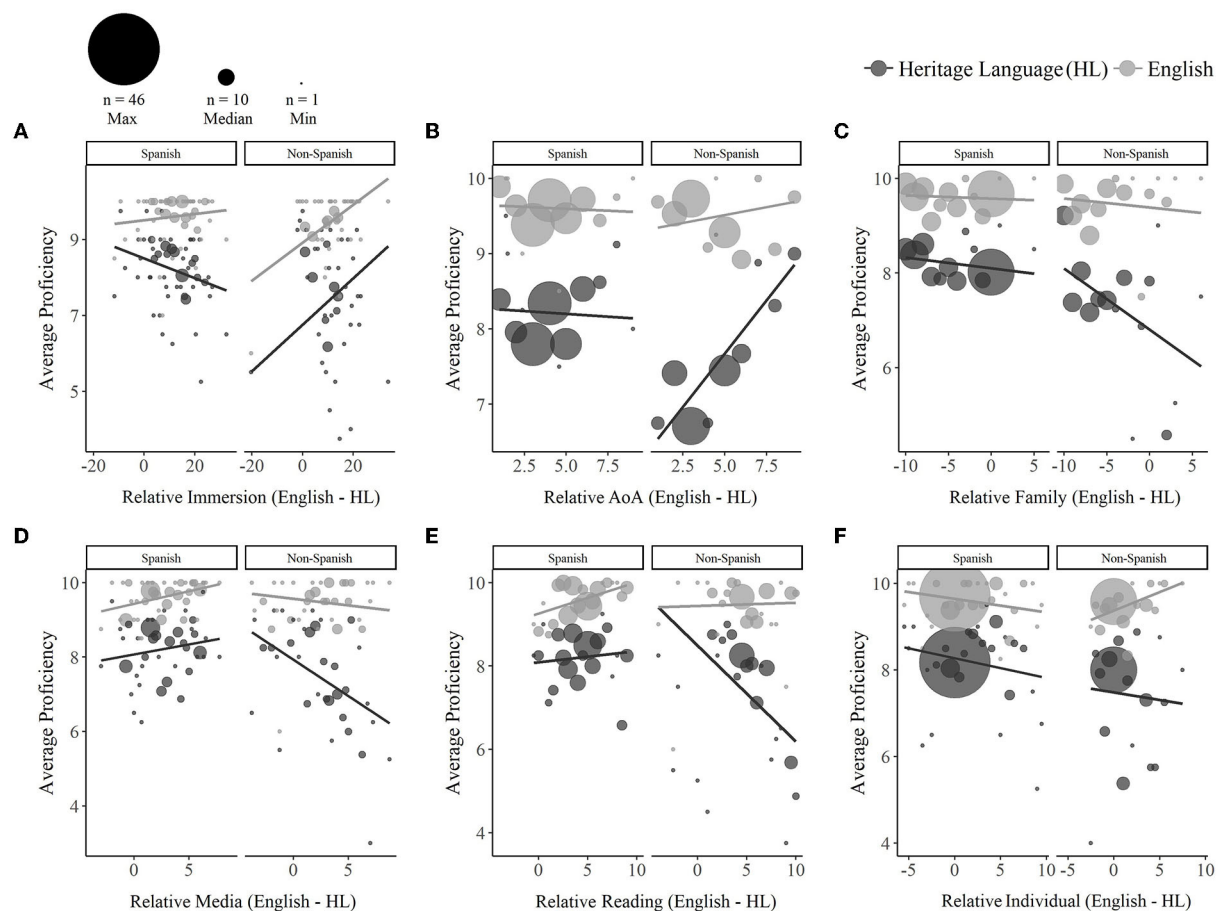


FIGURE 3

Effects of relative language experience (English—HL) on self-reported HL (dark gray) and English (light gray) proficiency. Among non-Spanish bilinguals, both HL and English proficiency increased with relatively greater English (vs. HL) immersion, while for Spanish bilinguals, HL proficiency decreased with relatively greater English immersion (A). Among non-Spanish bilinguals, lower HL proficiency was predicted by relatively earlier ages of English (vs. HL) acquisition (B), as well as relatively higher ratings of English (vs. HL) acquisition and exposure through family (C), media (D), and reading (E). Higher ratings of English (vs. HL) acquisition and exposure in individual contexts was non-significantly associated with higher English proficiency and lower HL proficiency (F). Dot sizes reflect the number of participants contributing to each aggregated value (max = 46, median = 10, min = 1).

Data analysis

Two sets of analyses were conducted to examine predictors of heritage language (HL) and English proficiency and cultural identification among Spanish and non-Spanish heritage speakers. To address issues of multicollinearity, we began by examining the correlational structure of LEAP-Q measures and created 7 composite measures for each language, which included **Age of Acquisition (AoA)**, **Duration of Immersion** (average number of years immersed in a country, school, or workplace in which each language was spoken), and five composite measures which each represented an aggregated measure of manner of acquisition and current exposure in different contexts. The included contexts were **Family Acquisition and Exposure**

(averaged across ratings of how much family contributed to the acquisition of each language and how much participants are currently exposed to each language through family), **Friends Acquisition and Exposure**, **Media Acquisition and Exposure** (e.g., through TV, radio), **Reading Acquisition and Exposure**, and **Individual Acquisition and Exposure** (e.g., through language tapes/language lab/self-instruction). In order to assess the impact of relative language experience, we additionally calculated a dominance score for each composite measure by subtracting the HL score from the English score. All fixed effects had VIF scores < 5, indicating minimal multicollinearity.

Effects of HL, English, and relative language experience measures within each set of analyses were examined with separate linear mixed-effects models, with variable numbers of

participants depending on the availability of relevant proficiency or cultural identification measures for individual subjects. Models therefore included effects of (1) HL experience on HL and English receptive (averaged across understanding and reading) and expressive (speaking) proficiency ($n = 126$), (2) English experience on HL and English receptive and expressive proficiency ($n = 127$), (3) relative language experience on HL and English receptive and expressive proficiency ($n = 126$), (4) HL experience on HL and English cultural identification ($n = 79$), (5) English experience on HL and English cultural identification ($n = 79$), and (6) relative language experience on HL and English cultural identification ($n = 79$).

Fixed effects for proficiency models included the 7 HL, English, or relative language experience composite measures plus all two-, three-, and four-way interactions with Heritage Group (Spanish vs. non-Spanish), Language (HL vs. English proficiency), and Measure (receptive vs. expressive proficiency). Cultural identification models included the 7 composite measures, receptive and expressive proficiency, plus all two- and three-way interactions with Heritage Group (Spanish vs. non-Spanish) and Language (HL vs. English cultural identification). All models included a random intercept for participant. Contrasts for Heritage Group (Spanish: -0.57 vs. Non-Spanish: $+0.43$), Language (HL: -0.5 vs. English: $+0.5$), and Measure (Expressive: -0.5 vs. Receptive: $+0.5$) were centered and weighted by the number of responses. Continuous fixed effects were mean-centered and scaled via z-score transformation.

Parameter estimates and significance of fixed effects were assessed with the Satterwhite method using the *lme4* (Bates et al., 2014) and *lmerTest* (Kuznetsova et al., 2017) R packages. Tukey-adjusted follow-up tests of simple effects were conducted using the *emmeans* and *emtrends* functions of the *emmeans* R package (Lenth et al., 2018).

Results

Predictors of heritage language and English proficiency

Effects of heritage language (HL) experience

Self-reported proficiency was significantly higher in English ($M = 9.55$, 95% CI [9.38, 9.73]) than in the heritage language ($M = 7.93$, 95% CI [7.76, 8.10]), $p < 0.001$. See Table 2 for full output. A two-way interaction between Language and Heritage Group ($p = 0.014$) indicated that Spanish bilinguals had significantly higher HL proficiency than non-Spanish bilinguals [*Estimate* = 0.64, *SE* = 0.17, $t_{(163.94)} = 3.70$, $p < 0.001$], whereas the two groups did not differ in English proficiency [*Estimate* = 0.27, *SE* = 0.17, $t_{(163.94)} = 1.53$, $p = 0.128$].

Across both groups, the composite measures of HL Acquisition and Exposure through **friends** (Language x Friends: $p = 0.023$; Figure 1A) and **reading** (Language x Reading: $p <$

0.001) predicted higher self-reported HL proficiency [Friends: *Estimate* = 0.16, *SE* = 0.09, $t_{(163.94)} = 1.81$, $p = 0.070$; Reading: *Estimate* = 0.68, *SE* = 0.09, $t_{(163.94)} = 7.33$, $p < 0.001$], but not English proficiency ($ps > 0.202$). The effect of reading experience was greater for receptive HL proficiency [*Estimate* = 0.89, *SE* = 0.11, $t_{(274)} = 8.18$, $p < 0.001$] compared to expressive proficiency [*Estimate* = 0.47, *SE* = 0.11, $t_{(274)} = 4.27$, $p < 0.001$; Language x Measure x Reading: $p = 0.017$].

In contrast, a two-way interaction between Language and Immersion ($p = 0.002$) revealed that a longer duration of HL **immersion** was associated with significantly lower English proficiency [*Estimate* = -0.17 , *SE* = 0.08, $t_{(163.94)} = -2.03$, $p = 0.044$], but not HL proficiency [*Estimate* = 0.05, *SE* = 0.08, $t_{(163.94)} = 0.53$, $p = 0.595$; Figure 1B]. Greater HL acquisition and exposure through **individual** contexts (e.g., self-instruction, language labs, and language tapes) was associated with lower proficiency overall ($p < 0.001$), which was particularly the case for HL proficiency [*Estimate* = -0.63 , *SE* = 0.10, $t_{(164)} = -6.25$, $p < 0.001$] compared to English proficiency [*Estimate* = -0.30 , *SE* = 0.10, $t_{(164)} = -2.91$, $p = 0.004$; Language x Individual: $p = 0.001$].

Finally, three-way interactions with Language and Heritage Group revealed that ratings of HL acquisition and exposure through **reading** ($p < 0.001$) and **individual** contexts ($p = 0.002$) were more predictive of HL proficiency for non-Spanish bilinguals [Reading: *Estimate* = 1.25, *SE* = 0.15, $t_{(164)} = 8.57$, $p < 0.001$; Individual: *Estimate* = -1.06 , *SE* = 0.17, $t_{(164)} = -6.36$, $p < 0.001$] than Spanish bilinguals [Reading: *Estimate* = 0.25, *SE* = 0.15, $t_{(164)} = 1.68$, $p = 0.095$; Individual: *Estimate* = -0.21 , *SE* = 0.12, $t_{(164)} = -1.78$, $p = 0.077$; see Figures 1C,D].

Effects of English experience

Earlier **ages of English acquisition** (Language x AoA: $p < 0.001$) and higher ratings of English acquisition and exposure through **family** (Language x Family: $p = 0.012$) predicted lower HL proficiency [AoA: *Estimate* = 0.36, *SE* = 0.10, $t_{(163)} = 3.50$, $p < 0.001$; Family: *Estimate* = -0.34 , *SE* = 0.11, $t_{(163)} = -3.04$, $p < 0.001$], but not English proficiency ($ps > 0.48$). Refer to Table 3 for full output. Although simple effects did not reach significance, an interaction between Language and Reading ($p = 0.016$) indicated that higher ratings of English acquisition and exposure through **reading** were associated with lower HL proficiency [*Estimate* = -0.08 , *SE* = 0.12, $t_{(163)} = -0.62$, $p = 0.536$], but higher English proficiency [*Estimate* = 0.18, *SE* = 0.12, $t_{(163)} = 1.52$, $p = 0.131$].

Three-way interactions with Language and Heritage Group revealed that the effects of **age of English acquisition** ($p = 0.046$) and **family** ($p = 0.029$) on HL proficiency were greater for non-Spanish bilinguals [AoA: *Estimate* = 0.65, *SE* = 0.15, $t_{(163)} = 4.30$, $p < 0.001$; Family: *Estimate* = -0.55 , *SE* = 0.19,

TABLE 4 Effects of relative language experience (English—HL) on self-reported heritage language and English proficiency.

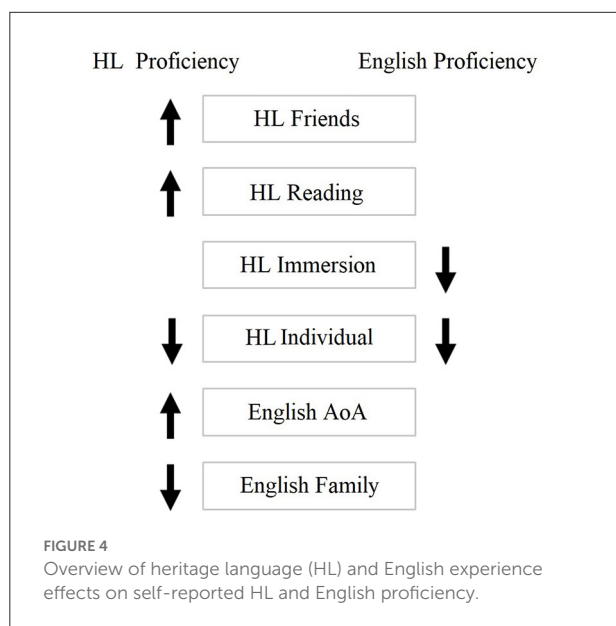
	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
(Intercept)	8.71	0.07	109	121.64	<0.001	***
Language	1.67	0.06	327	27.14	<0.001	***
Heritage	−0.45	0.14	109	−3.12	0.002	**
Measure	−0.05	0.06	327	−0.84	0.400	
AoA	0.13	0.08	109	1.60	0.113	
Immersion	0.16	0.09	109	1.81	0.072	~
Family	−0.16	0.08	109	−2.05	0.043	*
Friends	−0.01	0.09	109	−0.11	0.913	
Media	−0.04	0.08	109	−0.47	0.637	
Reading	−0.05	0.09	109	−0.60	0.547	
Individual	−0.03	0.08	109	−0.36	0.717	
Language:Heritage	0.65	0.12	327	5.22	<0.001	***
Language:Measure	0.29	0.12	327	2.32	0.021	*
Heritage:Measure	−0.35	0.12	327	−2.79	0.006	**
Language:AoA	−0.22	0.07	327	−3.01	0.003	**
Heritage:AoA	0.37	0.17	109	2.21	0.030	*
Measure:AoA	0.06	0.07	327	0.87	0.384	
Language:Immersion	0.12	0.08	327	1.61	0.108	
Heritage:Immersion	0.55	0.19	109	2.96	0.004	**
Measure:Immersion	−0.06	0.08	327	−0.81	0.416	
Language:Family	0.22	0.07	327	3.33	0.001	**
Heritage:Family	−0.23	0.16	109	−1.48	0.141	
Measure:Family	0.11	0.07	327	1.63	0.104	
Language:Friends	−0.03	0.07	327	−0.39	0.693	
Heritage:Friends	0.33	0.18	109	1.88	0.062	~
Measure:Friends	0.04	0.07	327	0.48	0.630	
Language:Media	0.20	0.07	327	2.82	0.005	**
Heritage:Media	−0.47	0.17	109	−2.76	0.007	**
Measure:Media	−0.04	0.07	327	−0.54	0.587	
Language:Reading	0.35	0.07	327	4.82	<0.001	***
Heritage:Reading	−0.44	0.17	109	−2.62	0.010	*
Measure:Reading	−0.04	0.07	327	−0.59	0.553	
Language:Individual	0.17	0.07	327	2.34	0.020	*
Heritage:Individual	0.18	0.18	109	1.00	0.317	
Measure:Individual	−0.10	0.07	327	−1.41	0.159	
Language:Heritage:Measure	0.54	0.25	327	2.16	0.031	*
Language:Heritage:AoA	−0.52	0.14	327	−3.60	<0.001	***
Language:Measure:AoA	0.00	0.14	327	0.03	0.975	
Heritage:Measure:AoA	0.07	0.14	327	0.49	0.621	
Language:Heritage:Immersion	−0.39	0.16	327	−2.41	0.016	*
Language:Measure:Immersion	0.23	0.16	327	1.48	0.140	
Heritage:Measure:Immersion	−0.27	0.16	327	−1.69	0.091	
Language:Heritage:Family	0.37	0.13	327	2.74	0.007	**
Language:Measure:Family	−0.12	0.13	327	−0.89	0.375	
Heritage:Measure:Family	−0.02	0.13	327	−0.15	0.880	

(Continued)

TABLE 4 (Continued)

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
Language:Heritage:Friends	−0.24	0.15	327	−1.61	0.109	
Language:Measure:Friends	−0.11	0.15	327	−0.73	0.465	
Heritage:Measure:Friends	0.06	0.15	327	0.42	0.673	
Language:Heritage:Media	0.39	0.15	327	2.68	0.008	**
Language:Measure:Media	0.03	0.15	327	0.21	0.831	
Heritage:Measure:Media	−0.08	0.15	327	−0.53	0.597	
Language:Heritage:Reading	0.52	0.15	327	3.56	<0.001	***
Language:Measure:Reading	0.09	0.15	327	0.59	0.557	
Heritage:Measure:Reading	−0.09	0.15	327	−0.59	0.555	
Language:Heritage:Individual	0.30	0.16	327	1.93	0.054	~
Language:Measure:Individual	0.11	0.15	327	0.73	0.464	
Heritage:Measure:Individual	−0.09	0.16	327	−0.56	0.579	
Language:Heritage:Measure:AoA	−0.36	0.29	327	−1.26	0.207	
Language:Heritage:Measure:Immersion	0.26	0.32	327	0.82	0.412	
Language:Heritage:Measure:Family	0.02	0.27	327	0.09	0.932	
Language:Heritage:Measure:Friends	−0.20	0.30	327	−0.66	0.512	
Language:Heritage:Measure:Media	0.15	0.29	327	0.53	0.600	
Language:Heritage:Measure:Reading	0.26	0.29	327	0.88	0.378	
Language:Heritage:Measure:Individual	0.10	0.31	327	0.33	0.743	

~ $p < 0.08$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.



$t_{(163)} = -2.97$, $p = 0.003$] than Spanish bilinguals [AoA: $Estimate = 0.07$, $SE = 0.13$, $t_{(163)} = 0.50$, $p = 0.62$; Family: $Estimate = -0.13$, $SE = 0.12$, $t_{(163)} = -1.02$, $p = 0.311$; see Figures 2A,B, respectively].

Effects of relative language experience (English—HL)

A significant three-way interaction between Language, Heritage Group, and **relative immersion** ($p = 0.016$) revealed that among non-Spanish bilinguals, relatively longer English (vs. HL) immersion predicted higher self-reported proficiency in both the HL [$Estimate = 0.53$, $SE = 0.17$, $t_{(151)} = 3.16$, $p = 0.002$] and in English [$Estimate = 0.43$, $SE = 0.17$, $t_{(151)} = 2.58$, $p = 0.011$]. Among Spanish bilinguals, relatively longer English immersion was associated with marginally lower HL proficiency [$Estimate = -0.22$, $SE = 0.12$, $t_{(151)} = -1.90$, $p = 0.059$], with no effect on English proficiency ($p = 0.549$; see Figure 3A and Table 4 for full output).

A series of three-way interactions additionally emerged for **relative age of acquisition** ($p < 0.001$) and the composite measures for relative acquisition and exposure through **family** ($p = 0.007$), **media** ($p = 0.008$), and **reading** ($p < 0.001$). Among non-Spanish bilinguals, HL proficiency was negatively predicted by more similar ages of HL and English acquisition [$Estimate = 0.60$, $SE = 0.14$, $t_{(151)} = 4.41$, $p < 0.001$] and relatively higher ratings of English (vs. HL) acquisition and exposure experience through family [$Estimate = -0.50$, $SE = 0.13$, $t_{(151)} = -3.85$, $p < 0.002$], media [$Estimate = -0.52$, $SE = 0.14$, $t_{(151)} = -3.76$, $p < 0.002$], and reading [$Estimate = -0.63$, $SE = 0.13$, $t_{(151)} = -4.82$, $p < 0.001$]. Relative AoA, manner of acquisition, and exposure did not predict HL proficiency

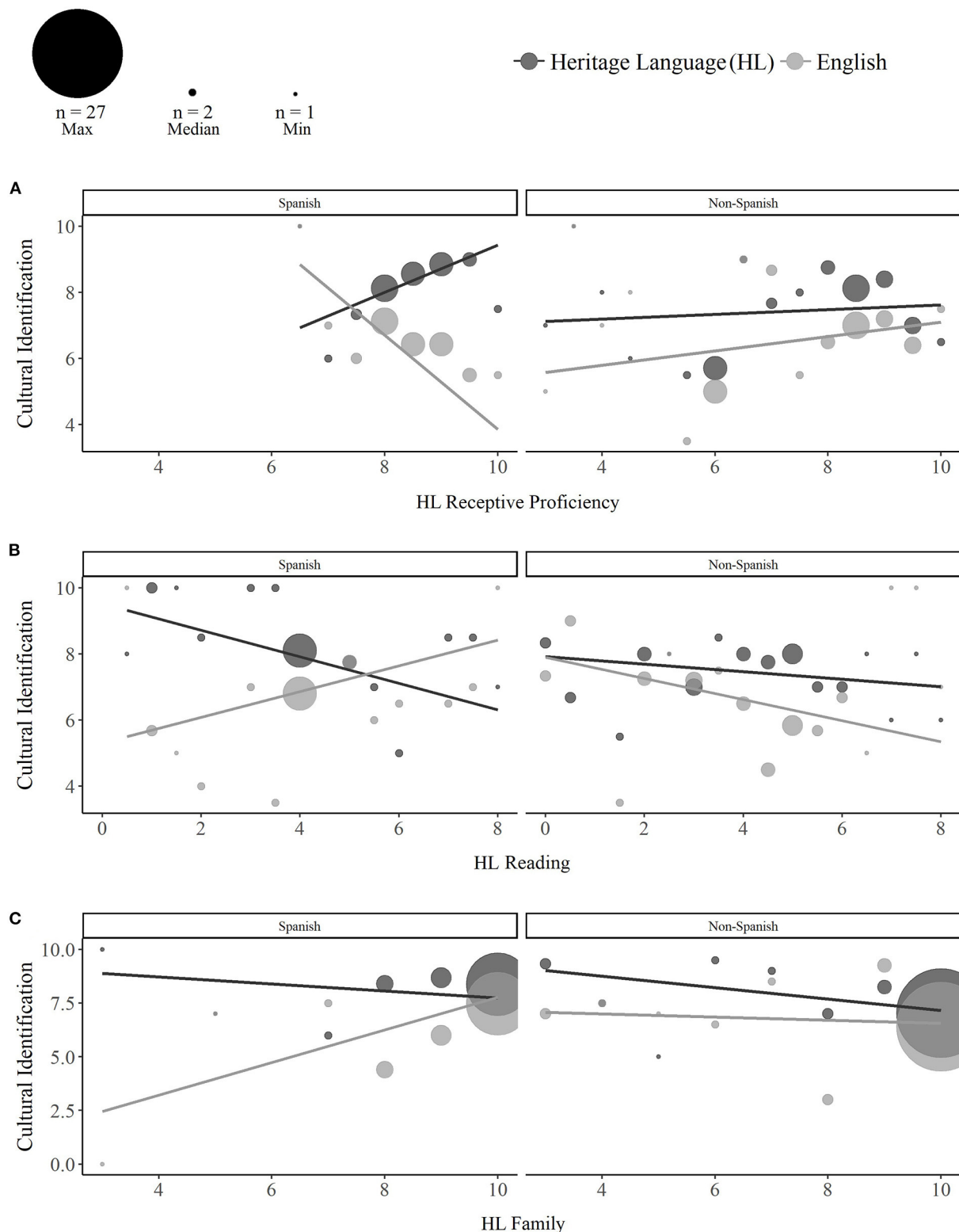


FIGURE 5

Effects of heritage language (HL) experience on cultural identification with the HL (dark gray) and English (light gray). Among Spanish bilinguals, cultural identification with English increased with lower HL receptive proficiency (A), as well as greater HL acquisition and exposure through reading (B) and family (C). Dot sizes reflect the number of participants contributing to each aggregated value (max = 27, median = 2, min = 1).

TABLE 5 Effects of heritage language experience on heritage language and English cultural identification.

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
(Intercept)	7.20	0.22	58	32.38	<0.001	***
Language	−0.96	0.44	58	−2.16	0.035	*
Heritage group	−0.33	0.46	58	−0.71	0.480	
Expressive proficiency	0.55	0.28	58	1.95	0.056	~
Receptive proficiency	−0.11	0.34	58	−0.33	0.746	
AoA	0.24	0.23	58	1.04	0.303	
Immersion	−0.28	0.21	58	−1.32	0.191	
Family	0.07	0.22	58	0.31	0.758	
Friends	0.34	0.24	58	1.42	0.161	
Media	−0.04	0.23	58	−0.17	0.862	
Reading	−0.25	0.25	58	−1.00	0.321	
Individual	0.28	0.26	58	1.08	0.284	
Language:Heritage	0.27	0.92	58	0.30	0.765	
Language:Expressive	0.47	0.56	58	0.83	0.410	
Heritage:Expressive	−0.17	0.60	58	−0.28	0.783	
Language:Receptive	−1.29	0.68	58	−1.88	0.065	~
Heritage:Receptive	0.76	0.73	58	1.04	0.302	
Language:AoA	0.13	0.47	58	0.27	0.786	
Heritage:AoA	1.27	0.45	58	2.83	0.006	**
Language:Immersion	−0.12	0.42	58	−0.29	0.777	
Heritage:Immersion	−0.36	0.42	58	−0.85	0.400	
Language:Family	1.02	0.44	58	2.34	0.023	*
Heritage:Family	−0.94	0.47	58	−2.01	0.050	~
Language:Friends	−0.09	0.47	58	−0.19	0.854	
Heritage:Friends	0.59	0.50	58	1.19	0.238	
Language:Media	−0.48	0.46	58	−1.04	0.302	
Heritage:Media	−0.65	0.47	58	−1.39	0.169	
Language:Reading	0.47	0.51	58	0.92	0.361	
Heritage:Reading	−0.43	0.49	58	−0.87	0.388	
Language:Individual	−0.01	0.52	58	−0.02	0.988	
Heritage:Individual	−0.13	0.51	58	−0.25	0.802	
Language:Heritage:Expressive	−0.43	1.19	58	−0.37	0.716	
Language:Heritage:Receptive	3.46	1.45	58	2.39	0.020	*
Language:Heritage:AoA	1.40	0.89	58	1.56	0.124	
Language:Heritage:Immersion	−1.31	0.84	58	−1.57	0.122	
Language:Heritage:Family	−1.45	0.93	58	−1.57	0.123	
Language:Heritage:Friends	0.71	0.99	58	0.72	0.477	
Language:Heritage:Media	−0.34	0.93	58	−0.37	0.715	
Language:Heritage:Reading	−2.03	0.98	58	−2.08	0.042	*
Language:Heritage:Individual	1.76	1.01	58	1.74	0.086	

~ $p < 0.08$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

for Spanish bilinguals ($ps > 0.242$) or English proficiency for either group ($ps > 0.118$). Finally, a significant two-way interaction between Language and relative individual experience [$Estimate = 0.17$, $SE = 0.07$, $t_{(327)} = 2.34$, $p = 0.020$] indicated that greater English (vs. HL) acquisition

and exposure in individual contexts was (non-significantly) associated with lower HL proficiency [$Estimate = -0.11$, $SE = 0.10$, $t_{(151)} = -1.15$, $p = 0.252$] and higher English proficiency [$Estimate = 0.08$, $SE = 0.10$, $t_{(151)} = 0.79$, $p = 0.438$; see Figures 3B–F].

TABLE 6 Effects of relative language experience (English—HL) on heritage language and English cultural identification.

	<i>Estimate</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>	
(Intercept)	7.31	0.22	116	33.05	<0.001	***
Language	−0.86	0.44	116	−1.95	0.053	~
Heritage	−0.41	0.46	116	−0.90	0.369	
Expressive proficiency	−0.29	0.31	116	−0.94	0.352	
Receptive proficiency	0.07	0.40	116	0.17	0.867	
AoA	0.12	0.24	116	0.51	0.611	
Immersion	0.64	0.27	116	2.39	0.019	*
Family	0.03	0.25	116	0.14	0.888	
Friends	−0.24	0.25	116	−0.93	0.352	
Media	0.06	0.25	116	0.23	0.815	
Reading	0.21	0.24	116	0.89	0.376	
Individual	−0.10	0.23	116	−0.45	0.657	
Language:Heritage	0.24	0.92	116	0.26	0.796	
Language:Expressive	−0.22	0.61	116	−0.36	0.717	
Heritage:Expressive	0.12	0.63	116	0.18	0.856	
Language:Receptive	1.15	0.79	116	1.45	0.151	
Heritage:Receptive	−1.15	0.84	116	−1.37	0.175	
Language:AoA	0.12	0.47	116	0.26	0.794	
Heritage:AoA	0.11	0.48	116	0.24	0.811	
Language:Immersion	0.46	0.53	116	0.86	0.393	
Heritage:Immersion	0.77	0.55	116	1.41	0.163	
Language:Family	−0.69	0.49	116	−1.40	0.165	
Heritage:Family	1.20	0.51	116	2.34	0.021	*
Language:Friends	0.21	0.51	116	0.42	0.678	
Heritage:Friends	−0.39	0.53	116	−0.72	0.470	
Language:Media	0.24	0.50	116	0.49	0.627	
Heritage:Media	0.58	0.50	116	1.15	0.252	
Language:Reading	−0.31	0.47	116	−0.66	0.509	
Heritage:Reading	0.18	0.47	116	0.37	0.709	
Language:Individual	0.32	0.46	116	0.70	0.486	
Heritage:Individual	−0.24	0.46	116	−0.53	0.598	
Language:Heritage:Expressive	0.79	1.27	116	0.62	0.534	
Language:Heritage:Receptive	−3.50	1.69	116	−2.07	0.041	*
Language:Heritage:AoA	−0.65	0.96	116	−0.68	0.499	
Language:Heritage:Immersion	0.87	1.10	116	0.79	0.430	
Language:Heritage:Family	0.80	1.02	116	0.78	0.435	
Language:Heritage:Friends	−1.35	1.07	116	−1.27	0.208	
Language:Heritage:Media	0.39	1.00	116	0.39	0.697	
Language:Heritage:Reading	1.59	0.94	116	1.69	0.093	
Language:Heritage:Individual	−0.33	0.92	116	−0.36	0.720	

~p < 0.08, *p < 0.05, ***p < 0.001.

In sum, higher HL proficiency was predicted by greater HL experience through reading and friends, later absolute and relative ages of English acquisition, less absolute and relative English experience through family, and less relative English experience through reading and media.

HL reading experience had a greater impact on HL receptive proficiency (understanding/reading) compared to expressive proficiency (speaking). English proficiency declined with longer durations of HL immersion, and proficiency in both languages increased with longer

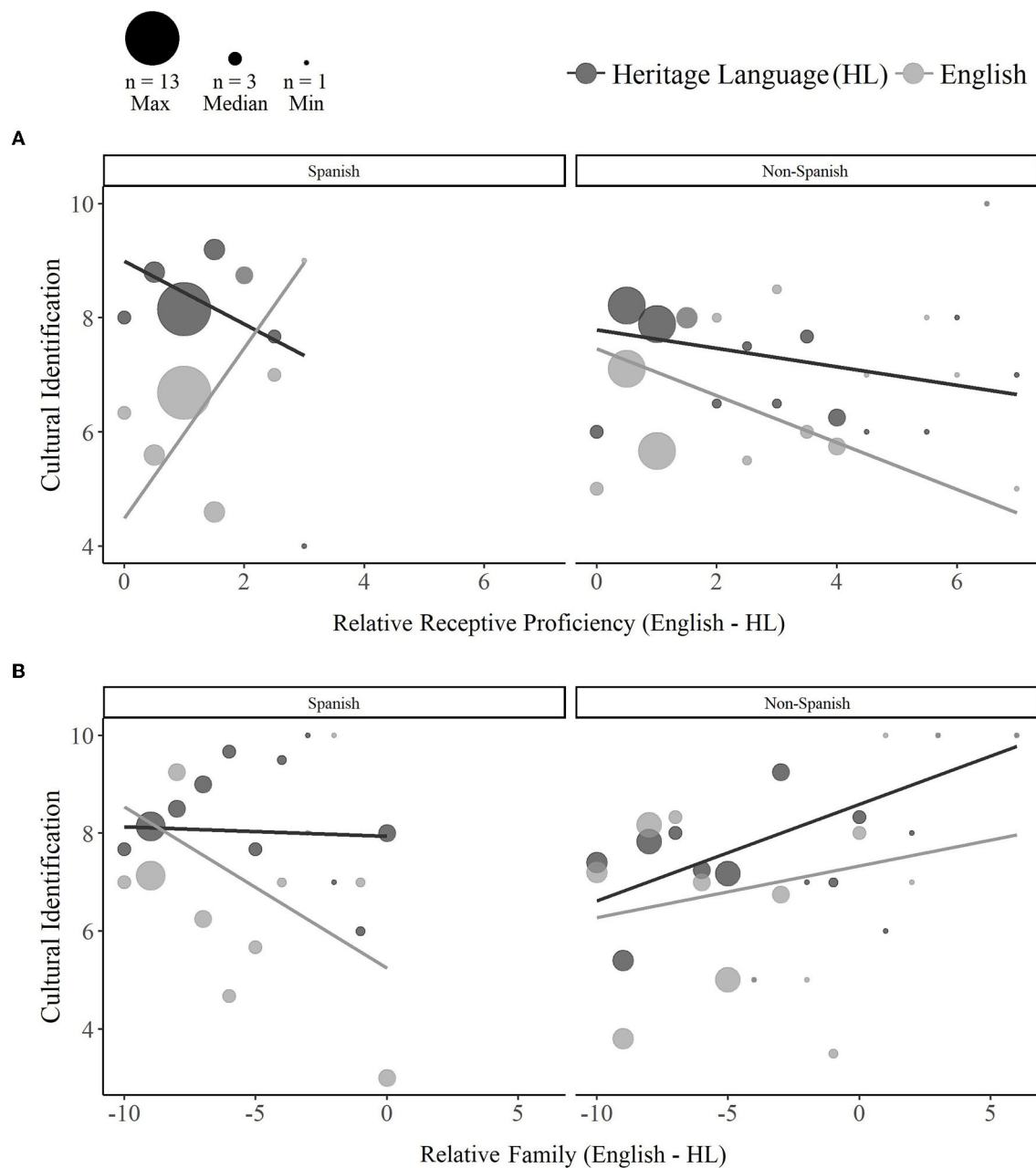


FIGURE 6

Effects of relative language experience (English—HL) on cultural identification with the HL (dark gray) and English (light gray). Among Spanish bilinguals, cultural identification with English increased with higher relative English (vs. HL) receptive proficiency (A). Relatively higher ratings of English (vs. HL) acquisition and exposure through family was associated with lower English identification among Spanish bilinguals and marginally higher HL identification among non-Spanish bilinguals (B). Dot sizes reflect the number of participants contributing to each aggregated value (max = 13, median = 3, min = 1).

relative durations of English (vs. HL) language immersion. Proficiency in both languages declined with greater HL experience in individual contexts (see Figure 4). Notably, effects of both HL and English experience were generally more robust among non-Spanish compared to Spanish bilinguals.

Predictors of heritage language and English cultural identification

Effects of heritage language experience

Cultural identification was significantly higher with the heritage language ($M = 7.71$, 95% CI [7.07, 8.35]) than English

($M = 6.74$, 95% CI [6.09, 7.38]), $p = 0.035$. See Table 5 for full output.

A three-way interaction between Language, Heritage Group, and self-reported Receptive Proficiency ($p = 0.020$) revealed that among Spanish bilinguals, cultural identification with English (but not the HL) declined with higher **receptive proficiency** in the HL [$Estimate = -2.16$, $SE = 0.89$, $t_{(116)} = -2.43$, $p = 0.017$; Figure 5A]. A three-way interaction with Reading ($p = 0.042$) revealed that among Spanish bilinguals, higher ratings of HL acquisition and exposure through **reading** [$Estimate = 0.79$, $SE = 0.43$, $t_{(116)} = 1.83$, $p = 0.062$] were associated with greater cultural identification with English, but lower cultural identification with the HL [$Estimate = -0.82$, $SE = 0.43$, $t_{(118)} = -1.89$, $p = 0.071$; Figure 5B]. Cultural identification with English was not moderated by receptive HL proficiency ($p = 0.834$) or reading experience ($p = 0.669$) among non-Spanish bilinguals.

Finally, the composite measure of **family** HL acquisition and exposure was unexpectedly associated with increased cultural identification with English [$Estimate = 0.69$, $SE = 0.33$, $t_{(116)} = 2.08$, $p = 0.040$], but not the HL ($p = 0.195$; Language \times Family: $p = 0.023$). Although the three-way interaction with Heritage Group was not significant ($p = 0.123$), simple effects revealed that the effect of HL family experience on English identification was driven by Spanish [$Estimate = 1.52$, $SE = 0.57$, $t_{(116)} = 2.65$, $p = 0.009$] rather than non-Spanish bilinguals [$Estimate = -0.15$, $SE = 0.46$, $t_{(116)} = -0.45$, $p = 0.65$; see Figure 5C]. No significant effects of English experience were observed for cultural identification (see Supplementary Table 1 for full output).

Effects of relative language experience (English—HL)

A significant main effect of relative **immersion** indicated that relatively longer durations of English (vs. HL) immersion predicted greater cultural identification with both languages ($p = 0.019$; see Table 6 for full output). Consistent with the effect of self-reported HL receptive proficiency, a three-way interaction between Language, Heritage Group, and relative self-reported **receptive proficiency** ($p = 0.041$) indicated that among Spanish bilinguals, cultural identification with English (but not the HL) increased with greater relative English (vs. HL) receptive proficiency [$Estimate = 2.28$, $SE = 1.04$, $t_{(116)} = 2.18$, $p = 0.031$; see Figure 6A]. Relative proficiency did not moderate cultural identification with either language for non-Spanish bilinguals ($ps > 0.42$). A significant interaction between Heritage Group and relative **family** acquisition and exposure ($p = 0.021$) indicated that relative English (vs. HL) family experience was a (non-significant) negative predictor of overall cultural identification among Spanish bilinguals [$Estimate = -0.64$, $SE = 0.42$, $t_{(58)} = -1.52$, $p = 0.135$] and a marginally positive predictor of overall identification

among non-Spanish bilinguals [$Estimate = 0.56$, $SE = 0.29$, $t_{(58)} = 1.93$, $p = 0.058$]. Although the three-way interaction with Language did not approach significance ($p = 0.435$), relatively greater English (vs. HL) family experience was associated with significantly lower cultural identification with English among Spanish bilinguals [$Estimate = -1.21$, $SE = 0.6$, $t_{(116)} = -2.03$, $p = 0.045$] and marginally greater cultural identification with the HL among non-Spanish bilinguals [$Estimate = 0.73$, $SE = 0.41$, $t_{(116)} = 1.78$, $p = 0.08$; see Figure 6B].

Discussion

The goal of the present study was to uncover linguistic predictors of self-reported language proficiency and cultural identification among different groups of adult heritage bilinguals. Self-reported proficiency in the majority language (English) was best predicted by the duration of immersion in the heritage language (HL). As expected, a longer cumulative duration of immersion in a country or school and/or work environment in which the HL was spoken was associated with lower reported English proficiency. Higher reported HL proficiency was predicted by higher ratings of HL acquisition and use through reading and friends, lower ratings of English acquisition and use through family, and later ages of English acquisition. Proficiency in both languages declined with greater HL experience in individual contexts (e.g., acquisition and exposure through self-instruction, language tapes, and language labs). Finally, despite higher self-reported English proficiency, cultural identification was higher with the HL, and this was especially true for Spanish heritage bilinguals. English cultural identification was negatively associated with subjective HL receptive proficiency, and to a lesser extent, positively associated with greater reliance on reading and family for HL acquisition and use. In addition to characterizing the factors that promote language proficiency and cultural identification, a critical finding from the present investigation is that the impact of heritage language and English language experience varied depending on heritage speakers' native languages.

Heritage group and self-reported language proficiency

First, we found that greater reliance on reading for HL acquisition and exposure predicted higher self-reported HL receptive proficiency among non-Spanish, but not Spanish bilinguals. One probable explanation for this finding is that the two groups differed in how much they could rely on English reading skills to support literacy in the HL. Unlike Spanish-English bilinguals, the non-Spanish bilinguals' heritage languages (Chinese, Hebrew, Korean, Russian, Tamil, and Thai) all utilized a different script from English, which may have

reduced the amount of cross-linguistic transfer in literacy (Huang and Hanley, 1995; Durgunoglu, 2002; Lindsey et al., 2003; Bialystok et al., 2005a,b; Luk, 2005) and other academic skills (Van der Slik, 2010; Zhang, 2013; Kostecká et al., 2015; Siu and Ho, 2015; see Koda, 2005; Genesee et al., 2006 for reviews). For instance, Bialystok et al. (2005a) observed that same-script bilinguals transferred literacy skills across languages, while different-script bilinguals did not. Because Spanish and English utilize the same script, the ability to comprehend written text in Spanish may be supported by English reading skills even without extensive exposure to Spanish text. In contrast, for different-script bilinguals, the ability to comprehend written text in the heritage language may be more contingent on dedicated exposure to HL text through reading. Consistent with this interpretation, non-Spanish bilinguals with minimal HL reading experience had significantly lower reading proficiency than matched Spanish bilinguals (-1 SD; $M_s = 4.33$ and 7.30 , respectively; $p < 0.001$). This gap closed among non-Spanish and Spanish bilinguals with greater HL reading experience ($+1$ SD; $M_s = 8.67$ and 7.80 , respectively; $p = 0.070$). This finding suggests that HL experience through reading may be particularly important for different-script bilinguals.

We additionally found that earlier ages of English acquisition and a more substantial role of family for English acquisition and use predicted lower self-reported HL proficiency among non-Spanish, but not Spanish bilinguals. Because bilinguals need to split their time between their two languages, time spent using one language leads to decreased use of the other language (Meir and Janssen, 2021). HL proficiency often declines with greater majority language use and less HL use (Jia and Aaronson, 2003; Gollan et al., 2015; Montrul, 2016; Vorobyeva and Bel, 2021). Our findings suggest that the negative impact of reduced HL use on HL proficiency may be minimized for speakers of more typologically similar languages, potentially because reading and conversational skills acquired from the majority language can transfer to the HL. In addition to orthographic similarities, the degree of lexical and grammatical overlap between English and Spanish (two Indo-European languages) is likely greater than between English and non-Spanish languages (primarily non-Indo-European) included in the present study. Consequently, even if time spent using English detracts from time spent using the HL, Spanish-English bilinguals may be better able to benefit from positive linguistic transfer between languages (Odlin, 1989; Bialystok et al., 2003, 2005a; Melby-Lervåg and Lervåg, 2011).

Due to the high number of Spanish speakers in the United States, Spanish-English bilinguals may also be able to benefit from greater HL experience outside of the home even if English is used more frequently with family. Indeed, Spanish bilinguals in the present study did report significantly greater HL exposure through music/radio, reading, and individual instruction, as well as numerically greater HL exposure through

family and TV relative to non-Spanish bilinguals (see Table 1). Supplementary analyses provide preliminary support for such a compensatory mechanism, as the negative effects of both English AoA and family use on HL proficiency declined with greater overall HL experience (aggregated across contexts of acquisition and exposure; see Supplementary Table 2 for details). Together, these findings suggest that the extent to which majority language experience helps vs. hinders HL acquisition and maintenance is subject to variability in linguistic similarity across languages, as well as the amount of HL use across different contexts.

Heritage group and cultural identification

Among Spanish bilinguals, identification with English-speaking cultures (e.g., American) increased with lower self-reported HL receptive proficiency, as well as with greater HL experience through reading and family. Prior work has demonstrated a robust relationship between cultural identification and language proficiency, most often showing a positive association between cultural identification and proficiency within a given language (e.g., between HL proficiency and HL ethnic identity; Bankston and Zhou, 1995; Cho, 2000; Pease-Alvarez, 2002; Oh and Fuligni, 2010; Yu, 2015; Arredondo et al., 2016; Schroeder et al., 2017). Our findings indicate that proficiency in one language can be inversely related to cultural identification with the other, and that the relationship between language experience and cultural identification varies across different groups of heritage speakers.

More unexpected was our finding that cultural identification with English increased with higher ratings of Spanish acquisition and exposure through reading and family. Effects of relative proficiency further indicated that while identification with the two languages was comparably high among Spanish bilinguals with substantially greater Spanish (vs. English) family experience, identification with English became progressively lower with more balanced use of the two languages at home. Although speculative, some heritage speakers may develop stronger or weaker identification with each culture to compensate for imbalances in language use and immersion at home. Cheryan and Monin (2005) found that Asian Americans expressed greater American cultural identification when their American identities were threatened. Additionally, while bilinguals primed with a particular language or culture often exhibit culturally-congruent behaviors and judgments (i.e., assimilation), there are also cases in which bilinguals instead respond in culturally-incongruent ways, particularly if they perceive their cultural identities to be threatened or in conflict with one another (Benet-Martínez et al., 2002; Cheng et al., 2006; Zou et al., 2008). The fact that a positive association between English identification and HL experience was found for Spanish bilinguals, but not for the non-Spanish bilinguals may potentially stem from differences in

the extent to which the two groups perceive their cultural identities to be compatible vs. in conflict. Indeed, an exploratory examination of the relationship between English and HL identification within the two groups provides tentative support for this interpretation. Specifically, while there was a non-significant negative correlation between English and HL cultural identification among Spanish bilinguals ($r = -0.22$, $p = 0.206$), there was a marginal positive correlation between identification with the two languages among non-Spanish bilinguals ($r = 0.29$, $p = 0.059$). Similar group differences were observed by Gong (2007) who found that while there was no correlation between ethnic (minority culture) identity and national (majority culture) identity among African Americans, ethnic and national identity were positively correlated among American-born Chinese Americans. A possible avenue for future research may therefore be to examine whether different groups of heritage speakers vary in the perceived compatibility of their two cultures, and whether such differences moderate the impact of language experience on cultural identification within and across languages.

Future research may additionally examine the extent to which the observed effects and predictors of self-reported language proficiency are replicated using objective measures of language ability. The inclusion of objective measures assessing a variety of linguistic domains (e.g., lexicon, syntax, pronunciation) will contribute to determining the generalizability of the present findings and for characterizing the impact of heritage and majority language experience on different aspects of language proficiency. Second, our understanding of systematic variability across different heritage speakers would benefit from the inclusion of a greater number of participants from a more diverse range of language backgrounds. In particular, the roles of script and cross-linguistic transfer could be more fully elucidated through the inclusion of same- and different-script bilinguals within (e.g., German and Italian vs. German and Greek) and across (e.g., German and Vietnamese vs. German and Mandarin) language families. Likewise, interactions between language experience and culture could be examined more fully by crossing linguistic and cultural similarity.

In conclusion, the present findings reveal that the relationships between language experience, self-reported language proficiency, and cultural identification systematically vary as a function of heritage speakers' native languages. We additionally provide preliminary evidence to suggest that such differences may partly stem from variability in the degree of linguistic (e.g., orthographic overlap) and cultural (e.g., cultural compatibility) similarity across languages, as well as in opportunities for HL exposure outside of the home. Together, our results demonstrate the complex interplay between heritage and majority language experience, and highlight the need

to consider individual measures within the broader context of bilinguals' linguistic environments and history. Greater sensitivity to the needs and abilities of different types of bilinguals can promote the development of more effective heritage bilingual curricula, and provide a more nuanced understanding of heritage bilinguals' language acquisition and identity.

Data availability statement

The dataset analyzed for this study is available from the corresponding author upon reasonable request.

Ethics statement

The studies involving human participants were reviewed and approved by Northwestern University's Institutional Review Board. The participants provided their written informed consent to participate in this study.

Author contributions

SH, AC-F-Y, and VM contributed to the conception of the study, revised the manuscript, and approved the final version. SH and AC-F-Y organized the data and wrote the first draft of the manuscript. SH performed the analyses and visualization. VM supervised the study and provided funding. All authors contributed to the article and approved the submitted version.

Funding

This research was supported in part by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under Award Number R01HD059858 to VM.

Acknowledgments

We thank current and past members of the *Bilingualism and Psycholinguistics Research Group*, whose studies were used to identify heritage bilinguals for inclusion in the present dataset.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Albirini, A. (2014). Toward understanding the variability in the language proficiencies of Arabic heritage speakers. *Int. J. Bilingual.* 18, 730–765. doi: 10.1177/1367006912472404
- Armon-Lotem, S., Rose, K., and Altman, C. (2021). The development of English as a heritage language: the role of chronological age and age of onset of bilingualism. *First Lang.* 41, 67–89. doi: 10.1177/0142723720929810
- Arredondo, M. M., Rosado, M., and Satterfield, T. (2016). Understanding the impact of heritage language on ethnic identity formation and literacy for U.S. Latino children. *J. Cogn. Cult.* 16, 245–266. doi: 10.1163/15685373-12342179
- Au, T. K., and Romo, L. F. (1997). "Does childhood language experience help adult learners?" in *The Cognitive Processing of Chinese and Related Asian Languages*, Ed H.-C. Chen (Hong Kong: Chinese University Press), 417–441.
- Bankston, C. L. III, and Zhou, M. (1995). Effects of minority-language literacy on the academic achievement of Vietnamese youths in New Orleans. *Sociol. Educ.* 68, 1–17. doi: 10.2307/2112760
- Bartolotti, J., Marian, V., Schroeder, S. R., and Shook, A. (2011). Statistical learning of a Morse code language is improved by bilingualism and inhibitory control. *Proc. Cogn. Sci. Soc.* 33, 885–890. Available online at: <https://escholarship.org/uc/item/6rt4964b>
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*. doi: 10.18637/jss.v067.i01
- Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., et al. (2012). The measure matters: language dominance profiles across measures in Spanish-English bilingual children. *Bilingualism* 15, 616–629. doi: 10.1017/S1366728912000090
- Benet-Martínez, V., Leu, J., Lee, F., and Morris, M. W. (2002). Negotiating biculturalism: cultural frame switching in biculturals with oppositional versus compatible cultural identities. *J. Cross Cult. Psychol.* 33, 492–516. doi: 10.1177/0022022102033005005
- Bialystok, E., Luk, G., and Kwan, E. (2005a). Bilingualism, biliteracy, and learning to read: interactions among languages and writing systems. *Sci. Stud. Read.* 9, 43–61. doi: 10.1207/s1532799xssr0901_4
- Bialystok, E., Majumder, S., and Martin, M. M. (2003). Developing phonological awareness: is there a bilingual advantage? *Appl. Psycholinguist.* 24, 27–44. doi: 10.1017/S014271640300002X
- Bialystok, E., McBride-Chang, C., and Luk, G. (2005b). Bilingualism, language proficiency, and learning to read in two writing systems. *J. Educ. Psychol.* 97, 580–590. doi: 10.1037/0022-0663.97.4.580
- Carreira, M., and Kagan, O. (2011). The results of the National Heritage Language Survey: implications for teaching, curriculum design, and professional development. *Foreign Lang. Ann.* 44, 40–64. doi: 10.1111/j.1944-9720.2010.01118.x
- Chabal, S., Hayakawa, S., and Marian, V. (2022). Language is activated by visual input regardless of memory demands or capacity. *Cognition* 222:104994. doi: 10.1016/j.cognition.2021.104994
- Chabal, S., Schroeder, S. R., and Marian, V. (2015). Audio-visual object search is changed by bilingual experience. *Attent. Percept. Psychophys.* 77, 2684–2693. doi: 10.3758/s13414-015-0973-7
- Chen, P., Bobb, S. C., Hoshino, N., and Marian, V. (2017). Neural signatures of language co-activation and control in bilingual spoken word comprehension. *Brain Res.* 1665, 50–64. doi: 10.1016/j.brainres.2017.03.023
- Chen, S. H., Zhou, Q., and Uchikoshi, Y. (2018). Heritage language socialization in Chinese American immigrant families: prospective links to children's heritage language proficiency. *Int. J. Biling. Educ. Biling.* 24, 1193–1209. doi: 10.1080/13670050.2018.1547680
- Chen, S. X., Benet-Martínez, V., and Bond, M. (2008). Bicultural identity, bilingualism, and psychological adjustment in multicultural societies: immigration-based and globalization-based acculturation. *J. Pers.* 76, 803–838. doi: 10.1111/j.1467-6494.2008.00505.x
- Cheng, C. Y., Lee, F., and Benet-Martínez, V. (2006). Assimilation and contrast effects in cultural frame switching: bicultural identity integration and valence of cultural cues. *J. Cross Cult. Psychol.* 37, 742–760. doi: 10.1177/0022022106292081
- Cheryan, S., and Monin, B. (2005). Where are you really from?: Asian Americans and identity denial. *J. Pers. Soc. Psychol.* 89, 717–730. doi: 10.1037/0022-3514.89.5.717
- Cho, G. (2000). The role of heritage language in social interactions and relationships: reflections from a language minority group. *Biling. Res. J.* 24, 369–384. doi: 10.1080/15235882.2000.10162773
- Daskalaki, E., Chondrogianni, V., Blom, E., Argyri, F., and Paradis, J. (2019). Input effects across domains: the case of Greek subjects in child heritage language. *Sec. Lang. Res.* 35, 421–445. doi: 10.1177/0267658318787231
- de Leeuw, E., Schmid, M. S., and Mennen, I. (2010). The effects of contact on native language pronunciation in an L2 migrant setting. *Biling. Lang. Cogn.* 13, 33–40. doi: 10.1017/S1366728909990289
- Durgunoglu, A. Y. (2002). Cross-linguistic transfer in literacy development and implications for language learners. *Ann. Dyslexia* 52, 189–204. doi: 10.1007/s11881-002-0012-y
- Felix-Ortiz, M., Newcomb, M. D., and Myers, H. (1994). A multidimensional measure of cultural identity for Latino and Latina adolescents. *Hispanic J. Behav. Sci.* 16, 99–115. doi: 10.1177/07399863940162001
- Freeman, M. R., Blumenfeld, H. K., Carlson, M. T., and Marian, V. (2022). First-language influence on second language speech perception depends on task demands. *Lang. Speech* 65, 28–51. doi: 10.1177/0023830920983368
- Freeman, M. R., Blumenfeld, H. K., and Marian, V. (2016). Phonotactic constraints are activated across languages in bilinguals. *Front. Psychol.* 7:702. doi: 10.3389/fpsyg.2016.00702
- Gatbonton, E., and Trofimovich, P. (2008). The ethnic group affiliation and L2 proficiency link: empirical evidence. *Lang. Aware* 17, 229–248. doi: 10.1080/09658410802146867
- Gathercole, V., and Thomas, E. (2009). Bilingual first-language development: dominant language takeover, threatened minority language take-up. *Biling. Lang. Cogn.* 12, 213–237. doi: 10.1017/S1366728909004015

Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomm.2022.994709/full#supplementary-material>

- Genesee, F., Geva, E., Dressler, C., and Kamil, M. (2006). "Synthesis: cross-linguistic relationships," in *Developing Literacy in Second-Language Learners: Report of the National Literacy Panel on Language-Minority Children and Youth*, Eds D. August and T. Shanahan (New York, NY: Lawrence Erlbaum Associates Publishers), 153–174.
- Gharibi, K., and Boers, F. (2017). Influential factors in incomplete acquisition and attrition of young heritage speakers' vocabulary knowledge. *Lang. Acquis.* 24, 52–69. doi: 10.1080/10489223.2016.1187613
- Giguere, D., and Hoff, E. (2020). Home language and societal language skills in second-generation bilingual adults. *Int. J. Bilingual.* 24, 1071–1087. doi: 10.1177/1367006920932221
- Gollan, T. H., Starr, J., and Ferreira, V. S. (2015). More than use it or lose it: the number-of-speakers effect on heritage language proficiency. *Psychon. Bull. Rev.* 22, 147–155. doi: 10.3758/s13423-014-0649-7
- Gollan, T. H., Weissberger, G., Runnqvist, E., Montoya, R. I., and Cera, C. M. (2012). Self-ratings of spoken language dominance: a multi-lingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Biling. Lang. Cogn.* 15, 594–615. doi: 10.1017/S1366728911000332
- Gong, L. (2007). Ethnic identity and identification with the majority group: relations with national identity and self-esteem. *Int. J. Intericult. Relat.* 31, 503–523. doi: 10.1016/j.ijintrel.2007.03.002
- Hakuta, K., and D'Andrea, D. (1992). Some properties of bilingual maintenance and loss in Mexican background high-school students. *Appl. Linguist.* 13, 72–99. doi: 10.1093/applin/13.1.72
- Hayakawa, S., Shook, A., and Marian, V. (2020). When it's harder to ignore than to ignore: evidence of greater attentional capture from a non-dominant language. *Int. J. Bilingual.* 24, 999–1016. doi: 10.1177/1367006920915277
- Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., and Parra, M. (2012). Dual language exposure and early bilingual development. *J. Child Lang.* 39, 1–27. doi: 10.1017/S0305000910000759
- Hovsepian, A. (2018). Vocabulary growth in Armenian-English bilingual kindergarteners. *First Lang.* 38, 47–71. doi: 10.1177/0142723717715985
- Huang, H. S., and Hanley, J. R. (1995). Phonological awareness and visual skills in learning to read Chinese and English. *Cognition* 54, 73–98. doi: 10.1016/0010-0277(94)00641-W
- Hur, E., Lopez Otero, J. C., and Lee, E. (2021). Attitudes and expectations towards heritage language instruction: evidence from Korean and Spanish in the US. *Languages* 6:14. doi: 10.3390/languages6010014
- Jia, G., and Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Appl. Psycholinguist.* 24, 131–161. doi: 10.1017/S0142716403000079
- Jia, R., and Paradis, J. (2015). The use of referring expressions in narratives by Mandarin heritage language children and the role of language environment factors in predicting individual differences. *Biling. Lang. Cogn.* 18, 737–752. doi: 10.1017/S1366728914000728
- Karayayla, T., and Schmid, M. S. (2019). First language attrition as a function of age at onset of bilingualism: first language attainment of Turkish-English bilinguals in the United Kingdom. *Lang. Learn.* 69, 106–142. doi: 10.1111/lang.12316
- Koda, K. (2005). "Learning to read across writing systems: transfer, metalinguistic awareness, and second language reading development," in *Second Language Writing Systems*, Eds V. Cook and B. Bassetti (Tonawanda, NY: Multilingual Matters), 311–314.
- Kostecká, Y., Kostecký, T., Vodičková, K., and Jančarik, A. (2015). Linguistic integration of middle school immigrant children in Czechia. *AUC Geogr.* 50, 181–192. doi: 10.14712/23361980.2015.97
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Bilingual.* 22, 564–582. doi: 10.1177/1367006916654355
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2018). Emmeans: estimated marginal means, aka least-squares means. R package version. 1.3. Available online at: <https://cran.r-project.org/web/packages/emmeans/emmeans.pdf>
- Lindsey, K. A., Manis, F. R., and Bailey, C. E. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *J. Educ. Psychol.* 95, 482–494. doi: 10.1037/0022-0663.95.3.482
- Luk, G. (2005). *Exploring the latent factors behind inter-language correlations in reading and phonological awareness* [Doctoral dissertation] Retrieval from ProQuest Dissertation (Accession No. 0612993523). University: York University, Toronto, Canada.
- Makarova, V., Terekhova, N., and Mousavi, A. (2019). Children's language exposure and parental language attitudes in Russian-as-a-heritage-language acquisition by bilingual and multilingual children in Canada. *Int. J. Bilingual.* 23, 457–485. doi: 10.1177/1367006917740058
- Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007)067
- Marian, V., Hayakawa, S., Lam, T. Q., and Schroeder, S. R. (2018). Language experience changes audiovisual perception. *Brain Sci.* 8:85. doi: 10.3390/brainsci8050085
- Marian, V., Hayakawa, S., and Schroeder, S. R. (2021). Memory after visual search: overlapping phonology, shared meaning, and bilingual experience influence what we remember. *Brain Lang.* 222:105012. doi: 10.1016/j.bandl.2021.105012
- McCarthy, K. M., and de Leeuw, E. (2022). Prosodic patterns in Sylheti-English bilinguals. *Stud. Sec. Lang. Acquisit.* 44, 562–579. doi: 10.1017/S027226312100036X
- Meir, N., and Janssen, B. (2021). Child heritage language development: an interplay between cross-linguistic influence and language-external factors. *Front. Psychol.* 12:651730. doi: 10.3389/fpsyg.2021.651730
- Meir, N., Walters, J., and Armon-Lotem, S. (2017). Bi-directional cross-linguistic influence in bilingual Russian-Hebrew children. *Linguist. Approach. Bilingual.* 7, 514–553. doi: 10.1075/lab.15007.mei
- Melby-Lervåg, and, M., and Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: a meta-analysis of the correlational evidence. *J. Res. Read.* 34, 114–135. doi: 10.1111/j.1467-9817.2010.01477.x
- Montrul, S. (2005). Second language acquisition and first language loss in adult early bilinguals: exploring some differences and similarities. *Sec. Lang. Res.* 21, 199–249. doi: 10.1191/0267658305sr247oa
- Montrul, S. (2008). *Incomplete Acquisition in Bilingualism: Re-examining the Age Factor. Studies in Bilingualism*, Vol. 39. Amsterdam: John Benjamins Publishing Company.
- Montrul, S. (2011). Interfaces and incomplete acquisition. *Lingua Spec. Issue Interfaces Lang. Acquisit.* 212, 591–604. doi: 10.1016/j.lingua.2010.05.006
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.
- Noels, K. A., Pon, G., and Clément, R. (1996). Language, identity, and adjustment: the role of linguistic self-confidence in the acculturation process. *J. Lang. Soc. Psychol.* 15, 246–264. doi: 10.1177/0261927X960153003
- Odlin, T. (1989). *Language Transfer*, Vol. 27. Cambridge: Cambridge University Press.
- Oh, J. S., and Fuligni, A. J. (2010). The role of heritage language development in the ethnic identity and family relationships of adolescents from immigrant backgrounds. *Soc. Dev.* 19, 202–220. doi: 10.1111/j.1467-9507.2008.00530.x
- Otwinowska, A., Meir, N., Ringblom, N., Karpava, S., and La Morgia, F. (2021). Language and literacy transmission in heritage language: evidence from Russian-speaking families in Cyprus, Ireland, Israel and Sweden. *J. Multiling. Multicult. Dev.* 42, 357–382. doi: 10.1080/01434632.2019.1695807
- Pease-Alvarez, L. (2002). Moving beyond linear trajectories of language shift and bilingual language socialization. *Hisp. J. Behav. Sci.* 24, 114–137. doi: 10.1177/0739986302024002002
- Phinney, J. S. (1990). Ethnic identity in adolescents and adults: review of research. *Psychol. Bull.* 108, 499–514.
- Phinney, J. S., Horenczyk, G., Liebkind, K., and Vedder, P. (2001). Ethnic identity, immigration, and well-being: an interactional perspective. *J. Soc. Issues* 57, 493–510. doi: 10.1111/0022-4537.00225
- Polinsky, M. (2015). Heritage languages and their speakers: State of the field, challenges, perspectives for future work, and methodologies. *Zeitschr. Fremdsprachwissenschaft* 26, 7–27. Available online at: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:37108756>
- Polinsky, M., and Kagan, O. (2007). Heritage languages: in the 'wild' and in the classroom. *Lang. Linguist. Compass.* 1, 368–395. doi: 10.1111/j.1749-818X.2007.00022.x
- Rodina, Y., Kupisch, T., Meir, N., Mitrofanova, N., Urek, O., and Westergaard, M. (2020). Internal and external factors in heritage language acquisition: evidence from heritage Russian in Israel, Germany, Norway, Latvia and the United Kingdom. *Front. Educ.* 5:20. doi: 10.3389/feduc.2020.00020

- Schmid, M. S., and Yilmaz, G. (2018). Predictors of language dominance: an integrated analysis of first language attrition and second language acquisition in late bilinguals. *Front. Psychol.* 9:1306. doi: 10.3389/fpsyg.2018.01306
- Schroeder, S. R., Lam, T. Q., and Marian, V. (2017). Linguistic predictors of cultural identification in bilinguals. *Appl. Linguist.* 38, 463–488. doi: 10.1093/applin/amv049
- Scontras, G., Fuchs, Z., and Polinsky, M. (2015). Heritage language and linguistic theory. *Front. Psychol.* 6:1545. doi: 10.3389/fpsyg.2015.01545
- Shook, A., and Marian, V. (2016). The influence of native-language tones on lexical access in the second language. *J. Acoust. Soc. Am.* 139, 3102–3109. doi: 10.1121/1.4953692
- Siu, C. T. S., and Ho, C. S. H. (2015). Cross-language transfer of syntactic skills and reading comprehension among young Cantonese–English bilingual students. *Read. Res. Q.* 50, 313–336. doi: 10.1002/rrq.101
- Stoehr, A., Benders, T., Van Hell, J. G., and Fikkert, P. (2017). Second language attainment and first language attrition: the case of VOT in immersed Dutch–German late bilinguals. *Sec. Lang. Res.* 33, 483–518. doi: 10.1177/0267658317704261
- Tao, L., Cai, Q., and Gollan, T. H. (2021). Effects of cumulative language exposure on heritage and majority language skills: Spanish and Mandarin heritage speakers in the USA. *Linguist. Approach. Bilingual.* 11, 168–191. doi: 10.1075/lab.18044.tao
- The Economist (2017). *Latinos Have Become Chicago's Second-Largest Ethnic Group*. Available online at: <https://www.economist.com/democracy-in-america/2017/10/05/latinos-have-become-chicagos-second-largest-ethnic-group> (accessed June 23, 2022).
- Thomas, E. M., Williams, N., Jones, L. A., Davies, S., and Binks, H. (2014). Acquiring complex structures under minority language conditions: bilingual acquisition of plural morphology in Welsh. *Biling. Lang. Cogn.* 17, 478–494. doi: 10.1017/S1366728913000497
- Torregrossa, J., Flores, C., and Rinke, E. (2022). What modulates the acquisition of difficult structures in a heritage language? A study on Portuguese in contact with French, German and Italian. *Biling. Lang. Cogn.* 1–14. doi: 10.1017/S1366728922000438
- U.S. Census Bureau (2020). *Language Spoken at Home (S1601)*. Available online at: <https://data.census.gov/cedsci/table?q=language&tid=ACST5Y2020.S1601>
- Unsworth, S. (2016). “Quantity and quality of language input in bilingual language development,” in *Lifespan Perspectives on Bilingualism*, Eds E. Nicoladis and S. Montanari (Berlin: Mouton de Gruyter, APA), 136–196.
- Valdés, G. (2000). “The teaching of heritage languages: An introduction for Slavic-teaching professionals,” in *The Learning and Teaching of Slavic Languages and Cultures*, Eds O. Kagan and B. Rifkin (Bloomington, IN: Slavica), 375–403.
- Van der Slik, F. W. (2010). Acquisition of Dutch as a second language: the explanative power of cognate and genetic linguistic distance measures for 11 West European first languages. *Stud. Sec. Lang. Acquisit.* 32, 401–432. doi: 10.1017/S0272263110000021
- Vorobyeva, T., and Bel, A. (2021). Factors affecting language proficiency in heritage language: the case of young Russian heritage speakers in Spain. *J. Lang. Contact.* 14, 304–330. doi: 10.1163/19552629-14020003
- Yu, S.-C. (2015). The relationships among heritage language proficiency, ethnic identity, and self-esteem. *Forum Int. Res. Educ.* 2, 57–71. doi: 10.18275/fire201502021039
- Zeigler, K., and Camarota, S. A. (2019). *67.3 Million in the United States Spoke a Foreign Language at Home in 2018*. Center for Immigration Studies. Available online at: https://cis.org/sites/default/files/2019-10/camarota-language-19_0.pdf
- Zhang, D. (2013). Linguistic distance effect on cross-linguistic transfer of morphological awareness. *Appl. Psycholinguist.* 34, 917–942. doi: 10.1017/S0142716412000070
- Zou, X., Morris, M. W., and Benet-Martínez, V. (2008). Identity motives and cultural priming: cultural (dis)identification in assimilative and contrastive responses. *J. Exp. Soc. Psychol.* 44, 1151–1159. doi: 10.1016/j.jesp.2008.02.001



OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway,
Norway

REVIEWED BY

Terje Lohndal,
Norwegian University of Science
and Technology, Norway
Seçkin Arslan,
Centre National de la Recherche
Scientifique (CNRS), France

*CORRESPONDENCE

Zuzanna Fuchs
zfuchs@usc.edu

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 02 June 2022

ACCEPTED 31 August 2022

PUBLISHED 30 September 2022

CITATION

Fuchs Z (2022) Eyetracking evidence
for heritage speakers' access
to abstract syntactic agreement
features in real-time processing.
Front. Psychol. 13:960376.
doi: 10.3389/fpsyg.2022.960376

COPYRIGHT

© 2022 Fuchs. This is an open-access
article distributed under the terms of
the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution
or reproduction in other forums is
permitted, provided the original
author(s) and the copyright owner(s)
are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does
not comply with these terms.

Eyetracking evidence for heritage speakers' access to abstract syntactic agreement features in real-time processing

Zuzanna Fuchs*

Department of Linguistics, University of Southern California, Los Angeles, CA, United States

This paper presents the results of an eyetracking study that uses the Visual World Paradigm to determine whether heritage speakers of Polish can use grammatical gender cues to facilitate lexical retrieval of the subsequent noun during real time processing. Previous work has investigated this question for heritage speakers of Spanish with gender cues located on definite articles, which are highly frequent in Spanish; the results are therefore consistent both with a grammatical account, wherein heritage speakers access abstract syntactic gender features during processing, and a probabilistic account, wherein facilitation is due to transition probabilities between frequently co-occurring elements. In Polish, gender cues appear on adjectives, which are optional and infrequent. Results of the present study show that heritage speakers of Polish can use gender on inflected adjectives to fixate on the target noun faster in trials where that gender cue uniquely identifies the target noun. This finding supports a grammatical rather than probabilistic account of the facilitative use of grammatical gender in this population: heritage speakers are able to access abstract syntactic information in real time to aid word recognition in a target-like manner.

KEYWORDS

heritage languages, grammatical gender, Polish, eyetracking, processing

Introduction

Heritage speakers (HSs) grow up speaking and hearing a minority language at home but ultimately become dominant in the majority language spoken by the community, with a clear shift in input and dominance around school-age, when children start spending significantly less time at home, where the heritage language is spoken (Valdés, 2000; Rothman, 2009; Montrul, 2016; Kupisch and Rothman, 2018; Polinsky, 2018). As an instance of unbalanced bilingualism, heritage languages are increasingly of interest to linguists for the questions they raise regarding the impact of reduced input on

the grammar and on language processing. Studies on HSs of a variety of languages have shown that in these conditions of reduced input to the heritage grammar, certain domains, such as morphosyntax, are more vulnerable and may show effects of attrition, transfer, or restructuring.

Within morphosyntax, grammatical gender has been shown to be particularly vulnerable to reduced input, with clear surface differences between heritage languages and their corresponding baseline languages, at least as evidenced by offline studies.¹ These differences occur both in gender assignment and in gender agreement: Heritage speakers have consistently been observed to assign nouns to gender categories differently than control speakers do, and to show non-target-like comprehension and/or production of gender agreement on articles, adjectives, and/or verbs (e.g., Hindi: Montrul et al., 2012; Russian: Polinsky, 2006, 2008; Hungarian: Bolonyai, 2007; Arabic: Albirini et al., 2011, 2013; Spanish: Montrul et al., 2014, 2008; Scontras et al., 2018; Swedish: Håkansson, 1995). In fact, evidence from divergent comprehension of gender agreement suggests that surface differences in gender agreement may even be a reflex of differences in the mental representation of grammatical gender in the heritage grammar as compared to the baseline grammar (Scontras et al., 2018).

Nevertheless, recent evidence from studies using online methodologies suggests that despite surface differences in production and comprehension of grammatical gender agreement, when one controls for knowledge of gender categorization, processing of gender by HSs may be qualitatively target-like. In an eyetracking study in the Visual World Paradigm, Fuchs (2021) found that HSs of Spanish were able to fixate on target items faster when a pre-nominal gender-marked article was sufficient to uniquely identify the target item than when it was not. Fuchs concluded that HSs were able to use gender information in real-time to facilitate lexical retrieval, in a manner qualitatively like the control group. These results may suggest that early and naturalistic acquisition of gender agreement is fundamental to developing the ability to use gender to facilitate lexical retrieval (Grüter et al., 2012; Montrul et al., 2014), an idea further supported by observations of first language acquisition of nouns and articles (as discussed in more detail in section “Discussion”).

However, the finding that HSs may use gender information to comprehend nouns more efficiently warrants further investigation. Previous work on facilitative use of gender agreement in the processing of nouns in other populations

has suggested that when the experimental method involves a gender cue located on an article that is frequent or obligatory in the language,² such results are consistent with two possible accounts: under a syntactic account, participants are in fact accessing abstract syntactic information on the article during processing of the noun phrase; under a probabilistic account, the results reflect a mechanism that relies on surface probabilities between frequently co-occurring article-noun pairs (van Heugten and Shi, 2009; Lew-Williams and Fernald, 2010; Melançon and Shi, 2015). Existing work on HSs in this domain (Fuchs, 2021) is consistent with either account, and therefore further work is needed to adjudicate between the accounts. Under a syntactic account, we should expect to observe HSs’ facilitative use of grammatical gender when the gender cue is located on a non-frequent, non-obligatory element within the nominal phrase. Under the probabilistic account, however, we might expect significant differences between heritage and control groups in a task that provides the gender cue on such an element.

The present paper presents an eye-tracking study in the Visual World Paradigm that tests whether HSs of Polish are able to use gender information on prenominal adjectives to facilitate lexical retrieval of the subsequent noun. In existing work, this methodology has been used extensively to investigate the processing of grammatical gender by monolingual children and adults (Lew-Williams and Fernald, 2007, 2010; van Heugten and Shi, 2009; Loerts et al., 2013; Melançon and Shi, 2015, among others), as well as by L2 and—more recently—heritage bilinguals (Lew-Williams and Fernald, 2010; Grüter et al., 2012; Dussias et al., 2013; Hopp, 2013, 2016; Sekerina, 2015; Lemmerth and Hopp, 2019; Fuchs, 2021, among others). Moreover, the nominal structure of Polish is best suited for these research purposes, as the language does not have overt articles and places adjectives prenominally in the unmarked word order. The results of this study suggest that HSs of Polish are also target-like in their processing of grammatical gender, which is in-line with the earlier findings regarding HSs of Spanish. The results therefore provide additional support for early naturalistic acquisition as instrumental in developing the ability to use gender to facilitate lexical retrieval in adulthood, but call into question hypotheses regarding the exact mechanism that leads to this ability, as articles—often assumed to be central to this process—do not exist in Polish.

1 The terms “baseline” and “control” are used herein to indicate the population of comparison, instead of terms such as “native” or “monolingual”. HSs are native speakers in their own right, by virtue of the nature of their acquisition process (for discussion see Pascual et al., 2012; Kupisch and Rothman, 2018; Polinsky, 2018, Chp. 2.3; Wiese et al., 2022).

2 While this may apply to both definite and indefinites articles, the discussion in the literature has primarily focused on definite articles, as these cooccur most frequently with nouns and are the most frequent gender cue in the input to child speakers of languages like Spanish that have gender agreement and obligatory articles (cf. Mariscal, 2009, p. 144).

Background

Grammatical gender in heritage languages

Grammatical gender is known to be one of the more vulnerable elements of the grammar of HSs. Robust evidence from many different heritage languages suggests that HSs are non-target like with respect to gender both in their production and in their comprehension (Håkansson, 1995; Montrul et al., 2008; Polinsky, 2008; Scontras et al., 2018). This is the case despite the fact that in most languages for which this has been observed, cues to gender are reliably available in the input in the form of agreement on articles, adjectives, and/or verbs. In fact, monolingual children are able to make use of these cues to acquire fairly target-like gender agreement by around age 2–3, and monolingual adults make virtually no errors in gender agreement in naturalistic speech production (e.g., Hernández Pina, 1984; Soler, 1984; Pérez-Pereira, 1991; Mariscal, 1996, 2001; Lleó, 1997; López-Ornat, 1997). For unbalanced bilinguals, however, this is not the case; although their acquisition of gender may be roughly on par with their monolingual peers in the early stages of acquisition (Pérez-Pereira, 1991; Mariscal, 1996; Lleó, 1997; López-Ornat, 1997; Mueller Gathercole, 2002; Kuchenbrandt, 2005; Eichler et al., 2013; Ticio Quesada, 2018), they appear to diverge around 1st or 2nd grade with higher error rates in their production of gender agreement (e.g., Mueller Gathercole, 2002). Various work has found that HS children tend to over-extend the default gender more so than do their monolingual peers (Sanchez-Sadek et al., 1975; Anderson, 1999; Montrul and Potowski, 2007; Cuza and Pérez-Tattam, 2015). These patterns persist into adulthood, and manifest in divergent production and comprehension of gender agreement, as mentioned above (Montrul et al., 2008; Alarcón, 2011; Boers et al., 2020; Hur et al., 2020). Although a substantial portion of the literature on gender in heritage languages has focused on the heritage gender system in the environment of a dominant language that lacks grammatical gender (i.e., English; cf. Scontras and Putnam, 2020), work on heritage gender systems in the environment of a dominant language with grammatical gender suggests that the effect of the dominant gender system is modest if at all present: HSs whose dominant language has gender are still consistently non-target-like in their production and comprehension of gender in the HL (Cornips and Hulk, 2008; van der Linden and Hulk, 2009; Brehmer and Rothweiler, 2012; Eichler et al., 2013; Schwartz et al., 2015; Meir et al., 2017; Egger et al., 2018; Kaltsa et al., 2019; Rodina et al., 2020).

However, a growing body of evidence suggests that task modality may play a role in HSs' performance on various experimental tasks that have been used to assess their knowledge of grammatical gender (Alarcón, 2011; Montrul et al., 2014),

and online methods may be advantageous in providing a more nuanced understanding of HSs' knowledge (Polinsky, 2018; Bayram et al., 2020; Fuchs, 2021). An example of work using online methods in this domain controls for *what* HSs know about grammatical gender in the HL and instead focuses on *how* they use that information in real time, demonstrating that HSs' processing of grammatical gender may be target-like, counter to expectations based on HSs' divergent performance on offline tasks. In an eye-tracking task in the Visual World Paradigm (Tanenhaus et al., 1995), Fuchs (2021) found that HSs of Spanish were able to access and deploy gender information on prenominal gender-marked articles to facilitate the processing of the subsequent noun. In the study, when viewing a display with two images representing lexical items of different genders ("mismatch" condition) and hearing a prompt that included a prenominal gender-marked cue (the masculine article *el* or the feminine article *la*), both HSs and the control group fixated on the target item faster than when viewing a display with two images of the same grammatical gender ("match" condition), for which the prenominal gender cue did not disambiguate the target item. Despite an absolute difference between the HSs and the controls in looking times across conditions, the HSs' faster looks to target items in the mismatch conditions were an indication that HSs can use gender information on the gender-marked article to narrow the search in the lexicon, thus facilitating lexical retrieval, in the same manner as control speakers of Spanish.

These findings mirror patterns in previous work on monolingual speakers of Spanish (Lew-Williams and Fernald, 2007, 2010; Grüter et al., 2012; Dussias et al., 2013), German (Hopp, 2013, 2016; Hopp and Lemmerth, 2016), and Dutch (Loerts et al., 2013), among others, as well as for child speakers of Spanish (Lew-Williams and Fernald, 2007), German (Lemmerth and Hopp, 2019), and French (van Heugten and Shi, 2009; Melançon and Shi, 2015). Also notable is the fact that the findings for the HSs contrast with findings for L2 learners of languages with grammatical gender, where the findings appear to be variable with respect to whether L2 learners can also use grammatical gender during real-time processing in this way, and whether or not this ability is modulated by proficiency in the L2 (Lew-Williams and Fernald, 2010; Grüter et al., 2012; Dussias et al., 2013; Hopp, 2013; Hopp and Lemmerth, 2016). HSs and L2 learners alike fall on a spectrum of proficiency in their non-dominant language, to which they have less input than to the dominant language, resulting in non-target-like gender categorization and gender agreement for both groups. Given this, HSs patterning with adult and child controls in their ability to use gender to facilitate lexical retrieval has implications for the understanding of how the nature of the acquisition process may impact processing abilities. Following Grüter et al. (2012) and Montrul et al. (2014), Fuchs (2021) suggests that early and naturalistic acquisition of grammatical gender in the speech stream may be crucial for developing robust associations

between pre-nominal gender cues and subsequent nouns, as discussed further in section “Discussion.”

However, there are outstanding questions with respect to the findings for HSs reported in Fuchs (2021). In that study, the HSs—much like the children and adults in Lew-Williams and Fernald (2007, 2010) and Grüter et al. (2012)—were prompted by auditory stimuli in which the gender cue was on a prenominal definite article. In Spanish, these articles are remarkably frequent in the input, as bare nominals are quite constrained in their distribution (cf. Mariscal, 2009; Paolieri et al., 2010). Acquirers, therefore, learn very early on that this cue to gender is both reliable and frequent. This has implications for the results of studies—whether targeting monolingual, child, or bilingual populations—investigating the facilitative use of grammatical gender specifically when the predictive gender cue is located on definite articles. What is interpreted as facilitative use of grammatical gender may on the one hand indeed be driven by a syntactic mechanism, by which participants access abstract gender agreement information in real time and use this to narrow their search within the lexicon to those items that match that gender information. However, these same results would also be consistent with a probabilistic account. Given the frequency of article-noun sequences, participants in these studies might be treating them as memorized phrases and using transitional probabilities between a given article and the candidate nouns in the VWP (Lew-Williams and Fernald, 2007, 2010; van Heugten and Shi, 2009; Melançon and Shi, 2015). This might be particularly likely for children and for HSs, who have accumulated less input in the language and are more likely to treat article-noun sequences units as unanalyzed chunks, similar to what was discussed above.

For monolingual adults and children, follow-up studies have been run to test this question, using methods that primarily involve manipulating the locus of the gender cue. That adult control speakers of Spanish are able to access gender information in real time was shown, among others, by Lew-Williams and Fernald (2010). In Experiments 2 and 3, participants learned novel nouns preceded by an indefinite gender-marked article, but were tested on those nouns using a definite gender-marked article. To succeed on the task, participants had to generalize from the information they were given in the learning phase, rather than just memorize article-noun sequences from the input. Crucially, L2 learners of Spanish in the same task were not able to generalize, suggesting that what appeared to be facilitative use of gender on articles in the initial task was driven by access to probabilistic rather than syntactic knowledge. The control group, however, was able to use gender to facilitate lexical retrieval in this version of the study, suggesting monolingual speakers do indeed access syntactic information on definite articles in real time. Testing use of gender cues on agreeing elements other than articles gets at the same issue from another approach. For instance, Hopp and Lemmerth (2016) showed that adult control speakers of

German were able to use gender information on pre-nominal adjectives to facilitate lexical retrieval. Adjectives are always optional and therefore far less frequent in the input; this makes it significantly less likely that they can be treated as memorized phrases in the experimental setting and aid faster word recognition via a probabilistic process. The syntactic vs. probabilistic account has also been tested for children's facilitative use of grammatical gender, with results in support of the syntactic account: Melançon and Shi (2015) trained 30-month-old French-speaking children on novel nouns by presenting them with gender-marked determiners and agreeing adjectives. In the testing session, they found that the children's comprehension of the nouns was facilitated by the presence of correctly gender-marked articles, suggesting the children had generalized abstract gender information for the nouns and that they accessed this information during processing.

Given results in favor of a syntactic account for monolingual adults and children, it remains to be seen whether HSs can indeed access abstract gender information to facilitate word recognition in real time, or whether their use of gender cues on prenominal elements relies on probabilistic knowledge, more in line with the L2 learners in Lew-Williams and Fernald (2010, Experiment 2). To test this, the present study investigates whether HSs of Polish—a Slavic language that, unlike Spanish, has adjectives that appear prenominally in the unmarked word order—can use grammatical gender information on adjectives to facilitate lexical retrieval of the subsequent noun. Relevant properties of the gender system and of gender agreement marking in Polish will be introduced in the next section before detailing the research question and predictions in section “Research question.”

Gender in Polish

Polish is generally considered to have three (global) genders—masculine, feminine, neuter—as there is evidence for three gender categories in the citation form in the nominative singular, as illustrated in (1). There are some subcategories within the masculine based on animacy, and while there has been some debate concerning the status of these subcategories, formal analyses of Polish as having three global genders and possible subgenders of masculine (Corbett, 1983) are the most widely accepted (for an overview, see Swan, 2015), and are assumed by existing work on the acquisition of Polish gender by monolingual and bilingual children (Brehmer and Rothweiler, 2012).

- (1)
- | | | |
|----|------------|----------------------|
| a. | ta koszula | “this shirt, fem.” |
| | ta książka | “this book, fem.” |
| b. | to jajko | “this egg, neut.” |
| | to okno | “this window, neut.” |

- c. ten stół “this table, masc.”
 ten wazon “this vase, masc.”

lektury] było jasn-e.
 book be.PST.3SG.N clear-N.SG
 “That Jaś had not read the school book was clear.”

As illustrated in (1), each of these three genders has morphological correlates on nouns: *-a* for feminine (ex. *książka* “book, fem.”); *-o*, *-e*, or *ę* for neuter (ex. *okno* “window, neut.” *imię* “name, neut.”), and consonants for masculine (ex. *stół* “table, masc.” *talerz* “plate, masc.”). Like in most gender systems, these correspondences are not one-to-one, and there are exceptions in each gender category: *rzecz* “thing, fem.” *coś* “something, neut.” *mężczyzna* “man, masc.” Given that nominal morphophonology does not uniquely determine the gender category of a noun, the most reliable cue to grammatical gender, as generally established in formal work in this domain, is the agreement patterns that a noun determines on “associated words” in the nominal phrase (Hockett, 1958). In Polish, agreement is pervasive within the noun phrase, as gender category—as well as number and case—determine inflectional marking on attributive adjectives, relative pronouns, and demonstratives, as well as outside of the nominal phrase on predicative adjectives and verbs in certain tenses (Swan, 2015) (2). The default gender agreement for inanimate nouns in Polish is neuter, as evidenced by inflectional morphology in instances with no referent (3a) or with a genderless nominal (3b)—environments in which gender information is either absent or underspecified and therefore default gender agreement rules are deployed (Corbett and Fraser, 1999; Haspelmath, 2006).

(2)

- a. Ten star-y wazon
 DEM.M.SG old-M.SG vase.M
 był w kuchni.
 be.PST.3SG.M in kitchen
 “That old vase was in the kitchen.”

- b. Ta star-a książka
 DEM.F.SG old-F.SG book.F
 była w kuchni.
 be.PST.3SG.F in kitchen
 “That old book was in the kitchen.”

- c. To star-e wiadro
 DEM.N.SG old-N.SG bucket.N
 było w kuchni.
 be.PST.3SG.N in kitchen
 “That old bucket was in the kitchen.”

(3)

- a. Było zimn-o.
 be.PST.3SG.N cold-N.SG
 “(It) was cold (outside).”

- b. [Że Jaś nie przeczytał
 COMP Jaś NEG read.PST.3SG.M

Although this study is restricted to the nominative singular, the remainder of this section will introduce other elements of the agreement paradigm in order to provide a fuller picture of the Polish nominal agreement system. As mentioned above, the inflection on elements agreeing with the head noun is determined not only by grammatical gender but also by number (singular or plural) as well as case. Polish has six syntactic cases—nominative, genitive, dative, accusative, instrumental, and locative—as well as a seventh case (vocative) that is generally considered to be extra-syntactic and even often replaced by the nominative by modern speakers of Polish (Luczynski, 2002). The full inflectional paradigm for the adjective *stary* “old” is presented for illustration in Table 1.

There are two things to note about this paradigm with respect to syncretism. First, the three genders are collapsed in the plural, making it impossible for plural agreement endings to distinguish between genders.³ Second, the masculine and the neuter are syncretic in the singular for all but the nominative and accusative cases. In other words, in the singular, the inflectional endings for the feminine gender are always unique from the other genders, but the masculine and the neuter are only distinguishable from each other in the nominative and accusative.

In the study presented below and in existing work on the acquisition of grammatical gender in Polish, the empirical domain has been narrowed to focus on nouns in the nominative singular. This is guided by the fact that, as mentioned above, this is one of the few parts of the inflectional paradigm where the three genders are both equally morphologically specified and unique from each other. It should be noted that unlike in other Slavic languages such as Russian, these inflectional endings do not undergo vowel reduction and are therefore reliably transparent cues to gender, a fact that may also be relevant to acquisition (Janssen, 2016). In addition, the present study investigates only gender as it occurs on inanimate nouns. Animate nouns occur in each of the three gender categories, although only within the masculine gender category does animacy determine (minimally) different inflectional paradigms. For example, animate masculine nouns take the inflectional ending *-a* in the accusative singular where all other masculine nouns take *-ø*; animate personal nouns take a unique inflectional ending in the nominative plural. Such minimal differences motivate some analyses of Polish gender to posit

³ The exception are certain subcategories of the masculine, which are restricted to animate nouns and are not discussed here. These subcategories have been the subject of debate in the formal literature on the number of Polish gender (sub)categories (see Corbett, 1983; Swan, 2015, among others, for overview and discussion).

subgenders of the masculine (Corbett, 1983).⁴ Given this, the present study restricts the empirical domain to inanimate nouns.

From an acquisitional perspective, the inflectional paradigms for nominal agreement form a critical part of the input to the learner in acquiring the gender system. Recall that, although there are morphophonological correlates to gender on nouns, these are correlations rather than reliable cues with one-to-one mappings between form and gender. Whereas in languages like Spanish this lack of reliable morphophonological cues on the noun is compensated for by the consistent presence of reliable cues on gender-inflected determiners (masculine *el* and feminine *la*) that are often cited as central to the acquisition of gender classes (e.g., Mariscal, 2009; for discussion, see Fuchs et al., in press), Polish lacks such obligatory elements. Thus, the marking on adjectives is one of the main reliable cues to the acquisition of gender for Polish children. However, adjectives are optional and therefore infrequent in the input to the learner (e.g., Behrens, 2006; Stolt et al., 2008; Tribushinina and Gillis, 2012; Tribushinina et al., 2014); the gender cues that they provide are thereby also not frequent in the input. Despite the relative scarcity of gender cues in Polish, work on L1 acquisition of gender distinctions suggests that children nevertheless acquire initial distinctions by around age 2;0 (Smoczynska, 1985), similar to what has been found for languages like Spanish (Hernández Pina, 1984; Soler, 1984; Pérez-Pereira, 1991; Mariscal, 1996, 2001; Lleó, 1997; López-Ornat, 1997). However, children's production of gender agreement suggests that the distinction between masculine and feminine may be acquired before the neuter (Dabrowska, 2006; Janssen, 2016). Krajewski (2005) argues based on evidence

from a corpus of child speech (ages 1;7–2;6) that children first distinguish between the three global genders and only subsequently make animacy distinctions within the masculine gender. Although it has been proposed that diminutivization of nouns—which employs morphological marking on the noun that is consistently transparent for gender category—may also aid in the acquisition of gender by Polish children (Dabrowska, 2006), Janssen (2014) found that diminutives are less frequent in corpora of child-directed speech than previously assumed: diminutives constituted 23% of nouns in the corpus prepared by Haman et al. (2011).

Bilingual acquisition of Polish gender has been less studied. Brehmer and Rothweiler (2012) conducted a production study of bilingual Polish-German children (2;11–6;5) and found that the children produced target-like agreement marking on adjectives agreeing with masculine and feminine nouns, but overextended the masculine in producing agreement with neuter nouns. These patterns were amplified in the children's production of agreement with nonce nouns designed to have morphophonological cues to gender consistent with the correlates of each gender category. More broadly, Haman et al. (2017) found that HSs of Polish aged 4;0–7;5 diverged from monolingual Polish speaking children in both vocabulary and grammatical knowledge, though the difference was more pronounced in production than comprehension. This is generally consistent with work that suggests HSs go through the same developmental stages in the acquisition of grammatical gender as do their monolingual peers, though with some delays (Sanchez-Sadek et al., 1975; Snyder et al., 2001; Kupisch et al., 2002; Mueller Gathercole, 2002; Kuchenbrandt, 2005; Eichler et al., 2013; Ticio Quesada, 2018).

⁴ Complications in how to treat animacy in this system come from both empirical and theoretical considerations. From an empirical perspective, evidence that recent loanwords pattern with traditionally animate non-human masculine nouns in their inflectional paradigm has called into question the status of animacy in the Polish gender system (Fuchs, 2014). From a theoretical perspective, there is also some debate as to whether animate gender features occur in the same locus in the nominal structure or whether they are representationally independent of inanimate gender information (Kramer, 2009, 2014; Steriopolo and Wiltshko, 2010; Bobaljik and Zocca, 2011; but see also Kramer, 2015).

TABLE 1 Inflectional paradigm of three global (inanimate) genders in the singular and plural of six cases.

	Singular			Plural		
	M	F	N	M	F	N
Nominative	stary	stara	stare	stare	stare	stare
Genitive	starego	starej	starego	starych	starych	starych
Dative	staremu	starej	staremu	starym	starym	starym
Accusative	stary	stara	stare	stare	stare	stare
Instrumental	starym	stara	starym	starymi	starymi	starymi
Locative	starym	starej	starym	starych	starych	starych

Research question

Existing work on grammatical gender in HLs has shown that, while adult HSs do show non-target-like knowledge of gender assignment as well as non-target-like production and comprehension of gender agreement in offline studies, their processing of gender agreement may be target-like: HSs of Spanish are faster to recognize a noun when it is preceded by a disambiguating gender-marked article (Fuchs, 2021). However, given the distribution and nature of the definite article in languages like Spanish, these findings are in fact compatible with both a grammatical account of facilitative use of grammatical gender—wherein participants access and integrate abstract gender features in real time word recognition—and a probabilistic account—wherein participants rely on transfer probabilities between article and noun derived from the frequency of the co-occurrence of these elements in the input. To test whether HSs can indeed access abstract gender information to facilitate lexical retrieval, the present study tests HSs of a language in which articles do not exist and

in which pre-nominal gender cues appear on optional and infrequent elements (adjectives). The research question asked and addressed by the present study is therefore the following:

Research question: Can HSs of Polish use grammatical gender information on prenominal adjectives to facilitate the lexical retrieval of the subsequent noun?

The two conditions for comparison are the mismatch condition—in which the items in the display are of different genders, and the gender marking on the adjective may therefore serve as a facilitative cue—and the match condition—in which the items are of the same gender, and therefore the onset of the lexical item is the first unique cue to the target item. Under the syntactic account of facilitative use of grammatical gender, HSs can access the abstract syntactic gender agreement feature during online processing and use it to narrow the list of candidates in the mental lexicon, and we should therefore expect the HSs of Polish recruited for this study to fixate faster on target items in mismatch condition trials than in match condition trials—in line with what was observed for HSs of Spanish using gender information on articles to facilitate lexical retrieval in Fuchs (2021). If, however, the probabilistic account is more accurate, then we should expect HSs to be unable to use gender cues on adjectives to facilitate lexical retrieval; in this study, this means we would expect their looks to the target item to occur at about the same time across trials in both conditions.

Materials and methods

Materials and design

Images of 36 picturable concrete items were selected to build visual displays for the study, equally split by gender: 12 masculine, 12 feminine, 12 neuter. Corresponding lexical items were at least two syllables long to allow for looking time. To ensure a clear word boundary between the gender cue on the prenominal adjective and the onset of the noun, all lexical items had a consonant as their first phoneme. A full list of target items is available in [Supplementary material](#). Within each gender category, 4 items were colored green, 4 red, and 4 blue. These colors were chosen because the corresponding color adjectives have an equal number of syllables. See [Table 2](#) for the appropriately inflected forms of each color adjective.

TABLE 2 Three color adjectives with equal number of syllables were selected for the study.

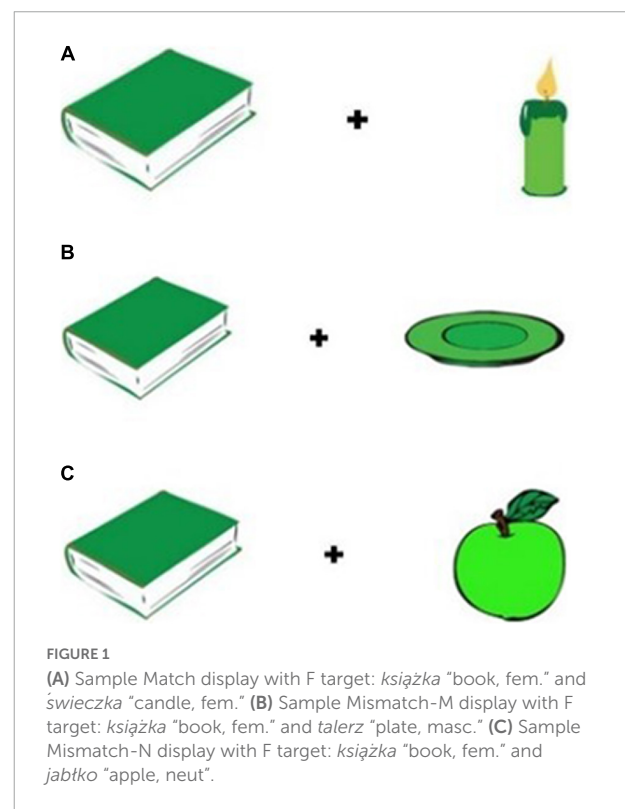
	M	F	N
Red	czerwon-y	czerwon-a	czerwon-e
Green	zielon-y	zielon-a	zielon-e
Blue	niebiesk-i	niebiesk-a	niebiesk-ie

The images were combined into 108 visual displays. Each display consisted of two images equidistant from a center fixation cross; one image was the target item, the other was the distractor. Because there are three genders, for each gender there was a match condition and two mismatch conditions based on the gender of the distractor, as schematized in [Table 3](#). Each image appeared as the target item three times: once in a match condition, and once in each mismatch condition. In total there were 36 match displays, 24 mismatch-M displays, 24 mismatch-F displays, and 24 mismatch-N displays. Sample visual displays are presented in [Figure 1](#).

Visual displays were paired with auditory stimuli of the form in (4), prompting the participant to direct their gaze to the target item. Given that Polish does not have overt articles and that the goal was to establish whether HSs can use gender information on adjectives to facilitate lexical retrieval, the gender cue was the inflectional suffix on a color adjective (Sekerina, 2015; Hopp and Lemmerth, 2016; Lemmerth and Hopp, 2019). The overall

TABLE 3 Experimental conditions for each target gender.

		Distractor gender		
		M	F	N
Target gender	M	match	mismatch-F	mismatch-N
	F	mismatch-M	match	mismatch-N
	N	mismatch-M	mismatch-F	match



structural simplicity of the sentence and light semantic load of the cue-carrying element were modeled after previous work on facilitative use of gender (Lew-Williams and Fernald, 2007, 2010; Grüter et al., 2012; Loerts et al., 2013; Sekerina, 2015; Hopp and Lemmerth, 2016; Lemmerth and Hopp, 2019; Fuchs, 2021).

- (4) Gdzie jest COLOR-GEN NOUN?
 where is color-GEN noun
 “Where is (the) green/red/blue noun?”

All sentences were first recorded by a male native speaker of Polish (L2 English) immigrated to the US within 1 year of the date of recording. The final auditory prompts were created by splicing a single token of *gdzie jest* “where is” with single tokens of each of the nine inflected forms of adjectives (cf. Table 2) and tokens of each lexical item. Splicing was intended to (a) eliminate possible effects of co-articulation that might give unintended cues to the target item, and (b) to ensure that the gender cue and the lexical item occurred at the same time across stimuli for ease of comparison and analysis. For all stimuli, the onset of the gender cue occurred at 1150 ms after the start of the auditory prompt, and the onset of the lexical item occurred 480 ms later. The average duration of lexical items was approximately 700 ms.

Visual and auditory stimuli were presented together, and each trial lasted 6 s. Each trial included 800 ms of looking time and an auditory signal that prompted participants to direct their gaze to the fixation cross before the auditory stimulus began at 1000 ms into the trial. Between each trial there was a 1-s break during which only the fixation cross was visible on the screen.

Participants

Fifty-five speakers of Polish participated in the study. Participants completed an abbreviated version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007; Kaushanskaya et al., 2020) to gather demographic information and self-reported proficiency measures. The LEAP-Q was also translated into Polish for the purposes of this study,⁵ and participants could choose to fill out the LEAP-Q in either English or Polish. Control speakers of Polish were identified as those who were born in Poland and lived at least 18 years in Poland ($n = 23$). HSs of Polish were those who reported (a) that Polish was either their sole first language or their first language acquired simultaneously with English and (b) that they had lived in Poland for 8 years or less ($n = 18$).⁶

A subset of the demographic data collected from the LEAP-Q is presented in Table 4. Self-reported proficiency scores were

collected, but, given the generally accepted lack of reliability of self-reported scores particularly for HSs, an oral lexical identification task was used to assess proficiency (Polinsky, 1997, 2006; Godson, 2003; Fuchs, 2021), discussed further in Section “Oral production task”.

Procedure

Participants were tested individually in a lab. They completed the LEAP-Q either in English or in Polish, then completed the oral production task used for data cleaning and as a proxy measure for proficiency. During this task, participants viewed a set of slides with each of the 36 images used as target items during the study. They were asked to orally label each image using a color adjective and a noun. In the event that a participant was unable to recall a word for a given image, they were allowed to move on to the next image without providing a response (these responses were marked as incorrect in both coding schemes discussed in section “Results”). Their response times for each individual image were not recorded, but the total time to complete the task was recorded. This occurred prior to the comprehension task in order to assess lexical knowledge prior to exposure to the lexical items in the comprehension task (Lieberman et al., 2018; Fuchs, 2021).

Participants then completed the eye-tracking comprehension task. Participants received oral and written instructions. They sat facing a 53.5-cm screen approximately 75 cm away from it, with their head in a chin-support apparatus that ensured minimal head movement during the task. Participants saw four practice trials, after which an SMR Eyelink 1,000 was calibrated, with the goal of achieving visual acuity below 0.5 degrees. Gaze position was recorded at 2,000 Hz. The task was split into two equal parts of 54 trials each; between the two parts, participants were given a break of self-determined duration. Calibration of the eye-tracker was repeated before the second half of the task. Participation in the entire study took approximately 45 min, and participants were compensated for their time.

Results

Oral production task

Responses to the oral production task were coded twice, once for the purposes of a measure of lexical proficiency, and once for the purposes of data cleaning (see section “Eye-tracking comprehension task”). To obtain a measure of lexical proficiency, participant responses were marked as correct if they produced an appropriate label for a given image (variants not intended by the experimental design were accepted, ex. *paczka* “package, fem.” for the intended *pudelko* “box, neut.”)

⁵ The Polish translation of the LEAP-Q is now available at <https://bilingualism.northwestern.edu/leapq/>.

⁶ Fourteen participants were excluded from data analysis. A majority of these left Poland after the age of 8 but before 18 and therefore did not reliably belong to either category. Two reported that their L1 was neither English nor Polish.

TABLE 4 Selected demographic information of the control and HS participants, as self-reported in the LEAP-Q.

			Time spent in Polish-speaking environment, in years (sd)			
	<i>n</i>	Age	Country	Family	Work/School	
Controls	23	31.8 (8.7)	25.0 (8.7)	24.2 (10.8)	19.8 (9.7)	
Heritage	18	26.1 (9.9)	0.8 (1.0)	14.4 (12.5)	1.7 (3.1)	
	Number of participants at each educational level					
	H.S.	Some coll.	College	Some Grad	Masters	PhD
Controls	1	2	3	0	9	8
Heritage	2	3	8	1	3	1

along with a correctly gender-marked color adjective. The resulting proportion of correctly labeled items (out of 36) is reported in [Figure 2](#). Control participants performed effectively at ceiling, with one or two exceptions. The HSs were able to correctly label on average approximately 28 (out of 36, $sd = 8.8$) of the images. There was a significant difference between the two groups (Wilcoxon Rank-Sum test, $p < 0.001$), and the HSs showed more variability (min. = 8, max. = 36).

Eyetracking comprehension task

The aim of the study was to observe use of grammatical gender while controlling for categorization, i.e., for those lexical items for which the HSs arguably know the correct grammatical gender. To achieve this, only trials for which participants knew both the lexical items in the corresponding visual display, along with their grammatical genders, were included for analysis. This required a second coding of the oral production task: in this version, responses were coded as correct only when

participants labeled a given image using the label intended by the experimental design and that label's corresponding grammatical gender (i.e., in this case, *paczka* “package, fem.” for the intended *pudelko* “box, neut.” coded as incorrect, which is especially important given that the produced lexical item belongs to a different grammatical gender than the one intended in the experiment). Removing—for each participant—trials in which they did not label or mislabeled one or both of the images excluded 40% of the trials for the HSs and 6% of the trials for the control group.⁷ For the remaining trials, time of first fixation (response time) was gathered for each participant and analyzed in R ([R Core Team, 2022](#)) using the *lme4* package ([Bates et al., 2015](#)). For linear mixed effects models, p -values were approximated using the Satterthwaite method implemented in the *lmerTest* package ([Kuznetsova et al., 2017](#)). Time of first fixation was defined as the earliest fixation on the interest area of the target item after the onset of the gender-marked adjectival suffix, which was 3250 ms after the start of each trial. The resulting times were trimmed to within two standard deviations of the mean, excluding approximately 4.6% of the data. Given the 3×3 experimental design, for ease of exposition the results will be presented according to the gender of the target item.

Feminine target noun results

The mean first fixation times for the heritage group and the control group on trials with feminine target items are presented in [Figure 3A](#). A mixed effects linear model was fit to the data, predicting time of first fixation by GROUP, CONDITION, and TRIAL, as well as their pairwise and three-way interactions, with random intercepts and slopes for CONDITION grouped by PARTICIPANT.⁸ The categorical CONDITION variable was Helmert contrast-coded to test for a significant difference between the two mismatch conditions (for ease of exposition this is referred to as Condition-Distractor below), and then to compare the match condition to the two mismatch conditions taken together (referred to as Condition-Match below). GROUP

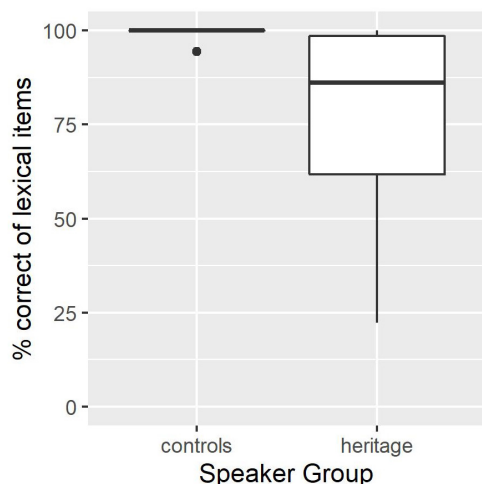


FIGURE 2
Percent of target items that participants in control and heritage groups labeled using an appropriate noun and a gender-marked adjective matching the gender of the noun.

⁷ These percentages of data removal are similar to the outcome of the same procedure applied in [Fuchs \(2021\)](#).

⁸ An additional model with random intercepts and slopes for CONDITION grouped by COLOR was attempted but resulted in a singular fit.

was a categorical variable with two levels and was contrast-coded. Since the order of trials was randomized for each participant, TRIAL was a continuous variable indicating the order in which a given stimulus occurred in the study for the given participant; the variable was centered and scaled.

The model found no significant effect of Condition-Distractor ($\beta = 2.33$, $SE = 12.51$, $t = 0.19$, $p = 0.852$), but did find a significant effect of Condition-Match ($\beta = 59.80$, $SE = 9.71$, $t = 6.16$, $p < 0.001$)—participants' mean time of first fixation was overall faster on mismatch conditions than on the match condition. The model also found a significant effect of TRIAL ($\beta = -47.41$, $SE = 8.84$, $t = -5.36$, $p < 0.001$): the average time of first fixation was faster in later trials. Crucially, the model did not find a significant difference between the heritage group and the control group ($\beta = 59.48$, $SE = 71.77$, $t = 0.83$, $p = 0.412$). A full summary of fixed effects for the model is reported in [Table 5](#). *Post hoc* models were fit to the data to probe the significant three-way interaction between GROUP, CONDITION-DISTRACTOR, and TRIAL in the original model. For the control group, the follow-up analysis revealed no significant interaction of CONDITION-DISTRACTOR and TRIAL. The follow-up analysis for the HSs did find a significant interaction, driven by a significant effect of TRIAL on response time on Mismatch-M trials ($\beta = -74.83$, $SE = 27.45$, $t = -2.73$, $p = 0.008$). This indicates that, over the course of the experiment, HSs were increasingly quick to fixate on the target item in trials in which the target was feminine and the distractor was masculine.

Masculine target noun results

The mean first fixation times for the heritage group and the control group on trials with masculine target items are presented in [Figure 3B](#). A mixed effects linear model was fit to data, predicting time of first fixation by GROUP, CONDITION, and TRIAL, as well as their pairwise and three-way interaction, with random intercepts and slopes for CONDITION grouped by PARTICIPANT and random intercepts and slopes for CONDITION grouped by adjective COLOR. The CONDITION variable was Helmert contrast-coded to test for the same contrasts as before. The full results for the fixed effects of the model are presented in [Table 6](#). The model found a significant difference between mismatch conditions with different distractor genders ($\beta = 27.33$, $SE = 11.93$, $t = 2.29$, $p = 0.032$), suggesting time of first fixation on masculine target items was faster when the distractor was feminine than when the distractor was neuter. The model also found a significant difference between the match condition and the two mismatch conditions taken together ($\beta = 38.36$, $SE = 9.43$, $t = 4.07$, $p = 0.003$). The model also found a significant effect of TRIAL ($\beta = -48.72$, $SE = 7.14$, $t = -6.84$, $p < 0.001$), but no significant effect of GROUP was identified ($\beta = 54.68$, $SE = 66.25$, $t = 0.83$, $p = 0.414$).

Neuter target noun results

The mean first fixation times for the heritage group and the control group on trials with neuter target items are presented in [Figure 3C](#). A mixed effects linear model was fit to data, predicting time of first fixation by GROUP, CONDITION, and TRIAL, as well as their pairwise and three-way interaction, with random intercepts and slopes for CONDITION grouped by PARTICIPANT and random intercepts and slopes for CONDITION grouped by adjective COLOR. The categorical CONDITION variable was Helmert contrast-coded as before. The output of this model for fixed effects is presented in [Table 7](#). The model found a significant effect of the gender of the distractor ($\beta = 50.78$, $SE = 10.21$, $t = 4.97$, $p < 0.001$) as well as a significant effect of match vs. mismatch condition ($\beta = 26.39$, $SE = 7.24$, $t = 3.65$, $p < 0.001$). The model also found a significant effect of TRIAL ($\beta = -40.13$, $SE = 7.16$, $t = -5.60$, $p < 0.001$) but no significant effect of GROUP ($\beta = 47.42$, $SE = 69.43$, $t = 0.68$, $p = 0.499$).

Subsequent visual analysis of [Figure 3C](#) motivated further questions regarding the contrasts between the two mismatch conditions with neuter targets. An additional *post hoc* linear mixed effects model was therefore fitted the data predicting time of first fixation by GROUP, CONDITION, and TRIAL, as well as their pairwise and three-way interactions, with random intercepts and slopes for CONDITION grouped by PARTICIPANT and random intercepts for CONDITION grouped by adjective COLOR⁹; this time, CONDITION was Helmert contrast-coded to first test for a difference between Match and Mismatch-M, and then to test for a difference between these two conditions combined as compared to the Mismatch-F condition. The model found no significant difference between Match and Mismatch-M ($\beta = -15.71$, $SE = 10.03$, $t = -1.56$, $p = 0.12$), but did find a significant difference between Mismatch-F and the other two conditions ($\beta = -38.98$, $SE = 7.51$, $t = -5.19$, $p < 0.001$). The model did not find a significant effect of GROUP ($\beta = 47.96$, $SE = 69.49$, $t = 0.69$, $p = 0.49$) but did find a significant effect of TRIAL ($\beta = -40.16$, $SE = 7.16$, $t = 5.61$, $p < 0.001$), consistent with the results of the earlier planned analysis. The output for fixed effects of this model is presented in [Table 8](#).

Discussion

The goal of this study was to address the following research question: Can heritage speakers of Polish use grammatical gender information on prenominal adjectives to facilitate the lexical retrieval of the subsequent noun? The prediction based on previous work on the facilitative use of grammatical gender was that, if HSs of Polish are able to use gender information

⁹ A similar model with added random slopes for CONDITION grouped by COLOR was attempted but resulted in a singular fit.

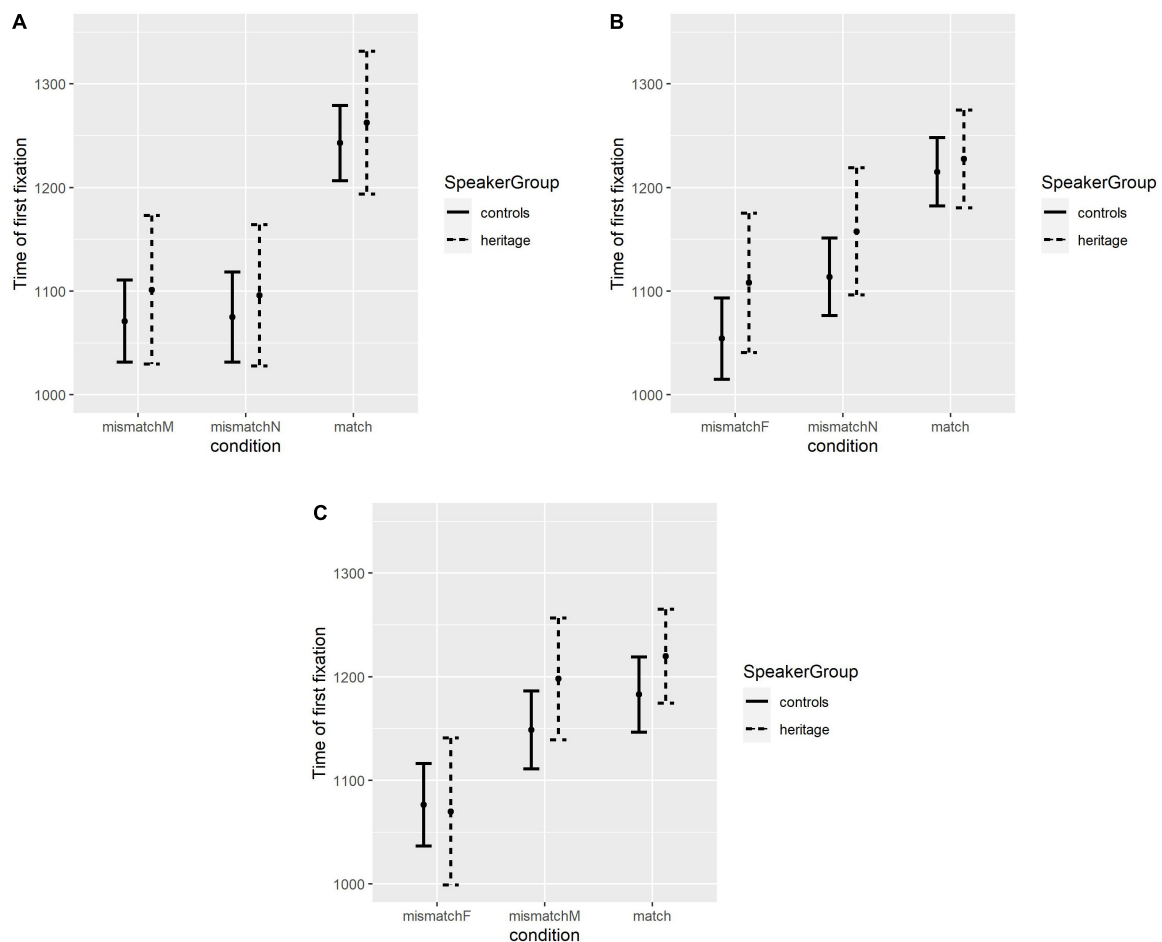


FIGURE 3

(A) Mean response times for conditions with F target items, split by group. Bars represent standard error. (B) Mean response times for conditions with M target items, split by group. Bars represent standard error. (C) Mean response times for conditions with N target items, split by group. Bars represent standard error.

to facilitate lexical retrieval, they should be able to fixate on target nouns faster in mismatch conditions, in which the agreement marking on the pre-nominal adjective provides a disambiguating cue to the subsequent noun, than on match conditions, in which the onset of the lexical item is the first available disambiguating cue. The results of the experimental study described above are consistent with this prediction: for target nouns in each of the three grammatical gender categories, participants in both groups fixated on target items faster in the mismatch conditions than, on average, in the match conditions. This indicates that, when knowledge of gender categorization is controlled for, HSs are able to access gender information in real-time and to use it to narrow the search in the mental lexicon to facilitate lexical retrieval of the target item.

The results lend support for the argument that HSs' processing of grammatical gender in real-time is qualitatively target-like (Fuchs, 2021). The results suggest only one quantitative difference between the HSs and the control group:

in conditions with feminine target nouns, HSs' speed in fixating on the target item increased over the course of the study, whereas for the control group there was no evidence of change over time. This is indicative of a learning effect specific to the HSs, and is also consistent with previous findings for heritage speakers (Fuchs, 2021).

It is also worth noting that in some cases it appears that use of grammatical gender to facilitate lexical retrieval may be modulated by the gender of the distractor. For conditions in which the target noun was masculine, looks to the target item were slower when the distractor was neuter than when the distractor was feminine. For conditions in which the target noun was neuter, looks to the target were slower when the distractor was masculine than when the distractor was feminine—in fact, results suggest that when the distractor was masculine, the speed of first fixation on the neuter target was comparable to that of first fixation on the target item in the corresponding match condition, in which no disambiguating cue was available on the

TABLE 5 Fixed effects of the linear model fit to data for trials in which the target item was feminine.

	Reaction time			
	β	SE	t	P
Group	59.48	71.77	0.83	0.412
Condition-Distractor	2.33	12.51	0.19	0.852
Condition-Match	59.80	9.71	6.16	<0.001***
Trial	-47.41	8.84	-5.36	<0.001***
Group: Condition-Distractor	8.36	25.01	0.33	0.738
Group: Condition-Match	3.94	19.42	0.20	0.840
Group: Trial	-11.92	17.68	-0.67	0.500
Condition-Distractor: Trial	15.27	10.57	1.44	0.149
Condition-Match: Trial	10.23	6.35	1.61	0.107
Group: Condition-Distractor: Trial	42.18	21.14	2.00	0.046*
Group: Condition-Match: Trial	-5.94	12.70	-0.47	0.640
Constant	4406.72	35.89	122.79	<0.001***
Observations	1046			
Akaike Inf. Crit.	14604.86			
Bayesian Inf. Crit.	14684.11			

* $p < 0.05$; ** $p < 0.01$; and *** $p < 0.001$.

TABLE 6 Fixed effects of the linear model fit to the data for trials in which the target item was masculine.

	Reaction time			
	β	SE	t	p
Group	54.68	66.25	0.83	0.414
Condition-Distractor	27.33	11.93	2.29	0.032*
Condition-Match	38.36	9.43	4.07	0.003**
Trial	-48.82	7.14	-6.84	<0.001***
Group: Condition-Distractor	-5.16	21.90	-0.24	0.814
Group: Condition-Match	-12.75	16.32	-0.78	0.438
Group: Trial	-12.02	14.28	-0.84	0.400
Condition-Distractor: Trial	10.44	8.75	1.19	0.233
Condition-Match: Trial	6.76	5.09	1.33	0.184
Group: Condition-Distractor: Trial	4.87	17.45	0.28	0.780
Group: Condition-Match: Trial	9.18	10.16	0.90	0.366
Constant	1155.05	34.51	33.47	<0.001***
Observations	1162			
Akaike Inf. Crit.	15975.2			
Bayesian Inf. Crit.	16071.3			

* $p < 0.05$; ** $p < 0.01$; and *** $p < 0.001$.

adjective. Notably, no such asymmetries between the gender of the distractor were evident when the target item was feminine.

While a thorough analysis of these patterns is beyond the scope of the present paper, these patterns suggest a hierarchical organization of the abstract gender features that leads to interference in access between some of them but not others (Fuchs, manuscript under revision). For present purposes, it suffices to say that the results suggest that these patterns are replicated in the heritage population, indicating that whatever drives the modulation of processing of the target gender by the distractor gender in Polish is equally active in the heritage grammar.

TABLE 7 Fixed effects of the linear model fit to data for trials in which the target item was neuter.

	Reaction time			
	β	SE	t	p
Group	47.42	69.43	0.68	0.499
Condition-Distractor	50.78	10.21	4.97	<0.001
Condition-Match	26.39	7.24	3.65	<0.001
Trial	-40.13	7.16	-5.60	<0.001
Group: Condition-Distractor	25.62	20.43	1.25	0.212
Group: Condition-Match	6.80	14.48	0.47	0.641
Group: Trial	4.12	14.33	0.29	0.774
Condition-Distractor: Trial	2.52	8.58	0.29	0.769
Condition-Match: Trial	6.25	5.23	1.19	0.232
Group: Condition-Distractor: Trial	-4.68	17.16	-0.27	0.785
Group: Condition-Match: Trial	15.86	10.47	1.52	0.130
Constant	1154.21	34.72	33.25	<0.001
Observations	1116			
Akaike Inf. Crit.	15306.5			
Bayesian Inf. Crit.	15391.8			

* $p < 0.05$; ** $p < 0.01$; and *** $p < 0.001$.

TABLE 8 Fixed effects of the *post hoc* linear model fit to data for trials in which the target item was neuter, with CONDITION Helmert-contrast-coded to test for a difference between Mismatch-M and Match, and then between those and Mismatch-F.

	Reaction time			
	β	SE	t	p
Group	47.96	69.49	0.69	0.494
Condition-Match-Mismatch-M	-15.71	10.03	-1.56	0.120
Condition-Mismatch-F	-38.98	7.51	-5.19	<0.001***
Trial	-40.16	7.16	-5.61	<0.001***
Group: Condition-Distractor	0.61	20.07	0.03	0.976
Group: Condition-Match	-15.72	15.02	-1.04	0.302
Group: Trial	3.36	14.32	0.23	0.815
Condition-Distractor: Trial	-7.60	8.67	-0.88	0.381
Condition-Match: Trial	-4.13	5.17	-0.80	0.425
Group: Condition-Distractor: Trial	-24.81	17.35	-1.43	0.253
Group: Condition-Match: Trial	-5.82	10.35	-0.56	0.574
Constant	1154.56	34.76	33.21	<0.001***
Observations	1116			
Akaike Inf. Crit.	15305.7			
Bayesian Inf. Crit.	15391.0			

* $p < 0.05$; ** $p < 0.01$; and *** $p < 0.001$.

Both in the overall facilitation effect and in the modulation effect, then, the HSs in this study performed qualitatively like the control group. This indicates that, despite surface differences in gender categorization and production of gender agreement, HSs' real-time processing of gender agreement is target-like. Combined with the results from Fuchs (2021), this underscores two important things. First, these findings echo the importance of online studies in achieving a more complete understanding of HSs' language abilities. The view from offline studies is only partial: yes, HSs across heritage languages consistently diverge from comparison groups in gender categorization and gender agreement, but online studies now demonstrate that when

one controls for gender categorization, HSs are able to access gender agreement features in real-time and integrate them into their word recognition process much like monolingual-speaking adults and children. The granularity of methods such as eye-tracking thus allows researchers to observe how speakers process linguistic information moment-by-moment, which adds critical nuance to our existing understanding of HSs' linguistic abilities.

Second, these findings are consistent with existing proposals for early and naturalistic experience with gender agreement in nominal phrases in the speech stream as central to developing target-like processing of agreement in adulthood. While both L2 learners and HSs show divergent production and comprehension of gender agreement in offline tasks, results from VWP studies suggest HSs are able to access abstract syntactic information to facilitate lexical retrieval, similar to what has been shown for monolingual adults and children (Lew-Williams and Fernald, 2007, 2010; van Heugten and Shi, 2009; Grüter et al., 2012; Dussias et al., 2013; Hopp, 2013, 2016; Loerts et al., 2013; Melançon and Shi, 2015; Hopp and Lemmerth, 2016; Lemmerth and Hopp, 2019). This suggests that HSs' real-time processing of gender agreement within the nominal phrase is more target-like than that of L2 learners.¹⁰

The explanation that has been offered for this lies in the nature of the L1 vs. the L2 acquisition processes (Lew-Williams and Fernald, 2010; Grüter et al., 2012; Fuchs, 2021). The logic is as follows: children acquire a language naturalistically from the speech stream. They encounter article-noun sequences in the input frequently, and it is thought that in an early stage of the acquisition process they treat these sequences as unanalyzed chunks, only subsequently segmenting them into an article and a noun (MacWhinney, 1978; Carroll, 1989; Pine and Lieven, 1997; Tomasello, 2000; Abu-Akel et al., 2004; Bassano et al., 2008; Mariscal, 2009). Evidence for this comes from children's early (age 1;6–2;0) production of “proto-determiners” on nouns—pre-nominal vowels whose phonology approximates the vowel of the correct definite article (López-Ornat, 1997). It has been suggested that this acquisition process facilitates the development of a tight link between the article and noun. By contrast, L2 learners' acquisition (at least in a traditional classroom setting), proceeds primarily from written material and is aided by a wealth of metalinguistic information, and may therefore not lead to the same robust associations between articles and nouns as in naturalistic acquisition. HSs, having acquired their heritage language in the home as children, share a naturalistic acquisition process with the L1 child and adult populations that have been investigated in these studies.

Per this hypothesis, this is why HSs—despite having non-target-like agreement production and comprehension like L2 learners—nevertheless pattern with monolingual adults and children in the processing of gender in noun phrases. Montrul et al. (2014) put forward a similar hypothesis to explain why HSs performed more like the control group than the L2 learners in an offline task targeting implicit knowledge of grammatical gender.

The present results are in line with the general observation: HSs pattern with monolingual adults and children in the processing of grammatical gender, which may point to the nature of the acquisition process as being instrumental in the development of the ability to use gender agreement to facilitate lexical retrieval. However, recall that Polish does not have overt determiners—no equivalent element to the *el/la* that is overt and obligatory in most contexts for Spanish (and the equivalent for French, Italian, etc.). Therefore, the finding that HSs' processing of gender agreement generalizes to agreement on (optional and infrequent in the input) adjectives suggests that our understanding of *what* exactly in the acquisition process is critical to the development of target-like processing of grammatical gender should also generalize beyond languages with obligatory articles.

Adjectival gender agreement has been far less studied than agreement on articles; the available existing evidence suggests it is learned later than agreement with determiners (Mariscal, 2009; Boloh and Ibernion, 2010). For children acquiring Polish naturalistically, acquiring gender agreement on adjectives poses additional challenges. While the unmarked word order is for adjectives to precede nouns in the noun phrase, adjectives may also follow the noun. Moreover, adjectives need not appear with an overt noun or even be linearly adjacent to the noun (a construction known as split nominals). Such long-distance dependencies are known to be harder to acquire than short-distance ones (Wilson et al., 2020). Polish-speaking children—whether on their way to becoming monolingual or bilingual adults—therefore have to learn grammatical gender from infrequent cues with irregular linear relationships to their target nouns.

Returning then to HSs' target-like performance on facilitative use of grammatical gender—now observed both in Spanish and in Polish: that the HSs pattern with baseline children and adults in these studies is still evidence that the nature of the L1 acquisition process, as opposed to L2 acquisition, may play an important role in determining target-like processing of grammatical gender agreement. However, Polish demonstrates that this need not be solely linked to the cooccurrence of article-noun sequences in the input to the acquirer, whether mono- or multilingual; rather, early and naturalistic acquisition likely entails generalizing gender information from other nominal elements in the speech stream such as adjectives as well, and one of the outcomes of this acquisition is an ability to access gender information in real

¹⁰ Dussias et al. (2013) present an exception, as they found that L1-Italian L2-Spanish learners could fixate on target items faster in mismatch than match conditions, but only when the target item was feminine. The authors speculate this may be driven by overlap in the gender systems of the two languages and/or by the similarity in form between feminine definite articles in the two languages.

time processing in adulthood that is robust to pressure from reduced input to the heritage language.

Conclusion

This paper presented the results of an eyetracking study using the Visual World Paradigm to assess the ability of heritage speakers of Polish to use gender agreement cues on prenominal adjectives to facilitate the lexical retrieval of the subsequent noun. The results showed that both HSs and the control group were able to fixate faster on target items in mismatch conditions, when the adjective inflected for gender served as a cue to the target item, than in match conditions, when the earliest disambiguating cue was the onset of the lexical item.

A previous study in this domain (Fuchs, 2021) similarly found that HSs of Spanish can use gender cues on prenominal articles to facilitate the lexical retrieval of the subsequent noun. However, the frequent and obligatory nature of these articles suggests that the results were in fact compatible with two accounts of what drove the facilitative effect. Under a syntactic account, the HSs accessed abstract gender agreement information on the pre-nominal element and integrated this information into word recognition; under a probabilistic account, HSs were not accessing gender information so much as relying on surface probabilistic properties of article-noun sequences. While the former has been found to be the case for monolingual children and adults (van Heugten and Shi, 2009; Lew-Williams and Fernald, 2010; Melançon and Shi, 2015), the latter has been found to be true for another group of unbalanced bilinguals: L2 learners (Lew-Williams and Fernald, 2010). The present paper tested these two accounts of processing of gender agreement in the noun phrase for HSs by using eye-tracking in the VWP to determine whether HSs of Polish can use gender information on inflected pre-nominal adjectives to facilitate lexical access of the subsequent noun.

The results indicate that HSs can indeed access and deploy abstract gender agreement information on pre-nominal elements during real-time processing in a target-like manner. Taken together with previous work on facilitative use of grammatical gender in monolingual and bilingual populations, these findings have implications for our understanding of what determines target-like processing of grammatical gender in adulthood. Although HSs are like L2 learners in their generally observed non-target-like gender categorization and gender agreement, they nevertheless pattern with monolingual adults and children in their facilitative use of grammatical gender. The results are consistent with the proposal that early and naturalistic acquisition of grammatical gender from the speech stream is likely central to the development of this ability (Lew-Williams and Fernald, 2010; Grüter et al., 2012; Montrul et al.,

2014; Fuchs, 2021), as it captures why HSs, L1 children, and baseline adults pattern together to the exclusion of adult L2 learners. However, it calls into question the assumption that it is precisely the frequent co-occurrence of articles and nouns that is central to the development of target-like processing of gender agreement in the noun phrase—this proposal is simply untenable for languages like Polish, which do not have overt articles and whose flexible word order implies gender cues do not necessarily appear linearly adjacent to nouns in the input. Nevertheless, the finding that HSs of both Spanish and Polish can use gender information on agreeing elements to facilitate the lexical retrieval of nouns in real time suggests that this ability is robust to reduced input to the heritage grammar.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the author, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the University of Maryland IRB. The participants provided their written informed consent to participate in this study.

Author contributions

ZF designed and conducted the study, completed the statistical analysis, and wrote the manuscript.

Funding

This work was supported in part by a Graduate Research Grant from the Institute for Quantitative Social Sciences at Harvard University.

Acknowledgments

For very helpful feedback and discussion the author grateful to Maria Polinsky, Jonathan Bobaljik, Holger Hopp, Kevin Ryan, and Irina Sekerina, as well as to the audience at the Workshop on Gender in Heritage Languages and Grammar hosted by the Multigender Project at the Center for Advanced Study in Oslo and the audience at the 94th Annual Meeting of the Linguistic Society of America. The author also grateful to my research assistants Michelle Min and Sadie Pate for their help with data collection. All errors are my own.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the author and do not necessarily represent those of her affiliated

organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.960376/full#supplementary-material>

References

- Abu-Akel, A., Bailey, A. L., and Thum, Y.-M. (2004). Describing the acquisition of determiners in English: a growth modeling approach. *J. Psycholinguist. Res.* 33, 407–424. doi: 10.1023/B:JOPR.0000039548.35396.c2
- Alarcón, I. (2011). Spanish gender agreement under complete and incomplete acquisition: Early and late bilinguals' linguistic behavior within the noun phrase. *Biling. Lang. Cogn.* 14, 332–350. doi: 10.1017/S1366728910000222
- Albirini, A., Benmamoun, E., and Chakrani, B. (2013). Gender and number agreement in the oral production of Arabic heritage speakers. *Biling. Lang. Cogn.* 16, 1–18. doi: 10.1017/S1366728912000132
- Albirini, A., Benmamoun, E., and Saadah, E. (2011). Grammatical features of Egyptian and Palestinian Arabic heritage speakers' oral production. *Stud. Second Lang. Acquisit.* 33, 273–303. doi: 10.1017/S0272263110000768
- Anderson, R. T. (1999). Loss of gender agreement in L1 attrition: preliminary results. *Biling. Res. J.* 23, 389–408. doi: 10.1080/15235882.1999.10162742
- Bassano, D., Maillochon, I., and Mottet, S. (2008). Noun grammaticalization and determiner use in French children's speech: A gradual development with prosodic and lexical influences. *J. Child Lang.* 35, 403–438. doi: 10.1017/S0305000907008586
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statist. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bayram, F., Rothman, J., Pisa, G., and Slabakova, R. (2020). Current trends and emerging methodologies in charting heritage language bilingual grammars. *PsyArXiv[preprint]* doi: 10.31234/osf.io/pa83g
- Behrens, H. (2006). The input-output relationship in first language acquisition. *Lang. Cogn. Proc.* 21, 2–24. doi: 10.1080/01690960400001721
- Bobaljik, J. D., and Zocca, C. L. (2011). Gender markedness: the anatomy of a counter-example. *Morphology* 21, 141–166. doi: 10.1007/s11525-010-9156-3
- Boers, I., Sterken, B., van Osch, B., Parafita Couto, M. C., Grijzenhout, J., and Tat, D. (2020). Gender in unilingual and mixed speech of Spanish heritage speakers in the Netherlands. *Languages* 5:68. doi: 10.3390/languages5040068
- Boloh, Y., and Ibernnon, L. (2010). Gender attribution and gender agreement in 4- to 10-year-old French children. *Cogn. Dev.* 25, 1–25. doi: 10.1016/j.cogdev.2009.09.011
- Bolonyai, A. (2007). (In)vulnerable agreement in incomplete bilingual L1 learners. *Int. J. Biling.* 11, 3–23. doi: 10.1177/13670069070110010201
- Brehmer, B., and Rothweiler, M. (2012). "The acquisition of gender agreement marking in Polish," in *Multilingual Individuals and Multilingual Societies*, eds K. Braunnmuller and C. Gabriel Amsterdam: John Benjamins, 81–100. doi: 10.1075/hsm.13.07bre
- Carroll, S. (1989). Second-Language acquisition and the computational paradigm. *Lang. Learn.* 39, 535–594. doi: 10.1111/j.1467-1770.1989.tb00902.x
- Corbett, G. (1983). The number of genders in Polish. *Papers Stud. Contrast. Linguist.* 16, 83–89.
- Corbett, G., and Fraser, N. (1999). "Default genders," in *Gender in Grammar and Cognition*, eds Barbara Unterbeck and Matti Rissanen Berlin: Mouton de Gruyter, 55–97. doi: 10.1515/9783110802603.55
- Cornips, L., and Hulk, A. (2008). Factors of success and failure in the acquisition of grammatical gender in Dutch. *Second Lang. Res.* 24, 267–295. doi: 10.1177/0267658308090182
- Cuza, A., and Pérez-Tattam, R. (2015). Grammatical gender selection and phrasal word order in child heritage Spanish: A feature re-assembly approach. *Bilingualism* 19, 50–68. doi: 10.1017/S1366728914000893
- Dabrowska, E. (2006). Low-level schemas or general rules? The role of diminutives in the acquisition of Polish case inflections. *Lang. Sci.* 28, 120–135. doi: 10.1016/j.langsci.2005.02.005
- Dussias, P., Valdés Kroff, J. R., Guzzardo Tamargo, R. E., and Gerfen, C. (2013). When gender and looking go hand in hand. *Stud. Second Lang. Acquisit.* 35, 353–387. doi: 10.1017/S0272263112000915
- Egger, E., Hulk, A., and Tsimpli, I. M. (2018). Crosslinguistic influence in the discovery of gender: The case of Greek-Dutch bilingual children. *Bilingualism* 21, 694–709. doi: 10.1017/S1366728917000207
- Eichler, N., Jansen, V., and Müller, N. (2013). Gender acquisition in bilingual children: French–German, Italian–German, Spanish–German and Italian–French. *Int. J. Biling.* 17, 550–572. doi: 10.1177/1367006911435719
- Fuchs, Z. (2014). Gender and analogical extension: from animacy to borrowings in Polish. *New Insights Slavic Linguist.* 3, 115–127.
- Fuchs, Z. (2021). Facilitative use of grammatical gender in Heritage Spanish. *Linguist. Approaches Biling.* doi: 10.1075/lab.20024.fuc [Epub ahead of print].
- Fuchs, Z., Polinsky, M., and Scontras, G. (in press). "Explaining gender: Lessons from heritage Spanish," in *Multifaceted multilingualism*, Ed. K. Grohmann (Amsterdam: John Benjamins).
- Godson, L. (2003). *Phonetics of language attrition: Vowel production and articulatory setting in the speech of Western Armenian heritage speakers*. Los Angeles, CA: University of California.
- Grüter, T., Lew-Williams, C., and Fernald, A. (2012). Grammatical gender in L2: A production or a real-time processing problem? *Second Lang. Res.* 28, 191–215. doi: 10.1177/0267658312437990
- Håkansson, G. (1995). Syntax and morphology in language attrition: a study of five bilingual expatriate Swedes. *Int. J. Appl. Linguist.* 5, 153–169. doi: 10.1111/j.1473-4192.1995.tb00078.x
- Haman, E. B., Etenkowski, M., Łuniewska, M., Szwabe, J., Dąbrowska, E., Szreder, M., et al. (2011). *Polish CDS Corpus*. Available online at: <https://childes.talkbank.org/access/Slavic/Polish/Polish-CDS.html>
- Haman, E., Wodniecka, Z., Marecka, M., Szweczyk, J., Bialecka-Pikul, M., Otwinowska, A., et al. (2017). How does L1 and L2 exposure impact L1 performance in bilingual children? Evidence from Polish-English migrants to the United Kingdom. *Front. Psychol.* 8:1444. doi: 10.3389/fpsyg.2017.01444
- Haspelmath, M. (2006). Against Markedness (And What to Replace It With). *J. Linguist.* 42, 25–70. doi: 10.1017/S0022226705003683
- Hernández Pina, F. (1984). Teorías psicosociolingüísticas y su aplicación a la adquisición del español como lengua materna. (la ed. en español). Madrid: Siglo Veintiuno.
- Hockett, C. (1958). *A course in modern linguistics*. London: Macmillan. doi: 10.1111/j.1467-1770.1958.tb00870.x

- Hopp, H. (2013). Grammatical gender in adult L2 acquisition: Relations between lexical and syntactic variability. *Second Lang. Res.* 29, 33–56. doi: 10.1177/0267658312461803
- Hopp, H. (2016). Learning (not) to predict: Grammatical gender processing in second language acquisition. *Second Lang. Res.* 32, 277–307. doi: 10.1177/0267658315624960
- Hopp, H., and Lemmerth, N. (2016). Lexical and syntactic congruency in L2 predictive gender processing. *Stud. Second Lang. Acquisit.* 40, 171–199. doi: 10.1017/S0272263116000437
- Hur, E., Otero, J. C. L., and Sanchez, L. (2020). Gender agreement and assignment in Spanish heritage speakers: Does frequency matter? *Languages* 5, 1–16. doi: 10.3390/languages5040048
- Janssen, B. (2014). Frequency effects on the acquisition of Polish and Russian gender morphology. *Stud. Slavic Gen. Linguist.* 40, 109–126. doi: 10.2307/26506035
- Janssen, B. (2016). *The acquisition of gender and case in Polish and Russian: a study of monolingual and bilingual children*. Amsterdam: Uitgeverij Pegasus.
- Kaltsa, M., Tsimpli, I. M., and Argyri, F. (2019). The development of gender assignment and agreement in English-Greek and German-Greek bilingual children. *Linguist. Approaches Biling.* 9, 253–288. doi: 10.1075/lab.16033.kal
- Kaushanskaya, M., Blumenfeld, H. K., and Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism* 23, 945–950. doi: 10.1017/S1366728919000038
- Krajewski, G. (2005). “The role of grammatical gender in the acquisition of noun inflection in Polish,” in *Psychology of Language and Communication*, Vol. 9, eds K. Beattie and A. Ellis London: Psychology Press.
- Kramer, R. (2009). *Definite Markers, Phi-features, and Agreement: A Morphosyntactic Investigation of the Amharic DP*. Los Angeles, CA: University of California.
- Kramer, R. (2014). Gender in Amharic: a morphosyntactic approach to natural and grammatical gender. *Lang. Sci.* 43, 102–115. doi: 10.1016/J.LANGSCI.2013.10.004
- Kramer, R. (2015). *The Morphosyntax of Gender*. Oxford: Oxford University Press, doi: 10.1093/acprof:oso/9780199679935.001.0001
- Kuchenbrandt, I. (2005). “Gender acquisition in bilingual Spanish,” in *Proceedings of the 4th International Symposium on Bilingualism*, Bilingualism 1252–1263.
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Kupisch, T., Müller, N., and Cantone, K. F. (2002). Gender in monolingual and bilingual first language acquisition: Comparing Italian and French. *Lingue Linguaggio* 1, 107–149. doi: 10.14187/7559
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *J. Statist. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13
- Lemmerth, N., and Hopp, H. (2019). Gender processing in simultaneous and successive bilingual children: Cross-linguistic lexical and syntactic influences. *Lang. Acquisit.* 26, 21–45. doi: 10.1080/10489223.2017.1391815
- Lew-Williams, C., and Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychol. Sci.* 18, 193–198. doi: 10.1111/j.1467-9280.2007.01871.x
- Lew-Williams, C., and Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *J. Memory Lang.* 63, 447–464. doi: 10.1016/j.jml.2010.07.003
- Lieberman, A., Borovsky, A., and Mayberry, R. (2018). Prediction in a visual language: real-time sentence processing in American Sign Language across development. *Lang. Cogn. Neurosci.* 33, 387–401. doi: 10.1080/23273798.2017.1411961
- Lleó, C. (1997). “Filler syllables, proto-articles and early prosodic constraints in Spanish and German,” in *Proceedings of the GALA \$97 conference on language acquisition*, eds A. Sorace, C. Heycock, and R. Shillcock (Edinburgh: Edinburgh University Press).
- Loerts, H., Wieling, M., and Schmid, M. S. (2013). Neuter is not common in Dutch: eye movements reveal asymmetrical gender processing. *J. Psycholinguist. Res.* 42, 551–570. doi: 10.1007/s10936-012-9234-2
- López-Ornat, S. (1997). “What Lies in between a Pre-Grammatical and a Grammatical Representation? Evidence on Nominal and Verbal Form-Function Mappings in Spanish from 1;7 to 2;1,” in *Contemporary Perspectives on the Acquisition of Spanish*, Vol. 1, eds A. T. Perez-Leroux and W. R. Glass Somerville, MA: Cascadia, 3–20.
- Luczynski, E. (2002). Zagadnienie kompetencji fleksyjnej dzieci w wieku przedszkolnym [na przykładzie rzeczownika] (The issue of inflectional competence of preschool children [case of noun]). *Język Polski*. 1, 43–50.
- MacWhinney, B. (1978). The acquisition of morphophonology. *Monogra. Soci. Res. Child Dev.* 43, 1–123. doi: 10.2307/1166047
- Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing Language Profiles in Bilinguals and Multilinguals. *J. Speech Lang. Hear. Res.* 50:940. doi: 10.1044/1092-4388(2007/067)
- Mariscal, S. (1996). “Adquisiciones morfosintácticas en torno al sintagma nominal: El género gramatical en español,” in *Estudios sobre la adquisición del Castellano, Catalan, Euskera y Gallego*, ed. M. Pérez-Pereira Santiago: Universidade de Santiago de Compostela, 263–272.
- Mariscal, S. (2001). Es “a pie” equivalente a DET + N? Sobre el conocimiento temprano de las categorías gramaticales. *Cognitiva* 13, 35–59. doi: 10.1174/021435501753635578
- Mariscal, S. (2009). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *J. Child Lang.* 36, 143–171. doi: 10.1017/S030500090808908
- Meir, N., Walters, J., and Armon-Lotem, S. (2017). Bi-directional cross-linguistic influence in bilingual Russian-Hebrew children. *Linguist. Approaches Biling.* 7, 514–553. doi: 10.1075/lab.15007.mei
- Melançon, A., and Shi, R. (2015). Representations of abstract grammatical feature agreement in young children. *J. Child Lang.* 42, 1379–1393. doi: 10.1017/S0305000914000804
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press, doi: 10.1017/CBO9781139030502
- Montrul, S., and Potowski, K. (2007). Command of gender agreement in school-age Spanish-English Bilingual Children. *Int. J. Biling.* 11, 301–328. doi: 10.1177/13670069070110030301
- Montrul, S., Bhatt, R. M., and Bhatia, A. (2012). Erosion of case and agreement in Hindi heritage speakers. *Linguist. Approaches Biling.* 2, 141–176. doi: 10.1075/lab.2.2.02mon
- Montrul, S., Davidson, J., de la Fuente, I., and Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Biling. Lang. Cogn.* 17, 118–138. doi: 10.1017/S1366728913000114
- Montrul, S., Foote, R., and Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition. *Lang. Learn.* 58, 503–553. doi: 10.1111/j.1467-9922.2008.00449.x
- Mueller Gathercole, V. C. (2002). “Grammatical Gender in Bilingual and Monolingual Children: A Spanish Morphosyntactic Distinction,” in *Language and Literacy in Bilingual Children*, eds R. Oller and R. Eilers Bristol: Multilingual Matters, 207–219. doi: 10.21832/9781853595721-010
- Paolieri, D., Lotto, L., Morales, L., Bajo, T., Cubelli, R., and Job, R. (2010). Grammatical gender processing in Romance languages: Evidence from bare noun production in Italian and Spanish. *Eur. J. Cogn. Psychol.* 22, 335–347. doi: 10.1080/09541440902916803
- Pascual, Y., Cabo, D., and Rothman, J. (2012). The (Il)logical problem of heritage speaker bilingualism and incomplete acquisition. *Appl. Linguist.* 33, 450–455. doi: 10.1093/applin/ams037
- Pérez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *J. Child Lang.* 18, 571–590. doi: 10.1017/S0305000900011259
- Pine, J. M., and Lieven, E. V. M. (1997). Slot and frame patterns and the development of the determiner category. *Appl. Psycholinguistics* 18, 123–138.
- Polinsky, M. (1997). Cross-linguistic parallels in language loss. *Southwest J. Linguist.* 14, 87–123.
- Polinsky, M. (2006). Incomplete acquisition: American Russian. *J. Slavic Linguist.* 14, 191–262.
- Polinsky, M. (2008). “Relative clauses in heritage Russian: Fossilization or divergent grammar,” in *Proceedings of the Formal Approaches to Slavic Linguistics: The Stony Brook Meeting 2007*, eds A. Antonenko, J. F. Bailyn, and C. Y. Bethin (Ann Arbor, MI: Michigan Slavic Publications), 5–21.
- Polinsky, M. (2018). Bilingual children and adult heritage speakers: The range of comparison. *Int. J. Biling.* 22, 547–563. doi: 10.1177/1367006916656048
- R Core Team (2022). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Rodina, Y., Kupisch, T., Meir, N., Mitrofanova, N., Urek, O., and Westergaard, M. (2020). Internal and external factors in heritage language acquisition: evidence from heritage Russian in Israel, Germany, Norway, Latvia and the United Kingdom. *Front. Educ.* 5:20. doi: 10.3389/educ.2020.00020

- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814
- Sanchez-Sadek, C., Kiraithe, J., and Villarreal, H. (1975). The acquisition of the concept of grammatical gender in monolingual and bilingual speakers of Spanish. *Paper presented at the annual meeting of the American educational research association*, Washington DC.
- Schwartz, M., Minkov, M., Dieser, E., Protassova, E., Moin, V., and Polinsky, M. (2015). Acquisition of Russian gender agreement by monolingual and bilingual children. *Int. J. Biling.* 19, 726–752. doi: 10.1177/1367006914544989
- Scontras, G., and Putnam, M. T. (2020). Lesser-studied heritage languages: An appeal to the dyad. *Heritage Lang. J.* 17, 152–155. doi: 10.46538/hlj.17.2.2
- Scontras, G., Polinsky, M., and Fuchs, Z. (2018). In support of representational economy: Agreement in heritage Spanish. *Glossa A J. Gen. Linguist.* 3:1. doi: 10.5334/gjgl.164
- Sekerina, I. A. (2015). Predictions, fast and slow. *Linguist. Approaches Biling.* 5, 532–536. doi: 10.1075/lab.5.4.16sek
- Smoczynska, M. (1985). “The acquisition of Polish,” in *The crosslinguistic study of language acquisition*, 1, ed. D. I. Slobin Mahwah, NJ: Erlbaum, 595–686.
- Snyder, W., Senghas, A., and Inman, K. (2001). Agreement morphology and the acquisition of noun-drop in Spanish. *Lang. Acquisit.* 9, 157–173. doi: 10.1207/S15327817LA0902_02
- Soler, R. (1984). “Acquisición y utilización del artículo,” in *Estudios sobre psicología del lenguaje infantil*, ed. M. Siguan Madrid: Piramide.
- Steriopolo, O., and Wiltschko, M. (2010). “The Distributed Gender Hypothesis,” in *Formal Studies in Slavic Linguistics*, eds G. Zybatow, P. Dudchuk, S. Minor, and E. Pshehotskaya Bern: Peter Lang, 155–172.
- Stolt, S., Haataja, L., Lapinleimu, H., and Lehtonen, L. (2008). Early lexical development of Finnish children: A longitudinal study. *First Lang.* 28, 259–279. doi: 10.1177/0142723708091051
- Swan, O. (2015). Polish gender, subgender, and quasi-gender. *J. Slavic Linguist.* 23, 83–122. doi: 10.1353/jsl.2015.0001
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., and Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 1632–1634. doi: 10.1126/science.7777863
- Ticio Quesada, E. (2018). “The emergence and type of functional categories,” in *The emergence of nominal expressions in Spanish-English early bilinguals: economy and bilingual first language acquisition*, Amsterdam: John Benjamins, 75–131. doi: 0.1075/sibil.56
- Tomasello, M. (2000). The item-based nature of children’s early syntactic development. *Trends Cogn. Sci.* 4, 156–163. doi: 10.1016/S1364-6613(00)01462-5
- Tribushinina, E., and Gillis, S. (2012). The acquisition of scalar structures: Production of adjectives and degree markers by Dutch-speaking children and their caregivers. *Linguistics* 50, 241–268. doi: 10.1515/ling-2012-0009
- Tribushinina, E., van den Bergh, H., Ravid, D., Aksu-Koç, A., Kilani-Schoch, M., Korecky-Kröll, K., et al. (2014). Development of adjective frequencies across semantic classes. *Lang. Interact. Acquisit.* 5, 185–226. doi: 10.1075/lia.5.2.02tri
- Valdés, G. (2000). “Introduction,” in *Spanish for native speakers*, San Diego, CA: Harcourt College, 1–32.
- van der Linden, E., and Hulk, A. (2009). The vulnerability of gender on determiners in L1, 2L1 and L2 acquisition. *Bucharest Working Papers Linguist.* 11, 97–107.
- van Heugten, M., and Shi, R. (2009). French-learning toddlers use gender information on determiners during word recognition. *Dev. Sci.* 12, 419–425. doi: 10.1111/j.1467-7687.2008.00788.x
- Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., et al. (2022). Heritage speakers as part of the native language continuum. *Front. Psychol.* 12:717973. doi: 10.3389/fpsyg.2021.717973
- Wilson, B., Spierings, M., Ravignani, A., Mueller, J. L., Mintz, T. H., Wijnens, F., et al. (2020). Non-adjacent dependency learning in humans and other animals. *Top. Cogn. Sci.* 12, 843–858. doi: 10.1111/tops.12381



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY

Natalia Meir,
Bar-Ilan University,
Israel
Catherine Caldwell-Harris,
Boston University,
United States

*CORRESPONDENCE

Alessandra Macbeth
amacbeth@apu.edu

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 13 July 2022

ACCEPTED 12 September 2022

PUBLISHED 06 October 2022

CITATION

Macbeth A, Atagi N, Montag JL,
Bruni MR and Chiarello C (2022) Assessing
language background and experiences
among heritage bilinguals.
Front. Psychol. 13:993669.
doi: 10.3389/fpsyg.2022.993669

COPYRIGHT

© 2022 Macbeth, Atagi, Montag, Bruni and
Chiarello. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Assessing language background and experiences among heritage bilinguals

Alessandra Macbeth^{1*}, Natsuki Atagi², Jessica L. Montag³,
Michelle R. Bruni⁴ and Christine Chiarello⁴

¹Department of Psychology, Azusa Pacific University, Azusa, CA, United States, ²Department of Child and Adolescent Studies, California State University, Fullerton, Fullerton, CA, United States,

³Department of Psychology, University of Illinois, Urbana-Champaign, Champaign, IL, United States, ⁴Department of Psychology, University of California, Riverside, Riverside, CA, United States

The language backgrounds and experiences of bilinguals have been primarily characterized using self-report questionnaires and laboratory tasks, although each of these assessments have their strengths and weaknesses. The Electronically Activated Recorder (EAR), an audio recording device, has recently become more prominent as a method of assessing real-world language use. We investigated the relationships among these three assessment tools, to understand the shared variance in how these measures evaluated various aspects of the bilingual experience. Participants were 60 Southern California heritage bilingual college students who spoke a variety of heritage languages and began to learn English between the ages of 0-to 12-years. Participants completed both self-report and laboratory-based measures of language proficiency and use, and they wore the EAR for 4 days to capture representative samples of their day-to-day heritage language (HL) use. The results indicated that self-reported HL use and English age of acquisition were significant predictors of real-world language use as measured by the EAR. In addition, self-reported HL proficiency and laboratory-based HL proficiency, as measured by verbal fluency, were mutually predictive. While some variability was shared across different assessments, ultimately, none of the measures correlated strongly and each measure captured unique information about the heritage bilingual language experience, highlighting the dissociation between language experience measured at a single point in time and an accumulated life history with a heritage language. These findings may provide guidance for bilingualism researchers about which assessment tool, or combination of tools, may be best for their specific research questions.

KEYWORDS

heritage bilinguals, electronically activated recorder, self-report questionnaires, language assessments, multilingual naming test, verbal fluency

Introduction

Bilinguals regularly encounter diverse linguistic experiences in their day-to-day lives, as a function of the ability to speak their two languages in different contexts and with different interlocutors. Although much of the psycholinguistic research over the past two decades has treated bilingualism categorically (but see also Gollan et al., 2011), there has been a recent push to consider bilingualism as a continuous spectrum of dynamic experiences that uniquely affect cognition and the neural indices of brain structure and function over time (Kaushanskaya and Prior, 2015; Luk and Pliatsikas, 2016; Takahesu Tabori et al., 2018; de Bruin, 2019; DeLuca et al., 2019). This updated perspective acknowledges that a given sample of bilinguals could vary widely with respect to their levels of proficiency, frequency of use, and experience with their given languages. As such, it has become increasingly important that current studies appropriately characterize the bilingual experiences of their participants.

The two most common methods employed in psycholinguistic research on bilingualism today are self-report questionnaires of language background and laboratory tests of language proficiency. However, neither measure provides objective insight into day-to-day bilingual language use. The inclusion of such real-world data in this field is important, because it may provide unique information about a bilingual's language experience and proficiency beyond what self-report or laboratory tasks can tell us. To address this issue and capture variability in day-to-day bilingual language use, the present study employed the Electronically Activated Recorder (EAR, Mehl et al., 2001; Mehl, 2017) to record real-world language use among a linguistically diverse sample of heritage bilingual undergraduates from Southern California. Studying heritage speakers represents a unique opportunity to examine diversity among adult bilinguals.

Heritage bilinguals are individuals who typically learned a home language from birth (i.e., their heritage language) and learned the community language through immersion during childhood (Rothman, 2009; Montrul, 2016), oftentimes when they started school. Although heritage bilinguals constitute over 75% of the bilingual population in the United States (American Academy of Arts and Sciences Commission on Language Learning, 2017), psycholinguistic research on bilingualism—even in the United States—has largely overlooked this population. Heritage bilinguals are a diverse group, who may or may not have had any formal education in their heritage language (Carreira and Kagan, 2011), tend to be more dominant in the community language (e.g., Rothman and Treffers-Daller, 2014; Sanz and Torres, 2018), and may vary widely in the skill with which they use their heritage language (e.g., Kupisch and Rothman, 2018; Polinsky, 2018). For instance, although heritage bilinguals often *understand* their heritage language “very well” (American Academy of Arts and Sciences Commission on Language Learning, 2017), their skill in *speaking* the heritage language may range from minimal to being highly proficient in the language (e.g., Valdés, 2001; Polinsky and Kagan, 2007). Thus, heritage

bilinguals are characterized by a wide range of language histories, language experiences, and language skills.

To better describe and understand the range of lived language experiences of heritage bilinguals, or other bilingual speakers, we must describe and document these ranges of experiences. Existing methods largely focus on early experience with language, such as the age at which an individual first began learning a language, or a description of the languages spoken in the home and at school during childhood (e.g., Surrain and Luk, 2019). These experiences, which aim to describe early and habitual experiences with a language, may be quite different from current patterns of language experience. For example, individuals with greater amounts of early childhood experience with a language may or may not maintain that experience into adulthood, and individuals who were less deeply immersed in a language through childhood may or may not use that language frequently in adulthood (e.g., Valdés, 2001; De Houwer, 2021). Current patterns of language use may be a characteristic of the bilingual experience that is both distinct from childhood language experience as well as various measures of language proficiency. Measuring current patterns of language use, and the frequencies and contexts in which a speaker uses their multiple languages, may be a way to further develop an understanding of the wide range of bilingual experiences.

Measuring current patterns of language use also allows us to document the prevalence of specific behaviors, such as code-switching. Code-switching, or the use of two or more languages within an utterance or conversation (Gumperz, 1977), is a linguistic phenomenon common in some bilingual speech that has been found to be strongly correlated with measures of language entropy or diversity in language use (Kalmala et al., 2022). Although once believed to indicate a bilingual speaker's “confusion” between their two languages, code-switching occurs systematically and in fact demonstrates high proficiency in both languages (e.g., Poplack, 1980; Bentahila and Davies, 1983; Otheguy and Lapidus, 2003). Moreover, code-switching may reflect various cognitive strategies that bilinguals use to effectively produce speech (for review, see Beatty-Martínez et al., 2020). For example, code-switching may be a cognitive tool for bilinguals to produce hard-to-retrieve lexical items (e.g., Sarkis and Montag, 2021). Code-switching behavior therefore reflects an array of linguistic, pragmatic, and cognitive skill in bilinguals' use of their two languages.

The present study had three goals. The first was to examine the relationships among common self-report language background items, laboratory tasks of language proficiency, and real-world language use (as measured by the EAR). Second, we studied how well self-report predicts day-to-day heritage language use, and vice versa, in a sample of 60 heritage bilinguals. Third, we examined all three measures (self-report, laboratory tasks, and actual, current language use) and the partial correlations that exist between them, to further investigate where variance is—or is not—shared among these three converging but distinct measures of the heritage bilingual language experience. We also report some

exploratory analyses with self-reported and EAR-captured code-switching data, which might provide further insight into the ability of heritage bilinguals to accurately gauge their own language use. Ultimately, we hope that the findings reported herein can help future researchers decide which tool (or combination of tools) to use when assessing the language background and experiences of heritage bilinguals, depending on the types of questions they hope to answer.

A brief overview of various language assessment tools

Self-report measures

Surveys in which individuals report information about their language history, use, and proficiency are one of the most commonly used methods to assess language background. Current self-report measures that are commonly used to assess the language backgrounds of bilinguals, in particular their proficiency/use of each language, include the Language History Questionnaire (LHQ; Li et al., 2006, 2014, 2020), the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007), and the Language and Social Background Questionnaire (LSBQ; Luk and Bialystok, 2013; Anderson et al., 2018)—all of which have been found to be valid and reliable measures of language backgrounds (Li et al., 2006; Marian et al., 2007; Luk and Bialystok, 2013). In particular, the LEAP-Q is thought to provide a fairly robust assessment of language proficiency, while the LSBQ is arguably the preferred self-report measure for those interested in language use (de Bruin, 2019). Such questionnaires are also generally quick and easy to administer.

Although these surveys often vary in their operationalizations of and the exact manner in which they ask about language history, use, and proficiency (see Kaščelan et al., 2021), there are commonalities in the types of questions asked. For example, to assess language history, surveys typically include questions about when (e.g., age of acquisition), how (e.g., immersion, formal education), and from whom (e.g., family, teachers) individuals learned each language in the past. To assess language use, surveys often include items about how frequently, in what contexts, and with whom individuals currently use each language and used each language in the past. And to assess language proficiency, surveys commonly ask individuals to rate their current level of fluency in speaking, understanding, reading, and/or writing in each of their languages. The combination of these questions is intended to provide insight into the ways in which a bilingual's current language use and proficiency may differ from their past language use.

However, these assessment tools are not perfect; they often depend on ordinal measures (e.g., Likert scales), which can lead to seemingly arbitrary responses that are difficult to interpret. For example, two individuals with the same self-assigned Likert ratings could, in reality, have entirely different levels of proficiency/

use, bringing into question the inherent utility of such measurements (Sechrest et al., 1996). Further, some commonly used questionnaires differ in the range of values on their scales. For example, the LEAP-Q and LSBQ have participants rate their proficiency on a scale of 0–10 (Marian et al., 2007; Anderson et al., 2018), whereas the LHQ utilizes a 1–7 scale (Li et al., 2020). In these cases, it is unclear whether participants would rate themselves equivalently on the two versions (e.g., Would one who rates themselves as 5/7 on English speaking proficiency when given the LHQ also assign themselves a 7/10 rating, a roughly equivalent score, on the LEAP-Q?). Although one proposed solution for creating consistency across such measures has been to develop a *Bilingualism Quotient* (Marian and Hayakawa, 2021), there is currently no single agreed upon way for language experience to be operationalized or measured (cf. Kałamała et al., 2022).

Additionally, self-report measures of second language (L2) proficiency/use may be swayed by participants' own biases. Some young adults are prone to overestimation of their L2 proficiency/use (MacIntyre et al., 1997; Gollan et al., 2012), and individuals who are more anxious about their L2 abilities may be prone to underestimation of L2 proficiency/use (MacIntyre et al., 1997). Self-assessment of one's language proficiency/use might also be skewed due to lack of interaction with an appropriate comparison group. For example, one might have an inflated sense of how proficient they are in a language if they have few native speakers in their local environments to compare themselves to (Blanche, 1988; Blanche and Merino, 1989). Thus, not only is the age of the participants important to consider when interpreting self-report measures of L2 proficiency/use, but the context in which participants use a language and the range of speakers with which they interact must also be considered.

Abundant evidence also suggests that individuals from different cultural backgrounds may interpret or respond to self-report questions in systematically different ways. Tomoschuk et al. (2019) found the way in which individuals from different cultural backgrounds respond to measures of proficiency can vary widely. Spanish-English bilinguals who rated themselves as highly proficient in Spanish (e.g., 7/7) scored lower on the Multilingual Naming Test (MINT; Gollan et al., 2012)—a laboratory task of language proficiency—than their Chinese-English bilingual counterparts who rated themselves equally highly on Chinese proficiency. Among participants who considered their proficiency in Spanish or Chinese low (e.g., 3/7), the opposite effect emerged: Chinese-English bilinguals' MINT scores in Chinese were significantly lower than Spanish-English bilinguals' MINT scores in Spanish. Such findings corroborate previous work by the same group showing that Chinese-English bilinguals are generally more accurate when self-reporting their non-English language proficiency than are Spanish-English bilinguals (Sheng et al., 2014). Additionally, Hoshino and Kroll (2008) found that despite similar lab-based English proficiency (measured by picture naming), Japanese-English bilinguals self-rated their English proficiency significantly lower than Spanish-English bilinguals.

Taken together, these studies suggest that while self-report is a commonly used method in the psycholinguistics literature, culture and other participant characteristics contribute non-random error to the observed data.

Laboratory tasks

One seemingly straightforward way to test the validity of participants' self-report on language background measures would be to employ laboratory-based measures of language ability. Some common measures include picture naming tasks such as the Peabody Picture Vocabulary Test (PPVT; Dunn and Dunn, 1997), the Boston Naming Test (BNT; Kaplan et al., 1983), LexTALE (Lemhöfer and Broersma, 2012), and the MINT (Gollan et al., 2012), all of which assess vocabulary knowledge as a proxy for language proficiency and have been normed in multiple languages. Verbal fluency measures, which assess lexical knowledge, retrieval, and production as a proxy for language proficiency, are frequently employed as well (Gollan et al., 2002; Portocarrero et al., 2007; Lezak et al., 2012; Friesen et al., 2015). Past research has suggested that moderate correlations exist between self-reports of language background and these laboratory measures (Marian et al., 2007; de Bruin et al., 2017). However, less than half of the studies recently published in the psycholinguistic bilingualism field include a laboratory measure of language proficiency or fluency (Hulstijn, 2012; Surraín and Luk, 2019), and the majority of studies rely solely on self-report measures of language proficiency.

In addition, both the MINT and verbal fluency also have their own shortcomings. Measures of vocabulary such as the MINT overlook other critical components of language skill such as syntax or sentence production (Paap et al., 2017), and assume that the words used on the test are unbiased and indeed a good index of vocabulary size. Further, performance on verbal fluency tasks can be influenced by participant variables such as age or executive control abilities that are independent of linguistic knowledge (Friesen et al., 2015). Of course, neither of these tasks measure actual, real-world language use, which may differ from either laboratory tasks or the self-report measures.

The EAR

The EAR—or Electronically Activated Recorder—is a free Android app that captures naturalistic data *via* audio snippets from a participant's day-to-day life (Mehl et al., 2001; Mehl, 2017). Importantly, the language data captured by the EAR reflects spontaneous, current speech use (Mehl et al., 2012) - participants cannot track when the EAR is recording, so there are no expectancy effects regarding when or how to speak. By using this time sampling approach, the EAR provides insight into what language(s) a participant is using day-to-day, the environmental and linguistic contexts in which those languages are being used,

and the frequency with which different languages are being spoken (Macbeth et al., 2022). The EAR can provide an objective measure of an individual's frequency of language use, which is important for understanding an individual's language habits and experiences.

Because the EAR measures language use, a characteristic commonly included in surveys of language background, we can use the EAR to determine how well self-report captures objective real-world behaviors and experiences, and to assess the validity of self-report data. One such study showed that participants are quite accurate at gauging how much they participate in behaviors such as listening to music, watching TV, or talking to others (Vazire and Mehl, 2008). In another study that compared the talkativeness of Mexican versus American individuals, cultural differences in the validity of self-report data emerged: Americans rated themselves as being more talkative and sociable, despite engaging in fewer conversations, spending less time with others, and talking less than the Mexican participants (Ramírez-Esparza et al., 2009). Similarly, a more recent study by Marchman et al. (2017) that used the LENA system¹ found that Spanish-English bilingual parents systematically underestimated the use of their dominant language (Spanish) and overestimated the use of their less dominant language (English) in their self-report of child-directed speech in each language. Altogether, these findings suggest that collecting objective measures of language use can improve our interpretation of self-reported language use.

The field has developed many different means to capture aspects of an individual's language history, current language use, and language proficiency. Each of these tools has advantages and disadvantages in their ease of implementation and interpretability of the data. In the present study, we propose that the EAR may be another tool to add to this list, and that observation of naturalistic language use outside the lab may be a useful means to better understand an individual's current patterns of language use. We believe the recordings captured by the EAR may reflect a construct not well captured by existing tools—current patterns of language use need not align with past or childhood measures of language use nor with measures of language proficiency. Understanding current day-to-day patterns of language use as a qualitatively different construct may be helpful for understanding the diverse range of experiences that bilingual speakers encounter.

The current study

In the current study, we examined how each of the three assessments used to measure heritage bilingual language abilities

¹ The LENA system is a digital language recording device designed to be worn by a young child that can record up to 16h of the auditory environment. Unlike the EAR, which uses time sampling methods to record random samples of the auditory and linguistic environment, the LENA records continuously.

and use (self-report, laboratory tasks, and the EAR) are related and what shared variance is (or is not) captured among them. This will allow us to better understand the language backgrounds and day-to-day language experiences among our sample of heritage bilingual speakers.

Because we anticipated that our sample of heritage bilinguals would be highly proficient in English, we expected the most variability in heritage language (HL) proficiency and use. Further, English age of acquisition (AoA) is also important to consider: The age at which heritage bilinguals learned English might affect when and how often they use their HL (e.g., De Houwer, 2021). Moreover, AoA is commonly examined in psycholinguistic studies of bilingualism and is frequently related to language proficiency (e.g., Johnson and Newport, 1989; Birdsong, 1992; Flege et al., 1999; Hakuta et al., 2003). Therefore, the self-report variables that we focused on in the present study were HL proficiency, HL use, and English AoA. We chose laboratory tasks that are commonly used to assess language proficiency in the existing literature, so we administered both the verbal fluency and MINT to participants. Although both tasks measure productive language skill, the MINT is a standardized measure of English vocabulary that could be used with all participants, regardless of the HL spoken by participants. Finally, we measured actual day-to-day HL use *via* the EAR.

We expected all three assessment tools to be significantly correlated, but to differing extents depending on the primary construct being measured by each tool. These measures are not necessarily different assessments of a single underlying construct, but rather likely represent different constructs that may relate to each other in interesting ways. For example, the EAR is primarily a measure of language use, so we hypothesized that self-reported HL use would best predict EAR-based HL use, and vice versa. Likewise, laboratory-based measures such as verbal fluency and the MINT are generally used as assessments of language proficiency, so we expected that self-reported HL proficiency would be mutually predictive of performance on these tasks. Further, we were interested in how self-reported AoA would relate to these other measures. If a heritage bilingual acquired their second language (English) later in life, then we would expect them to be more proficient in their HL and use their HL more. We were interested in the shared (and unshared) variance captured by each of these different assessment tools and reported the partial correlations through a series of regression models to demonstrate where these measures overlapped, and where they did not. Finally, we explored self-report items related to code-switching in order to determine which one(s) best predicted actual code-switching as measured by the EAR and gauge how well heritage bilinguals can assess their own code-switching frequency. From these results, our hope is that we can begin to understand the relative contributions of each bilingualism measure and how they can collectively contribute to a comprehensive understanding of bilingual language experience.

Materials and methods

Participants

Our sample consisted of 60 heritage bilingual participants (38 women, 22 men, $M = 19.25$ years) from the University of California, Riverside, and was a subset of the participants previously reported on by Macbeth et al. (2022). In addition to English (the predominant community language), the participants knew a variety of other heritage languages. The heritage languages captured in the recordings (n included in parentheses) included Amharic (1), Arabic (1), Burmese (1), Cantonese (1), Farsi (2), Hindi (1), Igbo (1), Korean (3), Mandarin (6), Portuguese (1), Punjabi (1), Spanish (24), Teochew (1), Thai (1), and Vietnamese (6). Nine participants did not use their heritage language during the recording period. All participants were exposed to their HL from birth and acquired English between birth and age 12 years ($M = 3.57$ years). Moreover, the majority of participants reported their HL being the language they used (76.67%) and heard (88.33%) the most during their childhood prior to entering elementary school. The study was advertised through the psychology department's participant pool. Participants were given \$25 and course credit for their participation.

Characteristics of participants' language environments

Participants reported exposure to various languages in their community. Southern California is a linguistically diverse region of the United States, where heritage bilinguals may have the opportunity to be exposed to the community language, their HL, as well as other languages. Such exposure to linguistic diversity may not only provide an environment in which bilingualism is supported but may also provide linguistic experiences that shape language and cognition (Bice and Kroll, 2019; Atagi and Sandhofer, 2020). Just over half of the participants reported hearing two or more languages in the communities in which they currently reside (53.33%), with 45% of participants hearing one or more languages other than English and their HL in their current communities. Additionally, 75% of participants reported hearing two or more languages on the campus of the university they currently attend, with 60% of participants hearing one or more languages other than English and their HL on campus. Moreover, 53.33% of participants reported hearing two or more languages in the communities in which they grew up, with 48.33% of participants reporting that they heard one or more languages other than English and their HL in those childhood communities. This retrospective self-reported data was corroborated by U.S. census data: Searching census data using the ZIP codes of the residences at which participants spent the majority of their childhood, we found that on average participants grew up in communities in which only 45.60% ($SD = 20.66\%$, range: 8.80–89.50%) of the population spoke only English (U.S. Census Bureau, 2018). Thus, the heritage bilinguals in this sample not only

had exposure to a HL in their homes, but also were exposed to various languages in their communities.

Materials

Language background questionnaire

This in-house questionnaire combined a variety of questions from the Language History Questionnaire (Li et al., 2014) and the LEAP-Q (Marian et al., 2007), and assessed various aspects of language history and current language use. Participants provided demographic information and reported proficiency in both English and their HL on a scale of 1–7 (1 = Very poor, 7 = Native-like; e.g., “Please rate your current ability in speaking, reading, writing, and understanding in each language”), their age of acquisition (AoA) of each language (e.g., “For each language, enter an age for when you first became exposed to this language”), and their current exposure to and use of each language on a scale of 0–10 (0 = Never, 10 = Always, e.g., “Please rate how much you are currently using each language”). Participants were also given four scenarios adapted from the Bilingual Switching Questionnaire (Rodriguez-Fornells et al., 2012). Code-switching frequency was measured on a scale of 1 (Never) to 5 (Always).

Laboratory tasks

Verbal fluency

Participants were presented with a category (e.g., vegetables) on a computer screen and asked to name as many examples of that category as they could within 1 min. Category names were always presented in English, but on half of the trials, participants provided examples of the category in their HL. After a practice trial (colors), participants completed two blocks consisting of four English trials and four HL trials. The categories of clothing, drinks, sports, and vegetables always appeared together in the same block (Block A), and the categories of furniture, modes of transportation, fruits, and words associated with the beach were always presented in the same block (Block B), though the categories appeared in a random order for each participant. The blocks were counterbalanced such that half of the participants received Block A first, and half received Block B first. Further, half of the participants completed Block A in English, and half completed Block A in their HL, and then Block B was completed in the language that was not used for Block A. A total score for each language, a proxy of language proficiency, was created from the sum of all valid responses given for each of the four categories. Similar procedures have been used in past studies (e.g., Gollan et al., 2002; Linck et al., 2009; Baus et al., 2013).

Multilingual naming test

In the MINT (Gollan et al., 2012), pictures of objects were presented on a computer screen one at a time and remained until the participant made a verbal response, or a maximum of 5 s. Participants were instructed to say the name of each object in

English, or if they did not recognize or know the name of the object, to say, “I do not know.” Most participants who identified as Spanish-English bilinguals also completed the MINT in Spanish ($n = 18$). There were five practice trials and 68 test trials. The total score on the test trials has been used in previous research as a measure of language proficiency and vocabulary knowledge (e.g., Tao et al., 2015).

The EAR

The “EAR on Android” app was downloaded from the Google Play store onto Motorola Moto E 2nd generation phones. Settings such as recording duration (length of recording) and interval (time between recordings) can be manually adjusted by the experimenter. In the current study, the EAR was set to record for 40 s every 12 min, with a six-hour blackout period at night based on when participants self-reported their typical bedtime. A more detailed description of the EAR and its utility for psycholinguistics-related research can be found in Macbeth et al. (2022).

Procedure

Data was collected in two waves. While the procedures across both waves were quite similar, the laboratory tasks (verbal fluency and MINT) were only completed by participants in Wave 2 ($n = 38$). The Spanish MINT was only completed by a Spanish-speaking subset ($n = 18$) of these Wave 2 participants. The Language Background Questionnaire and stimuli for verbal fluency and the MINT were presented on a Dell Precision 3,420 computer running Windows 7 Professional, and recordings of verbal responses on verbal fluency and the MINT were captured using a Marantz Professional PMD-561 handheld recorder. Visual stimuli for verbal fluency and the MINT were presented *via* E-Prime 2.0, and questionnaire data was collected through Qualtrics. All instructions, tasks, and questionnaires were conducted in English, except for the HL verbal fluency and Spanish MINT trials. For HL verbal fluency, instructions and category cues were given in English, and participants were asked to respond in their HL. For the Spanish MINT, instructions were presented in Spanish. Participants came to the lab for two data collection sessions, one before and one after the 4 days of EAR audio recording.

Session 1

Participants were informed about the nature of the study, what types of sounds the EAR is designed to record, and information about the recording duration and interval. They were asked to wear the EAR as much as they were comfortable with, including locations such as home, school, in class, and other public places like a park or mall. The only location participants were told to not wear the EAR was at work, to avoid potential conflict with employers.

After confirming understanding of the recording procedures, participants were consented. They then completed the MINT. The

EAR was programmed to begin recording immediately after the end of the testing session, and the recordings ended when the participant went to bed or at midnight on the fourth day of recording, whichever came first.

Interim recording period

For 4 days, either from Thursday–Sunday or Friday–Monday, the participants went about their daily lives while wearing the EAR. They could choose to wear the EAR either clipped to their waist or in an armband. Participants were, in general, quite compliant (Macbeth et al., 2022) and wore the EAR during approximately 80% of their waking hours.

Session 2

Participants returned the EAR on the day after recording was completed. While their audio files were uploaded to a secure server, participants completed a series of questionnaires including the Language Background Questionnaire. Following the self-report measures, participants completed the verbal fluency task. Once finished, participants were debriefed and compensated for their time and participation.

Data coding

The laboratory tasks

Participants' responses on verbal fluency and the MINT were audio-recorded and later coded in the laboratory by research assistants who were heritage speakers of those languages. For example, Spanish-English heritage bilingual research assistants coded Spanish MINT data. Moreover, because research assistants who worked on this study came from the same, linguistically diverse university from which participants were also recruited, all HL verbal fluency were coded by research assistants who were also heritage bilinguals of those HLs.

The EAR

Audio files were coded by at least two research assistants who spoke the languages contained in the audio files. Due to privacy and ethical considerations, only participants' speech—not conversation partners' speech—was examined. Detailed descriptions of coding procedures, as well as ethical considerations, can be found in Macbeth et al. (2022).

We coded the proportion of audio files in which participants spoke their HL out of their total number of audio files with speech. This proportion serves as an approximate measure of the amount of time engaged in day-to-day HL use and will be referred to hereafter as “EAR-based HL use.” Additionally, we coded the proportion of files in which participants code-switched, defined here as speaking both English and HL within the audio file, out of their total number of files with speech. Although this measure of code-switching does not make fine-grained distinctions among different types of code-switching, this measure captures instances of both inter- and intrasentential code-switching and serves as an approximate measure of

code-switching frequency; this measure will be referred to hereafter as “EAR-based code-switching.”

We chose to use the proportion of audio files with HL speech for each participant rather than the raw number of audio files with HL speech because we previously found that though the two values are strongly correlated, proportional, rather than absolute, predictors had greater variability which made them statistically more effective predictors (Macbeth et al., 2022). We chose to use file counts rather than word counts because these two values are also strongly correlated, but file counts allow researchers the option to forgo data transcription (a time-consuming process) and instead only code the language spoken in audio files. Only coding the language spoken in files also means that researchers do not have to develop a standard system of counting words across the various heritage languages that may be recorded by the EAR - this is beneficial since word boundaries vary across languages, and concepts may be represented by differing numbers of words across different languages (Macbeth et al., 2022).

Results

Sample characteristics

Table 1 reports descriptive statistics related to self-reported proficiency, AoA, and current use of the participants' two languages. Participants were generally English dominant or balanced bilinguals, and on average, they learned their HL at a significantly younger age than they did English.² They reported greater use of English in their day-to-day lives.

Self-reported code-switching was also fairly common. In general, participants rated themselves as moderate switchers during conversations ($M = 3.20$, $SD = 1.07$), in certain situations ($M = 3.57$, $SD = 1.08$), when discussing certain topics ($M = 3.07$, $SD = 1.30$), and also as moderate language mixers ($M = 3.28$, $SD = 1.20$).

Additionally, the laboratory measures of language proficiency were consistent with the self-report measures in that participants were highly proficient in English (see Table 1). Participants consistently performed better, on average, on the verbal fluency and MINT tasks in English compared to their HL.

There was variability in EAR-based measures of language use. Participants produced an average of 208.15 audio files ($SD = 61.64$, range = 44–318), of which an average of 75.37 audio

² The mean heritage language age of acquisition is not age 0; this reflects the fact that some participants self-reported simultaneously being exposed to their two languages at an age older than 0 (e.g., age 3). We left the data as is to reflect the participants' perceptions of when they acquired their languages, even though it is highly unlikely that the individual received no language input until several years after birth.

TABLE 1 Means, standard deviations, and *t*-tests for self-report items from the language background questionnaire, scores on laboratory-based proficiency tasks, and EAR-based measures.

	English		Heritage language		
	Mean (SD)	Range	Mean (SD)	Range	<i>t</i> -value
Self-Report Items					
Overall	6.73		5.43		
Proficiency	(0.53)	5–7	(1.31)	3–7	6.66*
Speaking	6.75		5.83		
	(0.63)	4–7	(1.11)	3–7	5.39*
Reading	6.78		5.02		
	(0.49)	5–7	(2.04)	1–7	6.32*
Writing	6.52		4.55		
	(0.81)	4–7	(2.03)	1–7	6.42*
Understanding	6.85		6.33		
	(0.44)	5–7	(0.88)	4–7	4.31*
Age of	3.57		0.94		
Acquisition	(2.89)	0–12	(1.39)	0–5	7.13*
Language Use	9.58		6.20		
	(1.25)	3–10	(2.44)	1–10	9.96*
Laboratory Tasks					
Verbal Fluency	50.61		29.91		
	(12.45)	25–87	(11.76)	10–72	7.53*
MINT	57.16		32.61 [†]		
	(5.42)	37–64	(10.32)	16–53	9.97*
EAR-Based Measure					
Proportion of	0.92		0.16		
Audio Files	(0.19)	0.13–1.00	(0.23)	0.00–0.98	14.25*

Proficiency was rated on a scale of 1–7. Overall Proficiency is the average of self-reported proficiency ratings for speaking, reading, writing, and understanding. Age of acquisition was reported in years. Rates of one's language use were rated on a scale of 0 (never)–10 (always). Proportion of audio files reflects the proportions of the number of audio files containing speech in each language out of the total number of audio files with speech. All comparisons between English and HL were statistically significant.

* $p < 0.001$.

[†]Indicates that the HL MINT only includes scores from participants who completed the MINT in Spanish ($n = 18$). The corresponding *t*-test only compares the English and Spanish scores from those 18 participants.

files contained speech ($SD = 33.65$, range = 15–169). Significantly more audio files contained English speech ($M = 71.23$, $SD = 36.09$, range = 2–168) than speech in the HL ($M = 9.63$, $SD = 13.41$, range = 0–88), and very few audio files contained code-switching ($M = 5.50$, $SD = 6.23$, range = 0–25)—that is, audio files that contained both English and the HL speech within a single audio file.

As captured by the EAR, our sample spoke in 37.1% ($SD = 13.7\%$, range = 8.2–76.1%) of their valid recorded audio files, on average. A file was considered valid if the participant was awake and wearing the EAR during the recording. Participants used their HL in 15.9% ($SD = 22.9\%$, median = 7.7%, range = 0–97.8%) of speech files, and code-switching occurred in 7.7% ($SD = 8.5\%$, median = 4.9%, range = 0–31.0%) of speech files. Of our participants, 15.0% never used a HL during the recording period, and 23.3% never code-switched.

TABLE 2 A correlation matrix showing the relationships between the self-report items, participants' scores on laboratory tasks, and day-to-day HL use derived from the EAR.

	1. HL Ov Prof	2. HL Use	3. Eng AoA	4. HL VF	5. Eng MINT
Self-Report Items					
1. HL Overall Proficiency					
2. HL Use	0.42*				
3. English AoA	0.27	0.44*			
Laboratory Tasks					
4. HL Verbal Fluency	0.57*	0.46*	0.04		
5. English MINT	−0.26	−0.31	−0.29	−0.22	
EAR-Based Measure					
6. HL Use	0.39*	0.51*	0.43*	0.60*	−0.65*

$N = 60$ for all correlations except those with English MINT ($n = 38$) and HL verbal fluency ($n = 35$).

*Indicates significance at a Bonferroni-corrected value of p of 0.008.

Relationships between self-report, laboratory tasks, and EAR

First, we conducted a series of correlations to examine the zero-order relationships between our measures (see Table 2). The goal of these analyses was to better understand individual characteristics that contribute to variable use of the HL. A series of noteworthy relationships emerged. Among our three self-report variables, self-reported HL use was moderately and positively related to both self-reported HL overall proficiency—the average of self-reported proficiency ratings for speaking, reading, writing, and understanding—and English AoA. The more a heritage speaker uses their HL, the more proficient they report themselves to be in their HL. Further, the more a heritage speaker uses their HL, the older they were when they learned English, suggesting they likely had more sole exposure to, and use of, their HL prior to English being introduced. However, self-reported English AoA and HL overall proficiency were not related to each other, which suggests that the age at which one acquired English has less bearing on their reported HL proficiency.

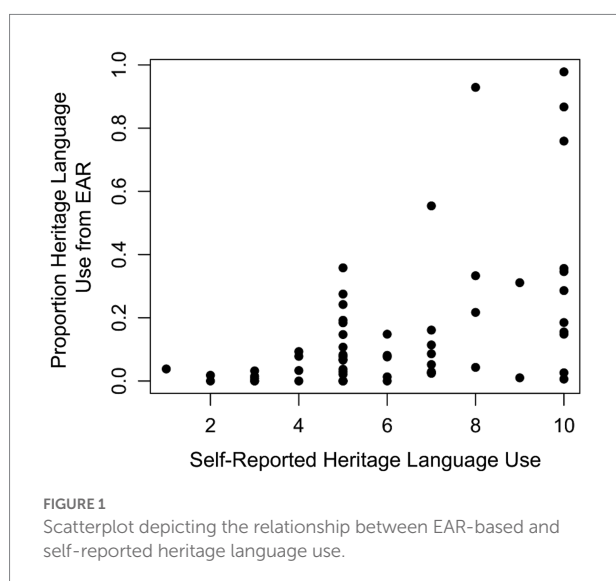
Among our laboratory-based tasks, we focused on HL verbal fluency and the English MINT. While HL performance on the MINT would arguably be more interesting given the greater variability in scores as compared to English performance, we did not examine the Spanish MINT due to the small number of participants who were able to complete this task ($n = 18$). The relationship between HL verbal fluency and English MINT was not significant, which suggests that a heritage speaker's HL and English abilities are independent of one another.³ Interestingly, HL verbal fluency was related to

³ To double-check this assumption, we also correlated HL verbal fluency and English verbal fluency scores, and found no relationship ($r = 0.12$,

self-reported HL overall proficiency, suggesting that heritage speakers are fairly, though not perfectly, accurate at assessing their skill in their HL. Further, HL verbal fluency was also related to self-reported HL use, indicating that the objective ability one displays in their HL appears to predict how much one uses the HL.

EAR-based HL use was significantly correlated with all the other self-report items and laboratory tasks. Of the self-report items, the strongest relationship existed between EAR-based HL use and self-reported HL use, as shown in Figure 1. Our self-report measure of HL use asked participants to rate how often they used their HL on a 0–10 scale from “never” to “always” so these ratings cannot be interpreted as proportion of time spent speaking, or proportion of utterances produced in the HL. Though the absolute self-report values may not map on to the utterance proportion values derived from the EAR, participants’ self-report values nonetheless provide a useful assessment of the prevalence of the HL in their lives. We expect that individuals who self-report higher values have greater exposure to their HL and individuals who report lower values have less exposure. The correlation between self-reported and EAR-based HL use is noteworthy because while both assessments claim to be measuring a similar construct, only about 25% of the variance in one variable is accounted for by the other. We see some, though imperfect, alignment between self-reported and EAR-derived estimates of HL use.

$p=0.497$), again suggesting that as a heritage bilingual, being proficient in one language does not guarantee better or worse proficiency in the other language. However, English verbal fluency scores and English MINT scores were significantly related ($r=0.50$, $p=0.001$), suggesting that both English verbal fluency and English MINT captured participants’ language skill in English.



Interestingly, it is evident from Figure 1 an asymmetry between self-reported and EAR-based language use emerged. Those who self-reported infrequent use of the HL tended to indeed use that language infrequently relative to other participants, but there was a great deal of variability among individuals who self-reported frequent use of the HL. These participants may have used the HL frequently, but they also may have hardly used it at all, as evidenced by the large vertical spread of scores on the right side of Figure 1. While the EAR might not have captured all possible instances of the participants using (or not using) their HL, such findings suggest that self-report is not entirely reliable on its own as an assessment of language use, particularly for individuals who self-report high rates of HL use.

EAR-based HL use was also strongly related to HL verbal fluency and English MINT, as shown in Figure 2. Participants who used their HL more often had higher HL verbal fluency scores and lower English vocabulary scores. Upon further examination of the data, there was one participant with a HL verbal fluency score greater than 3 standard deviations above the mean, and one other participant with an English MINT score greater than 3 standard deviations below the mean. After removing these two participants from the analysis, both the relationship between EAR-based HL use and HL verbal fluency as well as the relationship between EAR-based HL use and English MINT were weakened. These results suggest that HL verbal fluency and laboratory-based English proficiency are moderately correlated with EAR-based HL use. For the remainder of the analyses, HL verbal fluency and English MINT do not include these two outlier data points.

Predicting EAR-based HL use from self-report

The correlation matrix in Table 2 showed that multiple self-report items correlated with each other and with EAR measures of actual language use, so we aimed to better understand whether these correlations account for different or shared sources of variance. Because not all participants completed the laboratory tasks, we were first interested in the effects of self-report variables on their own, across the entire sample, prior to examining models that included all three instruments. For example, if both self-reported HL use and overall proficiency predict EAR-based HL use, do these two predictors account for similar variability within EAR-based HL use? Likewise, to what extent could one consider the two assessments to capture similar variance such that they could hypothetically be used interchangeably? To answer these questions, we conducted a multiple regression to determine which of these self-reported language background variables best predicted EAR-based HL use. All variables were z -scored prior to being entered in the model.

The simultaneous regression model included self-reported HL overall proficiency, HL use, and English AoA as predictors of EAR-based HL use. The model was significant, $F(3, 56)=9.72$, $p<0.001$, $R^2=0.34$, with only self-reported HL use serving as a

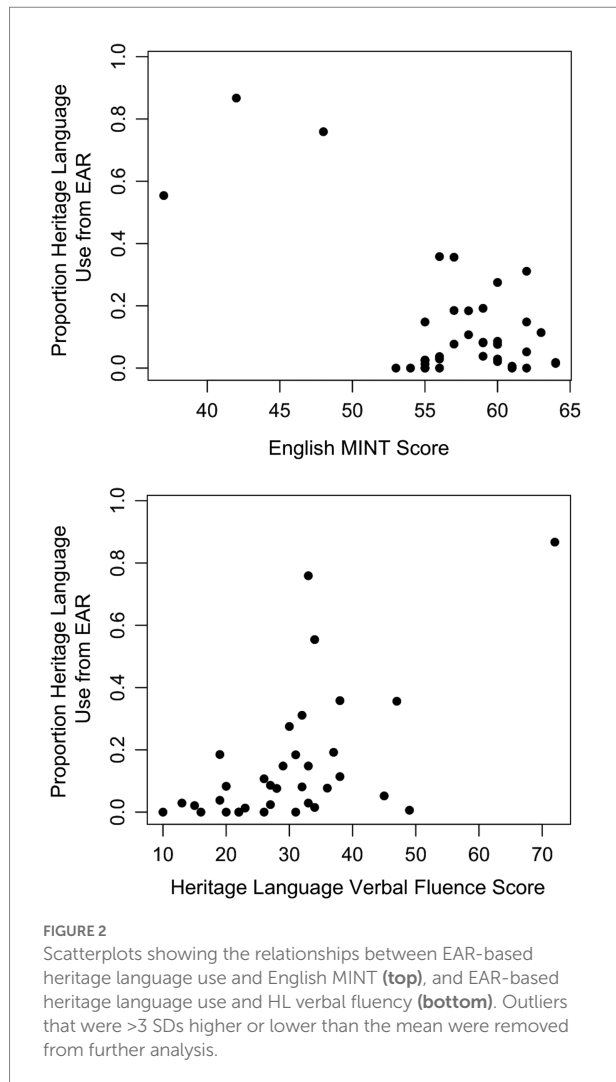


TABLE 3 Multiple regression analysis predicting EAR-based heritage language use from self-reported language background variables.

Self-Report Predictors	B	t-value	p-value	r	r _{partial}
HL Overall Proficiency	0.18	1.54	0.13	0.39	0.20
HL Use*	0.33	2.57	0.01	0.51	0.33
English AoA	0.24	1.96	0.06	0.43	0.25

Significant predictors are noted with an asterisk (*).

significant predictor of EAR-based HL use, as shown in Table 3 (although English AoA was marginally significant). In this model, much of the variance in EAR-based HL use is left unexplained by self-report, suggesting that the EAR is providing substantial unique information about real-world language use of heritage bilinguals that self-report is unable to capture.

Further, upon examining the partial correlation for each predictor (see Table 3), there is substantial shared variance among the self-report items. Without self-reported HL overall proficiency

and English AoA, the relationship between self-reported HL use and EAR-based HL use grows much weaker, going from $r=0.51$ to 0.33. The amount of HL use by a given individual appears to be partly predicted by how proficient they are in their HL, as well as when they began learning English. Thus, while self-report is overall not an entirely reliable means of capturing current language use, using multiple items to predict real-world language use appears to be beneficial.

Examining shared variance among self-report, laboratory tasks, and the EAR

Finally, we wanted to examine which predictors were contributing unique or overlapping variance among all three of our language assessments. For these analyses, we focused on our smaller subsample that completed the laboratory tasks in addition to the self-report and EAR portions of the study ($n=35$). Our hope was to provide further insight on what self-report, laboratory tasks, and the EAR tell us individually, and as a set, in the context of understanding the language experiences of heritage bilinguals. Self-reported HL overall proficiency, HL use, and English AoA were again the self-report variables of interest, HL verbal fluency was used as the laboratory-based variable of interest, and EAR-based HL use was our experience-based variable. The English MINT was not considered in these analyses because there was more variability in HL verbal fluency scores, and we were ultimately more interested in the relationships between self-reported and lab-based HL proficiency measures.

We used a series of simultaneous regression models to examine the shared and unshared variance among our constructs (see Table 4). The first three regression models examined how well HL verbal fluency and EAR-based HL use predicted each of our three self-reported outcome variables. The other self-report items were not included as predictors in the first three models (e.g., self-reported English AoA and HL use were not included as predictors of self-reported HL overall proficiency), because our research questions were inherently more focused on shared variance between assessments as opposed to shared variance among items on a single assessment. In Model 1, we found that HL verbal fluency significantly predicted self-reported HL overall proficiency, but EAR-based HL use did not, $F(2, 31)=8.43$, $p=0.001$, $R^2=0.35$. These findings suggest that a heritage bilingual's lab-based proficiency is consistent with how they self-report their HL proficiency. Further, the lack of relationship between HL verbal fluency and EAR-based HL use makes sense; the EAR measure of HL use is not a measure of language proficiency but rather day-to-day language use. It is not necessarily the case that the most proficient HL speakers tend to use their HL most frequently. Day-to-day language use and language proficiency are not the same construct and may not even be strongly related.

In Models 2 and 3, only EAR-based HL use significantly predicted self-reported HL use, $F(2, 31)=5.74$, $p=0.01$, $R^2=0.27$,

TABLE 4 Multiple regression analyses examining sources of shared and unshared variance between different predictors.

Predictor	B	t-value	p-value	r	r _{partial}
Model 1: Self-Reported HL Overall Proficiency					
HL Verbal					
Fluency*	0.54	3.05	0.005	0.56	0.48
EAR-Based HL					
Use	0.27	1.46	0.15	0.40	0.25
Model 2: Self-Reported HL Use					
HL Verbal					
Fluency	0.27	1.41	0.17	0.37	0.25
EAR-Based HL					
Use*	0.49	2.37	0.02	0.47	0.39
Model 3: Self-Reported English AoA					
HL Verbal					
Fluency	-0.27	-1.40	0.17	0.06	-0.24
EAR-Based HL					
Use*	1.00	5.02	< 0.001	0.64	0.67
Model 4: HL Verbal Fluency					
Self-Reported					
HL Overall					
Proficiency*	0.40	2.90	0.007	0.56	0.47
Self-Reported					
HL Use	0.19	1.35	0.19	0.37	0.24
Self-Reported					
English AoA*	-0.29	-2.06	0.05	0.06	-0.36
EAR-Based HL					
Use	0.34	1.65	0.11	0.36	0.29
Model 5: EAR-Based HL Use					
Self-Reported					
HL Overall					
Proficiency	0.04	0.29	0.78	0.40	0.05
Self-Reported					
HL Use	0.07	0.60	0.56	0.47	0.11
Self-Reported					
English AoA*	0.41	3.82	< 0.001	0.64	0.58
HL Verbal					
Fluency	0.25	1.65	0.11	0.36	0.29

The outcome variable for each regression model is bolded with predictors listed below. Significant predictors for each regression model are noted with an asterisk (*).

and English AoA, $F(2, 31) = 12.67$, $p < 0.001$, $R^2 = 0.45$, while HL verbal fluency was not a significant predictor in either model. This implies that heritage bilinguals' lab-based proficiency in their HL is not a predictor of their self-reported HL use, nor is it a predictor of the estimated age at which they acquired English. In this case, an experience-based measure of day-to-day language use is a better predictor of participants' estimate of their self-reported HL use and perhaps more surprisingly, AoA. In fact, it appears that in the model predicting English AoA, the zero-order relationship between English AoA and EAR-based HL use was statistically suppressed, as evidenced by the partial correlation between the two variables. This suggests that controlling for some of the

random variability in other variables within the regression model, such as HL verbal fluency, actually strengthens the relationship between English AoA and actual, EAR-based HL use.

Next, we sought to determine which self-report and experience-based variables best predicted HL verbal fluency scores. Model 4 shows that two self-report variables, HL overall proficiency and English AoA, are significant predictors of HL verbal fluency, $F(4, 29) = 5.47$, $p = 0.002$, $R^2 = 0.43$, while self-reported HL use and EAR-based HL use are not significant predictors. Again, this makes sense because day-to-day HL use, whether it is assessed through self-report or the EAR, is not the same underlying construct as language proficiency. Again, we see a dissociation between current day-to-day language use and measures of language proficiency such that it is not necessarily the most proficient speakers who currently use their heritage language most often. Therefore, if researchers are interested in approximating a participant's language abilities from self-report, the participant's self-reported proficiency and AoA may be the most indicative items. Interestingly, suppression was also present in this model for English AoA, which implies that the relationship between verbal fluency and AoA is typically obscured by the shared variance between AoA and other self-report items.

Finally, Model 5 predicted EAR-based HL use from self-report and HL verbal fluency. English AoA was the only variable to significantly predict EAR-based HL use, $F(4, 29) = 8.10$, $p < 0.001$, $R^2 = 0.53$. Importantly, self-reported HL use was not an important predictor of EAR-based HL use, even though the opposite was true in Model 2 (see Table 4). This suggests that much of the variance in self-reported HL use and English AoA is shared, with the partial correlation between self-reported HL use and EAR-based HL use growing much weaker (from $r = 0.47$ to 0.11) after controlling for English AoA and the other proficiency-related variables. Therefore, if a heritage bilingual has a good estimation of when they began speaking their non-heritage language (English in this case) and if other sources of variance (e.g., proficiency, use) are also accounted for, AoA may relate most strongly to day-to-day HL use, if such experience-based measures such as the EAR are not feasible or available.

Exploratory code-switching analyses

Next, as an exploratory measure, we examined how well self-reported code-switching frequency captured the variability in EAR-based code-switching. The self-reported tendency to switch during a conversation was significantly correlated with EAR-based code-switching, $r(58) = 0.52$, $p < 0.001$. In addition, switching more during particular situations, $r(58) = 0.27$, $p = 0.04$, switching more when discussing certain topics, $r(58) = 0.30$, $p = 0.02$, and mixing more frequently, $r(58) = 0.30$, $p = 0.02$ were also moderately related to EAR-based code-switching. When these four variables (z-scored) were entered into a simultaneous regression model, the model significantly predicted EAR-based code-switching frequency, $F(4, 55) = 5.24$, $p = 0.001$, $R^2 = 0.28$. However, the self-reported tendency to switch during a conversation was the only

significant predictor of EAR-based code-switching ($B = 0.60$, $p < 0.001$). Therefore, of the four self-reported code-switching questions, self-rated frequency of switching was selected as the item that best served as a proxy for actual EAR-based code-switching and is the predictor we use and report in subsequent analyses.

We also wondered how frequency of HL use related to code-switching frequency. In other words, if someone uses their HL a lot, are they more likely to code-switch, or are these two aspects of language use independent? Interestingly, the relationship between self-reported HL use and self-reported code-switching frequency [$r(58) = 0.30$, $p = 0.02$] as well as the relationship between EAR-based HL use and EAR-based code-switching [$r(58) = 0.60$, $p < 0.001$] were both significant. Together, these findings suggest that participants who use their HL more often also code-switch more frequently. However, it is possible that the EAR may provide better estimates of “true” code-switching frequency compared to self-report: Participants may not be fully aware of how much or how little they code-switch and have a harder time estimating that for themselves. It should be noted that like the relationship between EAR-based HL use and self-reported HL use, EAR-based measures of code-switching frequency reflect a true proportion of utterances containing code-switching, whereas self-reported code-switching frequency is a rating scale that may not directly map onto the true proportion of code-switched utterances as measured by the EAR.

We then asked whether any other self-report or laboratory measures of language proficiency that we examined previously (e.g., self-reported HL proficiency, HL use, English AoA, or HL verbal fluency) would aid in predicting EAR-based code-switching frequency. The regression model was significant, $F(5, 28) = 9.73$, $p < 0.001$, $R^2 = 0.64$. In addition to self-reported code-switching frequency, English AoA and HL verbal fluency were also significant predictors of EAR-based code-switching (see Table 5). Neither self-reported HL proficiency nor self-reported HL use were predictors of EAR-based code-switching. These results suggest that, in general, how well a bilingual believes they know one of their languages or how often they use it, is independent of switching frequency. However, acquiring English at an older age and being more proficient in the HL are associated with more frequent code-switching.

Discussion

The goals of this study were threefold: to examine the relationships among three different measures of a heritage bilingual’s language background, to determine how well self-report measures predict real-world HL use, and to investigate the extent to which various self-report items, laboratory tasks, and objective assessments of day-to-day language use “hang together” and serve as mutually predictable information about a heritage bilingual’s linguistic experiences. Further, we were also interested in how well heritage bilinguals can gauge their own frequency of

TABLE 5 Multiple regression analysis predicting EAR-based code-switching frequency from self-report and laboratory-based variables.

Predictor	<i>B</i>	<i>t</i> -value	<i>p</i> -value	<i>r</i>	<i>r</i> _{partial}
Self-Reported Code-Switching*	0.50	4.19	<0.001	0.63	0.58
Self-Reported HL Proficiency	−0.10	−0.62	0.54	0.31	−0.12
Self-Reported HL Use	−0.02	−0.15	0.89	0.38	−0.03
Self-Reported English AoA*	0.43	3.14	0.004	0.55	0.51
HL Verbal Fluency*	0.45	2.38	0.03	0.41	0.41

Significant predictors are noted with an asterisk (*).

code-switching, and whether other measures of bilingual proficiency or use can aid in predicting real-world code-switching tendencies above and beyond assessing code-switching through self-report.

We generally found moderate to strong relationships between the self-report items, laboratory tasks, and EAR-based measure of HL use, suggesting that they are all assessing similar, though not entirely overlapping, constructs. An interesting trend that emerged was that in reflecting upon one’s own language use, many heritage bilinguals tended to overestimate *via* self-report how frequently they used their HL, in comparison to how frequently they actually used their HL as measured by the EAR.

These findings are in line with past studies that have shown that young adults tend to show an enhancement bias (MacIntyre et al., 1997; Gollan et al., 2012), which might lead to overestimation of their HL proficiency and by proxy, their HL use too. Further, nearly half of our sample consisted of Spanish-English bilinguals, who have been shown to be less accurate in self-rating their proficiency compared to other cultural groups (Sheng et al., 2014), particularly when they are less proficient in Spanish compared to English (Tomoschuk et al., 2019). Overestimation of the less dominant language—in the case of the present study, the HL—is also consistent with past work by Marchman et al. (2017).

Our second set of results demonstrated that out of three commonly used self-report items (overall proficiency, frequency of language use, and AoA), self-reported frequency of HL use was a significant predictor—and English AoA, a marginally significant predictor—of day-to-day EAR-based HL use. However, when examining our simultaneous regression model involving all three measures (self-report, laboratory, and EAR-based measures), it was evident that English AoA accounted for the most unique variance of EAR-based HL use, above and beyond self-reported HL use. These two models likely yielded different results due to the inclusion of the laboratory task (i.e., HL verbal fluency) in the latter model: HL verbal fluency was significantly correlated with both self-reported overall proficiency and self-reported HL use, likely “soaking up” the variance associated with those two

self-reported measures. Thus, among heritage bilinguals, the age at which one acquired the majority, community language (in this case, English) appears to be particularly important in understanding the frequency with which one uses their HL in everyday life.

It is intuitive that we found older English AoA to be coupled with more frequent use of the HL. The heritage bilingual likely had greater practice with the HL and greater exposure across their lifetime to other speakers of their HL. Moreover, there may be characteristics of heritage bilinguals who are later-learners of the community language that are associated with more frequent HL use (e.g., their family members may speak the community language less). This may be one reason why we found English AoA—rather than self-reported HL use—to be a unique predictor of real-world HL use. English AoA in this study may be indexing aspects of a heritage bilingual's language history that is not captured in other self-report measures but is relevant for how the heritage bilingual currently uses their HL. For example, English AoA here may be tapping into the ways in which a heritage bilingual's family uses English and the HL: If family members used the HL more frequently in the past, it is possible that the HL may still be spoken more frequently with family members. Such language history characteristics may not be captured by self-reports of current language proficiency or use. Another reason might be that self-reporting one's AoA is more objective than self-reporting HL use. When asked about AoA, a bilingual may be able to recall some milestone in their lives associated with acquisition of the given language (e.g., immigrating to a new country or starting school). However, it is arguably more difficult to gauge the frequency with which you use a language because there are many situational or contextual factors (e.g., who you are with, where you are, or what you are doing) that may influence the amount of a language used on any given day. This may be another reason why the frequency of HL use is difficult to self-report. Therefore, AoA may be easier for participants to report and may capture other characteristics of heritage bilinguals' language experiences that make AoA an informative indicator of real-world language usage.

Turning to measures of proficiency, self-reported HL proficiency and HL verbal fluency appear to be mutually predictive, and one's self-reported proficiency in their HL was not indicative of how often they used the HL day-to-day. These results suggest that proficiency and frequency of use of a given language are separable constructs and largely independent of one another, a finding that is consistent with past research (Gollan et al., 2015). Just because someone is highly proficient in a HL—perhaps as a function of past immersion or exposure—does not mean that the language is being used often in the current context being captured. On the other hand, while previous studies have posited that laboratory-based measures are not the best assessments of linguistic skill (Friesen et al., 2015; Paap et al., 2017), our results with heritage bilinguals are in line with other studies (e.g., Marian et al., 2007; Shi, 2011, 2013) that show that self-reported language

proficiency and laboratory measures of language proficiency are moderately correlated and may explain similar variation in language proficiency.

With regard to the exploratory code-switching analyses, only the self-report item that asked participants about their frequency of code-switching during a conversation significantly predicted EAR-based code-switching. It seems that self-report items which ask about the contextual aspects of code-switching, such as whether particular situations or topics may induce more code-switching, do not predict EAR-based code-switching frequency above and beyond what self-reported conversational switching frequency tells us. It is not clear whether the poor predictive power of the contextual or situational effects of code-switching stem from participant's challenge to accurately report these behaviors, or if these contextual and situational effects genuinely do not relate to overall code-switching frequency. Additionally, it should be noted that the measure of EAR-based code-switching frequency reported here was a proportion of audio files containing speech in both English and the HL in a single audio file. It is therefore possible that a more fine-grained analysis of EAR-based code-switching (e.g., examining transcriptions of code-switched speech) may show real-life code-switching to be predicted by self-report items about the contextual or situational aspects of code-switching.

Another interesting result that emerged from the code-switching data was the finding that EAR-based code-switching and EAR-based HL use were more strongly related to one another than self-reported code-switching and self-reported HL use. Such findings support past work suggesting that bilinguals are often not aware of when they code-switch (Gumperz, 1982), which may be influencing the strength of the relationship between self-reported code-switching and self-reported HL use. Since the EAR-assessed measures of code-switching and HL use are arguably more objective, these results suggest that self-reported code-switching frequency might not be a strong proxy for real-world code-switching, and that the EAR may be more accurate at gauging such behavior.

Further, we found that English AoA and HL verbal fluency predicted code switching such that later English AoAs and higher HL verbal fluency scores were associated with more code-switching. Both later English AoA and higher HL verbal fluency scores are associated with greater proficiency in the HL. The finding that HL proficiency *positively* predicts real-world code-switching is consistent with work emphasizing that code-switching is used by bilinguals who are highly proficient in their two languages, rather than individuals who lack skill in one or both languages (Poplack, 1980). In fact, some argue that for highly proficient bilinguals, one's two languages are so integrated that code-switching becomes an opportunistic, almost effortless process (Green and Abutalebi, 2013). Our naturalistic data provides converging information that code-switching is associated with highly skilled language use. All participants in our sample were highly proficient English speakers, and greater proficiency in the HL was associated with higher rates of code-switching.

Given the effectiveness of the EAR as a tool to assess frequencies of HL use, could the naturalistic speech samples collected by the EAR be used to assess language proficiency as well? While this may be possible, some challenges arise. As we describe in past work (Macbeth et al., 2022; cf. Montag et al., 2018), given boundaries associated with forming sensible natural language sentences (e.g., function words must appear alongside content words), there is remarkably little variability in the lexical diversity of participant speech as captured by the EAR. At the sample sizes at which the EAR is typically used, there is not nearly enough meaningful variability in lexical diversity for it to be useful measure of vocabulary size or other aspects of word use. Hypothetically, researchers could code speech for various types of errors, but error rates may be quite low, again at the sample sizes typically collected. Likewise, researchers could hypothetically code utterances for syntactic complexity but given the rarity of complex syntax in spoken relative to written language (e.g., Biber, 1988) and the small sample sizes of speech, this method also may not yield stable estimates of complex language use. We are certainly open to the idea that naturalistic speech samples might be used to compute measures of language proficiency, so long as researchers avoid clear pitfalls associated with limits on spoken language lexical diversity and various consequences of the small size of EAR speech samples.

Limitations and future directions

As with any study, the work reported here is not without limitations. The laboratory tasks—verbal fluency and the MINT—were not added to the study protocol until partway through data collection, resulting in a smaller sample for those measures compared to the self-report and EAR assessments. Because of this, we could not reliably examine the relationships between HL (Spanish) MINT scores with other variables of interest. Future work would benefit from examining variability in laboratory-based language measures in heritage bilinguals. For example, in addition to Spanish, the MINT has also been normed in other non-English languages such as Mandarin and Hebrew (Gollan et al., 2015). For studies examining English learners, measures such as LexTALE (Lemhöfer and Broersma, 2012) may also be a valid measure of language skill in select languages, such as Dutch and German. Beyond such lexical tasks, including laboratory-based measures of morphosyntax (e.g., grammaticality judgment tasks), phonology (e.g., phonemic discrimination tasks), semantics (e.g., semantic relatedness judgment tasks), and/or pragmatics (e.g., perspective taking tasks) may provide a more detailed account of language proficiency than lexical measures alone. Using such measures would allow for deeper investigations of the relationships between non-English performance in these laboratory-based language measures, self-ratings, and EAR-based language use.

While the EAR methodology provides an important window into the day-to-day linguistic experiences of bilinguals, certain

limitations exist with naturalistic data collection. First, one of the most appealing aspects of the EAR, that it captures an intermittent sample of language use, can also mean that certain linguistic characteristics might be missed if they are not occurring frequently. Further, EAR can only capture spoken language use, and in today's digital world, much communication is written. An undergraduate heritage bilingual might text or email in a HL, another aspect of real-world HL use, but this cannot be captured *via* audio recording. It is unclear whether rates of HL speech among heritage bilingual undergraduates would match the rates of HL text they produce day-to-day, but this would be an interesting avenue to pursue.

There are also limitations to the aspects of language history that our survey items were designed to assess. The items on our language background measure were primarily drawn from the LEAP-Q (Marian et al., 2007) and LHQ (Li et al., 2014). However, our measure did not capture fine-grained information about non-English language use in the home and in various social settings like school and religious activities (cf. LSBQ; Anderson et al., 2018). It is possible that capturing such information would have allowed us to explain more of the variability in real-world language use examined in the present study. Further, the EAR is capable of providing researchers with information about the number of audio files with speech in different contexts, since it is fairly easy to discern where a participant is and what they are doing throughout the recording period. Future work using the EAR and LSBQ in tandem could yield interesting findings regarding frequency of language use in more specific contexts and with specific interlocutors.

Additionally, we encourage replication and expansion of this study methodology with other bilingual populations. The undergraduate sample of Southern California heritage bilinguals used in the present study differs from other bilingual populations in many ways, the most important being that our sample was primarily English-dominant (i.e., dominant in the majority language), as evidenced *via* both self-report and lab-based proficiency measures. Future work should investigate how the conceptualization of proficiency among heritage bilinguals, as well as their patterns of language use, might differ from self-ratings or speech patterns produced by individuals who identify as more dominant in their HL.

Conclusion

Overall, the results of this study suggest that each of the three measures examined in this study—self-report, laboratory-based tasks, and EAR-based assessments—capture some unique variability in the experiences of heritage bilinguals. For example, it is evident that the EAR provides an estimate of day-to-day language use that is just not possible to attain *via* self-report items or laboratory-based proficiency scores. As such, the EAR should be used for any study where researchers wish to sample naturalistic patterns of bilingual speech and understand how a bilingual's languages are being used

in the real world. Similarly, self-report and lab-based tasks have the benefit of being quick and easy to administer, appear to be fairly consistent with each other, and provide unique information about language proficiency that cannot be obtained using the EAR. Therefore, one or both of these measures should be used when information about an individual's linguistic knowledge and abilities is paramount. While none of these measures strongly correlated—nor did we expect them to—it was evident that they capture information about the heritage bilingual language experience that is shared in some aspects and unique in others. We suggest they be used in tandem to yield the most important insights for the particular research questions being addressed.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: The Open Science Repository (OSF): osf.io/mpjzy.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board (Socio-Behavioral) at the University of California, Riverside. The participants provided their written informed consent to participate in this study.

Author contributions

AM: primary author and editor of the manuscript, conducted data analysis, participated in project conceptualization and implementation, as well as overseeing coding and transcription of raw data. NA: contributed to writing of original manuscript as well as proofreading and editing, created supplementary materials. JM: contributed to proofreading and editing the manuscript, created figures. MB: participated in project conceptualization and implementation,

as well as overseeing coding and transcription of raw data. CC: initial project conceptualization, provided the resources and supervision for project implementation, contributed to proofreading and editing the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was partially supported by a James S. McDonnell Foundation Scholar Award to JM, as well as a National Science Foundation Postdoctoral Research Fellowship under Grant No. SBE-1714925 and CSUF Junior Grant to NA.

Acknowledgments

We would like to thank all of our dedicated research assistants for their invaluable help with testing participants, data coding, and audio transcription. We would also like to acknowledge Jacqueline Erens for her assistance with writing code for preliminary data analysis.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- American Academy of Arts and Sciences Commission on Language Learning (2017). America's languages: Investing in language education for the 21st century. American Academy of Arts and Sciences.
- Anderson, J. A., Mak, L., Chahi, A. K., and Bialystok, E. (2018). The language and social background questionnaire: assessing degree of bilingualism in a diverse population. *Behav. Res. Methods* 50, 250–263. doi: 10.3758/s13428-017-0867-9
- Atagi, N., and Sandhofer, C. M. (2020). Early language environments predict aspects of explicit language awareness development. *Lang. Learn.* 70, 464–505. doi: 10.1111/lang.12381
- Baus, C., Costa, A., and Carrieras, M. (2013). On the effect of second language immersion on first language production. *Acta Psychol.* 142, 402–409. doi: 10.1016/j.actpsy.2013.01.010
- Beatty-Martínez, A. L., Navarro-Torres, C. A., and Dussias, P. E. (2020). Codeswitching: a bilingual toolkit for opportunistic speech planning. *Front. Psychol.* 11:1699. doi: 10.3389/fpsyg.2020.01699
- Bentahila, A., and Davies, E. E. (1983). The syntax of Arabic-French code-switching. *Lingua* 59, 301–330. doi: 10.1016/0024-3841(83)90007-4
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Bice, K., and Kroll, J. F. (2019). English only? Monolinguals in linguistically diverse contexts have an edge in language learning. *Brain Lang.* 196:104644. doi: 10.1016/j.bandl.2019.104644
- Birdsong, D. (1992). Ultimate attainment in second language acquisition. *Language* 68, 706–755. doi: 10.2307/416851

- Blanche, P. (1988). Self-assessment of foreign language skills: implications for teachers and researchers. *RELC J.* 19, 75–93. doi: 10.1177/003368828801900105
- Blanche, P., and Merino, B. J. (1989). Self-assessment of foreign-language skills: implications for teachers and researchers. *Lang. Learn.* 39, 313–338. doi: 10.1111/j.1467-1770.1989.tb00595.x
- Carreira, M., and Kagan, O. (2011). The results of the National Heritage Language Survey: implications for teaching, curriculum design, and professional development. *Foreign Lang. Ann.* 44, 40–64. doi: 10.1111/j.1944-9720.2010.01118.x
- de Bruin, A. (2019). Not all bilinguals are the same: a call for more detailed assessments and descriptions of bilingual experiences. *Behav. Sci.* 9:33. doi: 10.3390/bs9030033
- de Bruin, A., Carreiras, M., and Duñabeitia, J. A. (2017). The BEST dataset of language proficiency. *Front. Psychol.* 8:522. doi: 10.3389/fpsyg.2017.00522
- De Houwer, A. (2021). *Bilingual development in childhood*. Cambridge: Cambridge University Press. Cambridge University Press.
- DeLuca, V., Rothman, J., Bialystok, E., and Pliatsikas, C. (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *PNAS* 116, 7565–7574. doi: 10.1073/pnas.1811513116
- Dunn, L. M., and Dunn, L. M. (1997). *Peabody picture vocabulary test—third edition*. Circle Pines, MN: AGS.
- Flege, J. E., Yeni-Komshian, G. H., and Liu, S. (1999). Age constraints on second-language acquisition. *J. Mem. Lang.* 41, 78–104. doi: 10.1006/jmla.1999.2638
- Friesen, D. C., Luo, L., Luk, G., and Bialystok, E. (2015). Proficiency and control in verbal fluency performance across the lifespan for monolinguals and bilinguals. *Lang. Cogn. Neurosci.* 30, 238–250. doi: 10.1080/23273798.2014.918630
- Gollan, T. H., Montoya, R. I., and Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology* 16, 562–576. doi: 10.1037/0894-4105.16.4.562
- Gollan, T. H., Salmon, D. P., Montoya, R. I., and Galasko, D. R. (2011). Degree of bilingualism predicts diagnosis of Alzheimer's disease in low-education but not in highly educated Hispanics. *Neuropsychologia* 49, 3826–3830. doi: 10.1016/j.neuropsychologia.2011.09.041
- Gollan, T. H., Starr, J., and Ferreira, V. S. (2015). More than use it or lose it: the number-of-speakers effect on heritage language proficiency. *Psychon. Bull. Rev.* 22, 147–155. doi: 10.3758/s13423-014-0649-7
- Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., and Cera, C. M. (2012). Self-ratings of spoken language dominance: a multilingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Biling. Lang. Cogn.* 15, 594–615. doi: 10.1017/S1366728911000332
- Green, D. W., and Abutalebi, J. (2013). Language control in bilinguals: the adaptive control hypothesis. *J. Cogn. Psychol.* 25, 515–530. doi: 10.1080/20445911.2013.796377
- Gumperz, J. J. (1977). The sociolinguistic significance of conversational code-switching. *RELC Journal* 8, 1–34. doi: 10.1177/003368827700800201
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge: Cambridge University Press. Cambridge University Press.
- Hakuta, K., Bialystok, E., and Wiley, E. (2003). Critical evidence: a test of the critical-period hypothesis for second-language acquisition. *Psychol. Sci.* 14, 31–38. doi: 10.1111/j.1467-9280.01415
- Hoshino, N., and Kroll, J. F. (2008). Cognate effects in picture naming: does cross-language activation survive a change of script? *Cognition* 106, 501–511. doi: 10.1016/j.cognition.2007.02.001
- Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Biling. Lang. Cogn.* 15, 422–433. doi: 10.1017/S1366728911000678
- Johnson, J. S., and Newport, E. L. (1989). Critical period effects in second language learning: the influence of maturational state on the acquisition of English as a second language. *Cogn. Psychol.* 21, 60–99. doi: 10.1016/0010-0285(89)90003-0
- Kalamala, P., Senderecka, M., and Wodniecka, J., Senderecka, Z. (2022). On the multidimensionality of bilingualism and the unique role of language use. *Bilingualism: Language and Cognition*, 25, 471–483. doi: 10.1017/S1366728921001073
- Kaplan, E., Goodglass, H., and Weintraub, S. (1983). *The Boston naming test*. Philadelphia, PA: Lea & Febiger.
- Kaşçelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., and De Cat, C. (2021). A review of questionnaires quantifying bilingual experience in children: do they document the same constructs? *Biling. Lang. Cogn.* 25, 29–41. doi: 10.1017/S1366728921000390
- Kaushanskaya, M., and Prior, A. (2015). Variability in the effects of bilingualism on cognition: it is not just about cognition, it is also about bilingualism. *Biling. Lang. Cogn.* 18, 27–28. doi: 10.1017/S1366728914000510
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Lemhöfer, K., and Broersma, M. (2012). Introducing Lex TALE: a quick and valid lexical test for advanced learners of English. *Behav. Res. Methods* 44, 325–343. doi: 10.3758/s13428-011-0146-0
- Lezak, M. D., Howieson, D. B., Bigler, E. D., and Tranel, D. (2012). *Neuropsychological assessment*. 5th Edn. New York: Oxford University Press.
- Li, P., Sepanski, S., and Zhao, X. (2006). Language history questionnaire: a web-based interface for bilingual research. *Behav. Res. Methods* 38, 202–210. doi: 10.3758/BF03192770
- Li, P., Zhang, F., Tsai, E., and Puls, B. (2014). Language history questionnaire (LHQ 2.0): a new dynamic web-based research tool. *Biling. Lang. Cogn.* 17, 673–680. doi: 10.1017/S1366728913000606
- Li, P., Zhang, F., Yu, A., and Zhao, X. (2020). Language history questionnaire (LHQ3): an enhanced tool for assessing multilingual experience. *Biling. Lang. Cogn.* 23, 938–944. doi: 10.1017/S1366728918001153
- Linck, J. A., Kroll, J. F., and Sunderman, G. (2009). Losing access to the native language while immersed in a second language. *Psychol. Sci.* 20, 1507–1515. doi: 10.1111/j.1467-9280.2009.02480.x
- Luk, G., and Bialystok, E. (2013). Bilingualism is not a categorical variable: interaction between language proficiency and usage. *J. Cogn. Psychol.* 25, 605–621. doi: 10.1080/20445911.2013.795574
- Luk, G., and Pliatsikas, C. (2016). Converging diversity to unity: commentary on “the neuroanatomy of bilingualism,” *Lang. Cogn. Neurosci.* 31, 349–352. doi: 10.1080/23273798.2015.1119289
- Macbeth, A., Bruni, M., De La Cruz, B., Erens, J. A., Atagi, N., Robbins, M. L., et al. (2022). Using the electronically activated recorder (EAR) to capture the day-to-day linguistic experiences of young adults. *Collabra. Psychology* 8:36310. doi: 10.1525/collabra.36310
- MacIntyre, P. D., Noels, K. A., and Clément, R. (1997). Biases in self-ratings of second language proficiency: the role of language anxiety. *Lang. Learn.* 47, 265–287. doi: 10.1111/0023-8333.81997008
- Marchman, V. A., Martinez, L. Z., Hurtado, N., Gruter, T., and Fernald, A. (2017). Caregiver talk to young Spanish-English bilinguals: comparing direct observation and parent-report measures of dual-language exposure. *Dev. Sci.* 20:e12425. doi: 10.1111/desc.12425
- Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007)067
- Marian, V., and Hayakawa, S. (2021). Measuring bilingualism: the quest for a “bilingualism quotient.” *Appl. Psycholinguist.* 42, 527–548. doi: 10.1017/s0142716420000533
- Mehl, M. R. (2017). The electronically activated recorder (EAR): a method for the naturalistic observation of daily social behavior. *Curr. Dir. Psychol. Sci.* 26, 184–190. doi: 10.1177/0963721416680611
- Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., and Price, J. H. (2001). The electronically activated recorder (EAR): a device for sampling naturalistic daily activities and conversations. *Behav. Res. Methods Instrum. Comput.* 33, 517–523. doi: 10.3758/BF03195410
- Mehl, M. R., Robbins, M. L., Deters, F., and große, (2012). Naturalistic observation of health-relevant social processes: the electronically activated recorder (EAR) methodology in psychosomatics. *Psychosom. Med.* 74, 410–417. doi: 10.1097/PSY.0b013e3182545470
- Montag, J. L., Jones, M. N., and Smith, L. B. (2018). Quantity and diversity: simulating early word learning environments. *Cogn. Sci.* 42, 375–412. doi: 10.1111/cogs.12592
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.
- Otheguy, R., and Lapidus, N. (2003). “An adaptive approach to noun gender in New York contact Spanish,” in *A romance perspective on language knowledge and language use: Selected papers from the 31st linguistic symposium on the romance languages*. eds. R. Nuñez-Cedeño, L. López and R. Cameron Amsterdam/Philadelphia: Benjamins, 209–229.
- Paap, K. R., Myuz, H. A., Anders, R. T., Bockelman, M. F., Mikulinsky, R., and Sawi, O. M. (2017). No compelling evidence for a bilingual advantage in switching or that frequent language switching reduces switch cost. *J. Cogn. Psychol.* 29, 89–112. doi: 10.1080/20445911.2016.1248436
- Polinsky, M. (2018). Bilingual children and adult heritage speakers: the range of comparison. *Int. J. Biling.* 22, 547–563. doi: 10.1177/1367006916656048
- Polinsky, M., and Kagan, O. (2007). Heritage languages: in the “wild” and in the classroom. *Lang. Linguist. Compass* 1, 368–395. doi: 10.1111/j.1749-818X.2007.00022.x

- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en Español: toward a typology of code-switching. *Linguistics* 18, 581–618. doi: 10.1515/ling.1980.18.7-8.581
- Portocarrero, J. S., Burright, R. G., and Donovick, P. J. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Arch. Clin. Neuropsychol.* 22, 415–422. doi: 10.1016/j.acn.2007.01.015
- Ramírez-Esparza, N., Mehl, M. R., Álvarez-Bermúdez, J., and Pennebaker, J. W. (2009). Are Mexicans more or less sociable than Americans? Insights from a naturalistic observation study. *J. Res. Pers.* 43, 1–7. doi: 10.1016/j.jrp.2008.09.002
- Rodríguez-Fornells, A., Kramer, U. M., Lorenzo-Seva, U., Festman, J., and Munte, T. F. (2012). Self-assessment of individual differences in language switching. *Front. Psychol.* 2:388. doi: 10.3389/fpsyg.2011.00388
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814
- Rothman, J., and Treffers-Daller, J. (2014). A prolegomenon to the construct of the native speaker: heritage speaker bilinguals are natives too! *Appl. Linguis.* 35, 93–98. doi: 10.1093/applin/amt049
- Sanz, C., and Torres, J. (2018). “The prior language experience of heritage bilinguals,” in *The handbook of advanced proficiency in second language acquisition*. eds. P. A. Malovrh and A. G. Benati Hoboken, NJ: Wiley Blackwell, 179–198.
- Sarkis, J. T., and Montag, J. L. (2021). The effect of lexical accessibility on Spanish-English intra-sentential codeswitching. *Memory and Cognition*, 49, 163–180. doi: 10.3758/s13421-020-01069-7
- Sechrest, L., McKnight, P., and McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *Am. Psychol.* 51, 1065–1071. doi: 10.1037/0003-066X.51.10.1065
- Sheng, L., Lu, Y., and Gollan, T. H. (2014). Assessing language dominance in mandarin-English bilinguals: convergence and divergence between subjective and objective measures. *Bilingualism* 17, 364–383. doi: 10.1017/S1366728913000424
- Shi, L. F. (2011). How “proficient” is proficient? Subjective proficiency as a predictor of bilingual listeners' recognition of English words. *Am. J. Audiol.* 20, 19–32. doi: 10.1044/1059-0889(2011/10-0013)
- Shi, L. F. (2013). How “proficient” is proficient? Comparison of English and relative proficiency rating as a predictor of bilingual listeners' word recognition. *Am. J. Audiol.* 22, 40–52. doi: 10.1044/1059-0889(2012/12-0029)
- Surraín, S., and Luk, G. (2019). Describing bilinguals: a systematic review of labels and descriptions used in the literature between 2005–2015. *Biling. Lang. Cogn.* 22, 401–415. doi: 10.1017/S1366728917000682
- Takahesu Tabori, A. A., Mech, E. N., and Atagi, N. (2018). Exploiting language variation to better understand the cognitive consequences of bilingualism. *Front. Psychol.* 9:1686. doi: 10.3389/fpsyg.2018.01686
- Tao, L., Taft, M., and Gollan, T. H. (2015). The bilingual switching advantage: sometimes related to bilingual proficiency, sometimes not. *J. Int. Neuropsychol. Soc.* 21, 531–544. doi: 10.1017/S1355617715000521
- Tomoschuk, B., Ferreira, V. S., and Gollan, T. H. (2019). When a seven is not a seven: self-ratings of bilingual language proficiency differ between and within language populations. *Biling. Lang. Cogn.* 22, 516–536. doi: 10.1017/S1366728918000421
- U.S. Census Bureau (2018). 2018 American community survey 5-year estimates subject tables: Language spoken at home. Available at: <https://data.census.gov/cedsci/table?q=Language%20Spoken%20at%20Home&tid=ACST5Y2018.S1601&moe=false&tp=false&hidePreview=true> (Accessed July 1, 2022).
- Valdés, G. (2001). “Heritage language students: profiles and possibilities,” in *Heritage languages in America: Preserving a national resource*. eds. J. K. Peyton, D. Ranard and S. McGinnis, Washington, DC: Center for Applied Linguistics and Delta Systems, 37–77.
- Vazire, S., and Mehl, M. R. (2008). Knowing me, knowing you: the accuracy and unique predictive validity of self-ratings and other-ratings of daily behavior. *J. Pers. Soc. Psychol.* 95, 1202–1216. doi: 10.1037/a0013314



OPEN ACCESS

EDITED BY

Eloi Puig-Mayenco,
King's College London, United Kingdom

REVIEWED BY

Nicoletta Biondo,
University of California,
Berkeley, United States
Zuzanna Fuchs,
University of Southern California,
United States

*CORRESPONDENCE

Grazia Di Pisa
grazia.di-pisa@uni-konstanz.de
Maki Kubota
maki.kubota@uit.no

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 10 June 2022

ACCEPTED 05 September 2022

PUBLISHED 12 October 2022

CITATION

Di Pisa G, Kubota M, Rothman J and
Marinis T (2022) Effects of markedness in
gender processing in Italian as a heritage
language: A speed accuracy tradeoff.
Front. Psychol. 13:965885.
doi: 10.3389/fpsyg.2022.965885

COPYRIGHT

© 2022 Di Pisa, Kubota, Rothman and
Marinis. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Effects of markedness in gender processing in Italian as a heritage language: A speed accuracy tradeoff

Grazia Di Pisa^{1*}, Maki Kubota^{2*}, Jason Rothman^{2,3} and
Theodoros Marinis^{1,4}

¹Department of Linguistics, University of Konstanz, Konstanz, Germany, ²Department of Language and Culture, UiT The Arctic University of Norway, Tromsø, Norway, ³Department of Language and Education, Universidad Nebrija, Madrid, Spain, ⁴School of Psychology and Clinical Language Sciences, University of Reading, Reading, United Kingdom

This study examined potential sources of grammatical gender variability in heritage speakers (HSs) of Italian with a focus on morphological markedness. Fifty-four adult Italian HSs living in Germany and 40 homeland Italian speakers completed an online Self-Paced Reading Task and an offline Grammaticality Judgment Task. Both tasks involved sentences with grammatical and ungrammatical noun-adjective agreement, manipulating markedness. In grammatical sentences, both groups showed a markedness effect: shorter reading times (RTs) and higher accuracy for sentences containing masculine nouns as compared to sentences with feminine nouns. In ungrammatical sentences, although both groups were sensitive to ungrammaticality, only HSs showed a markedness effect, that is, they had significantly longer RTs and higher accuracy when violations were realized on feminine adjectives. Proficiency in the HL was a significant predictor of accuracy and RTs at the individual level. Taken together, results indicate that HSs acquire and process gender in a qualitatively similar way to homeland native speakers. However, RT evidence seems to suggest that at least under particular experimental methods, markedness considerations are more prevalent for HSs resulting in a speed-accuracy tradeoff.

KEYWORDS

grammatical gender, heritage languages, Italian, markedness, speed-accuracy tradeoff

Introduction

Heritage speakers (hereafter HSs) are early bilinguals who grow up using a language at home that is distinct from the majority language spoken in the society in which they are raised (see, e.g., [Montrul, 2008](#); [Rothman, 2009](#)). In childhood, there is typically a significant shift in exposure from the heritage language (HL), usually coinciding with the start of school, to the societal majority language (ML). As a result, HSs often become dominant in the ML and their adult competence in the HL can vary considerably from homeland native speakers.

Grammatical gender (hereafter *gender*) is an inherent property of the noun reflected in agreement with other elements of the sentence (i.e., articles, determiners, and adjectives; Corbett, 1991). In most languages that have gender, assignment, and agreement are acquired early by monolingual children (cf. Chini, 1995 for Italian children; Müller, 1994 for French and German). Evidence from empirical research has shown that gender can be (particularly) vulnerable in heritage language acquisition (i.e., Montrul et al., 2008; Polinsky, 2008). However, there is also ample evidence showing that *some* HSs converge on a grammar for which gender is seemingly represented (and/or processed) in the same way it is for homeland speakers (see Alarcón, 2011; Bianchi, 2013; Kupisch et al., 2013; Van Osch et al., 2014; Irizarri van Suchtelen, 2016; Fuchs, 2019, 2021). Thus, HSs are fully capable of acquiring underlying syntactic gender systems; however, the ultimate representation of gender systems might not develop to be entirely the same as in homeland native speakers' grammars. For example, HSs of Romance languages—where feminine is marked relative to the default masculine—tend to make more errors with feminine nouns (i.e., Montrul et al., 2008; Alarcón, 2011; Bianchi, 2013; Hur et al., 2020), suggesting that factors such as morphological markedness play a role.

It is also worth highlighting that much of what we know so far about the acquisition and processing of gender in adult HSs is based on behavioral offline methods (but see, e.g., Fuchs, 2019, 2021; Keating, 2022), such as acceptability judgement, comprehension, and recognition tasks. These provide significant insights into HS behavior. However, as recently pointed out by Bayram et al. (2021), offline methods alone can be problematic with regard to the kind of knowledge they are targeting, soliciting, and capturing. Behavioral tasks can be influenced by (unconscious and conscious) metalinguistic and affective variables. Since HSs are more likely to have less (and/or qualitatively distinct) metalinguistic knowledge (Rothman, 2007; Montrul et al., 2014; Bayram et al., 2019) and/or be more apprehensive to give definitive judgments (Polinsky, 2018), offline tasks alone could introduce noise that obscures HSs underlying competence.

With the above in mind, the main goal of the present study is to combine offline judgments with automatic processing responses (reaction times while reading) to determine whether HSs of Italian in the German context, like Romance homeland speakers (Alemán Bañón and Rothman, 2016; Alemán Bañón et al., 2017), are sensitive to morphological markedness when processing gender agreement violations during sentence comprehension.

Gender in Italian and German

In Italian, there are two gender values: masculine and feminine. Gender assignment is largely transparent and follows both semantic and morpho-phonological rules. Canonical endings in Italian are -o and -a; thus, nouns ending in -o are typically masculine (*albero* “tree_M”), while those ending in -a are usually feminine (*casa* “house_F”; Schwarze, 2009). There are, of course, exceptions, for

example, *problema_M* (“problem” ending in -a but masculine) and *mano_F* (“hand” ending in -o but feminine). Nouns with non-canonical endings are gender ambiguous and less frequent. In nouns ending in -e, for example, gender is not clearly marked, as these nouns could be either masculine (*pane_M* “bread”) or feminine (*notte_F* “night”). Nevertheless, some derivational suffixes can help to determine the gender of the noun, since they regularly co-occur with one of the two genders. For example, words that end in -trice and -zione (*calcolatrice_F* “calculator,” *posizione_F* “position”) are reliably feminine, whereas those ending in -ale and -one (*pugnale_M* “dagger,” *cotone_M* “cotton”) are masculine (Chini, 1995).

Italian requires gender agreement between the noun and its determiners, most modifying adjectives, and pronouns. In this study, we focus on the mastery of gender agreement on predicative adjectives; therefore, examples of gender agreement on adjectives are provided in (1) (a,b) for feminine nouns, and (c,d) for masculine nouns.

(1) a. *La_F luna_F rossa_F.*

‘The red moon.’

b. *La_F volpe rossa_F.*

‘The red fox.’

c. *Il_M libro_M rosso_M.*

‘The red book.’

d. *Il_M pesce rosso_M.*

‘The red fish.’

As shown in 1(b) and (d), the nouns *volpe* “fox” and *pesce* “fish” have no overt ending corresponding to feminine and masculine gender, as in 1(a) and (c). Rather, their lexical entries include a specification for feminine 1(b) and masculine 1(d) gender, respectively. In all cases, 1 (a–d) there is gender agreement between the definite article, the lexical gender feature of the head noun, and the agreeing (predicative) adjective.

As alluded to above, although reliable morphological marking in Italian is helpful, as in 1(a) and (c) above, all nouns have grammatical gender even in the absence of an unambiguous morphological ending on the noun, as in 1(b) and (d). Since gender is an inherent part of the noun's entry in the mental lexicon, it brings together lexical and syntactic aspects (Corbett, 1991; Kramer, 2015). At the lexical level, learners need to first assign gender to nouns (*assignment*); then, when used (or processed) in a sentential context, the syntactic reflexes of agreement come to bear (*agreement on adjectives*).

Unlike Italian, German has a three-way gender system with masculine, feminine, and neuter nouns (Durrell, 2011). With respect to gender assignment, nouns are largely opaque. Even though there are some semantic, morphological, and phonological

patterns, there are also many exceptions (Köpcke, 1982). This makes the German system much less transparent compared to the Italian one.

As for agreement, in contrast to Italian, gender in German is not marked on the noun itself, but rather on determiners and adjectives occurring within the same DP (or referring to it elsewhere). However, the gender of determiners and adjectives can sometimes be ambiguous since agreement also depends on definiteness (definite vs. indefinite), case (nominative, accusative, dative, and genitive), and number (singular vs. plural; Kunkel-Razum et al., 2009).

Grammatical gender in heritage speakers

In the last two decades, HSs knowledge of gender systems has been the object of considerable research. Within the available literature there is a juxtaposition of findings, sometimes even for the same HL (e.g., Spanish) depending on the study/method used. While some show that HSs struggle with gender assignment and/or subsequent agreement in production and comprehension in various HLs (i.e., Russian: Polinsky, 2008; Spanish: Montrul et al., 2008), others demonstrate that HSs do not differ qualitatively from homeland native speakers (Italian: Bianchi, 2013; French: Kupisch et al., 2013; Spanish: Alarcón, 2011; Montrul et al., 2013; and Russian: Laleko, 2018). This suggests that mastery of gender systems in HLs that are qualitatively the same as in homeland varieties in HLs is attainable, although they can be vulnerable under specific conditions.

A closer look at these studies shows that gender in HSs is significantly affected by the HL-ML combination, the level of HL proficiency, the HL use, and the age of onset (AoO) of bilingualism. Most of the studies investigating gender have tested HSs whose ML was a non-gendered language, most often English (i.e., Spanish: Montrul et al., 2008; Russian: Polinsky, 2008). Fewer studies have been conducted in language pairs in which both languages have gender, but differ with respect to the properties of their gender systems (i.e., Italian and German: Bianchi, 2013; French and German: Kupisch et al., 2013). Regarding gender and proficiency, findings are controversial with some studies reporting higher error rates for HSs with lower proficiency level (i.e., Montrul et al., 2008), and other studies testing HSs with a higher level of proficiency finding no differences in terms of performance between HSs and matched homeland native speakers (i.e., Alarcón, 2011; Bianchi, 2013; Kupisch et al., 2013). However, even advanced HSs often are different compared to monolinguals with respect to grammatical gender when tested on non-canonical nouns (i.e., Bianchi, 2013; Montrul et al., 2013). Previous studies measuring HSs' relative amount (and quality) of exposure and use of their HL have shown that variation in HL exposure has consequences for HL development in children (i.e., Gagarina and Klassert, 2018; Torregrossa et al., 2021) and maintenance in adults (i.e., Lloyd-Smith et al., 2019, 2020). Some studies on gender have shown that HL exposure and/or use has an effect on HSs'

performance (i.e., Bianchi, 2013); however, others (i.e., Fuchs, 2021) found no evidence. Therefore, it is still an open question to what extent one's individual experiences with the HL modulate gender processing in HSs. Furthermore, previous studies have shown that AoO of bilingualism plays a role in the acquisition and maintenance of HLs, usually leading to more variable outcomes in simultaneous bilinguals (i.e., Montrul, 2008; Montrul et al., 2014; Giancaspro, 2017). However, few studies on gender in adult HSs have examined effects of AoO of bilingualism revealing controversial results (for Italian: Bianchi, 2013; for Spanish: Montrul, 2008; Keating, 2022), thus leaving open the question of the extent to which the syntax of gender is really affected by AoO.

In line with the inconsistent findings of the above factors, and in light of recent turns in various literature examining bilingual language and cognitive systems that advocate for regressing factors pertaining to exposure and, crucially, dynamic engagement with language in various contexts (DeLuca et al., 2019; Titone and Tiv, 2022), we collected detailed information on all these factors. The logic in doing so is to be able to unpack the conditions under which general observations are more or less true. In other words, it could be the case, for example, that morphological markedness affects HS processing more or less under specific conditions for individual HSs.

Morphological markedness

As it is the case that markedness can be understood differentially (morphologically, semantically, and frequency based), let us start by being explicit as to what we take to be marked and why in the present context. We take the position that in Italian gender, feminine is marked relative to masculine. Given the robust associations that the classical morphemes (-o, -a, -i, and -e) have with their respective gender, in one sense of markedness, it would be reasonable to argue that each is equally morphologically marked. This, however, is not the sense we mean. Claiming that feminine is marked and, relatedly, that masculine is the default is supported by both a frequentist position and various facts. In Italian, masculine nouns by far outnumber feminine ones: 60% are masculine and 40% are feminine (Costa et al., 2003). Furthermore, when one considers some classical diagnostics, it is easy to see how masculine is the default. For example, when nominalizing (and/or conceptually abstracting) anything new or novel in Italian, masculine is the gender assigned (D'Achille, 2003) as shown when a verb is made into a noun: $Il_M/*La_F$ *fumare è dannoso alla salute* "Smoking is harmful to health." Another example is the case of lexical borrowings that mostly take the masculine gender whether or not they are incorporated into Italian morpho-phonological or remain as lexical insertions, for example, *il film*, *il software*, *lo smartphone*.

Morphological markedness theory (Battistella, 1990) assumes that feature values, e.g., masculine vs. feminine for gender, are asymmetrically represented and have a hierarchical structure, with the more general or *unmarked* element (masculine in the Romance

case) being the “default value,” indicating just the presence of a grammatical feature (gender), and the most specific or *marked* version(s) (feminine in Romance languages) indicating a specific feature value (or specification; Battistella, 1990). In Italian, masculine is the most frequent gender (Cacciari, 2011) and it is also the least-marked; thus, masculine is considered the “default gender,” while feminine is regarded as marked (D'Achille, 2003). In German, however, the default gender is not as clear considering the presence of a third gender (neuter) in its system; however, convincing evidence exists to suggest that masculine is also the default gender in German (e.g., Steinmetz, 2006).

Previous research on homeland native speakers and L2 speakers examining noun–adjective agreement in Romance languages like French (Vigliocco and Franck, 1999, 2001), Italian (Vigliocco and Franck, 2001), and Spanish (e.g., Antón-Méndez et al., 2002; McCarthy, 2008) has shown that agreement errors were more frequent when the head noun was feminine (marked). This tendency to overuse the default gender (masculine) on agreement targets suggests the use of masculine as a default agreement strategy (i.e., McCarthy, 2008). Furthermore, some studies have shown that agreement violations realized on marked elements are detected more easily, consistent with the claim that marked features are more disruptive, thus more costly to process and consequently more recognizable during the processing of agreement (e.g., Deutsch and Bentin, 2001; Nevins et al., 2007).

In a recent set of neuroimaging studies, Alemán Bañón and Rothman (2016) and Alemán Bañón et al. (2017) found that both homeland Spanish native speakers and Spanish L2 learners were sensitive to markedness asymmetries, such that the P600 for gender violations emerged earlier and it was larger for *feature clash (marked) errors* (masculine noun+*feminine adjective) than *default (unmarked) errors* (feminine noun+*masculine adjective). This is consistent with the possibility that errors that involve mismatching marked features are more disruptive and easily detectable. While the two groups differed quantitatively, neither showed any systematic evidence of reliance on morphological defaults, although their online processing was sensitive to markedness in a native-like manner.

Previous studies focusing on the linguistic factors underlying gender errors in HSs have reported the tendency for HSs to be more accurate on gender assignment and agreement with the language-specific unmarked form, for example, in Spanish masculine nouns compared to feminine ones (i.e., Montrul et al., 2008; Van Osch et al., 2014; Irizarri van Suchtelen, 2016; Goebel-Mahrle and Shin, 2020; Hur et al., 2020). This over-reliance on the masculine in the above cases could be explained in terms of morphological markedness. Nevertheless, as the tasks used in previous studies were offline, we do not know how (or if) markedness affects online sentence processing.

Depending on the research question, homeland native comparisons are not always necessary or particularly illuminating in HS studies. Herein, however, we are interested in the comparison for a few reasons. To begin, we do not know if markedness matters for online gender agreement processing with

this type of method in any group—the studies we referenced showing such effects in homeland native language processing are not reading RT studies. While we realize that homeland Italian speakers are not necessarily the baseline for our Italian HSs, it would be interesting to see the extent to which markedness plays a role for the homeland group with this method to best contextualize/interpret what we observe for the present HSs. There is good reason to anticipate that HSs will show considerable markedness effects, above and beyond what the homeland speakers may or may not show, precisely because HS grammars have been shown to be particularly reliant, if not magnify (morphological) defaults (Polinsky, 2018). If so, in the present context, one might expect marked agreement (a)symmetries to be even more salient for HSs.

Research questions and hypotheses

Given the previous discussion, the present study aims to answer the following research questions:

RQ1: Are HSs sensitive to morphological markedness, and if yes, how does markedness affect the processing of agreement violations in HSs as compared to homeland speakers?

Based on previous research (Alemán Bañón and Rothman, 2016), we expect homeland speakers to be sensitive to morphological markedness (feature clash being more marked: masculine noun+*fem. adjective). Behaviorally, evidence in support of this would be obtained if they are more accurate with feature clash errors than default ones, although given that accuracy is assessed *via* offline judgment there could be a ceiling effect in accuracy. Conversely, in terms of the online measure, we would definitely expect sensitivity to markedness shown *via* speakers' slowing down with feature clash errors, indicating their grammatical system has detected an error. Regarding specific error type, we would expect RT slowdowns in the SPRT and higher accuracy in the Grammaticality Judgment Task (GJT) for feature clash (marked) errors (masculine noun+*feminine adjective), as opposed to no RT slowdowns and lower accuracy for default (unmarked) errors (feminine noun+*masculine adjective). As this is the first study to test markedness in this domain in HSs, we are unsure what to expect precisely although there is no reason, *a priori*, to not expect them to be equally sensitive to markedness. After all, we know that other sets of bilinguals are, even non-natives ones (Alemán Bañón et al., 2017). We might expect Italian HSs to be over-reliant on defaultness (masculine as default gender) as well as more sensitive to feature clash (marked) errors as compared to default (unmarked) ones given the heightened role that defaults can play in HS grammatical systems (Polinsky, 2018).

RQ2: Do proficiency and extra-linguistic factors (i.e., type of bilingualism, quantity and quality of input) affect accuracy and RTs in HSs?

In order to understand whether specific background variables affect HSs' performance in both tasks, we will consider the variables that have been shown to affect HL acquisition (HL proficiency, HL use in the home and in the society, type of bilingualism—simultaneous vs. sequential). We expect HSs' overall performance to benefit from higher proficiency in the HL (i.e., [Bianchi, 2013](#); [Kupisch et al., 2013](#)) and more HL use (i.e., [Bianchi, 2013](#)). Regarding AoO, there are two possible scenarios: in line with [Montrul \(2008\)](#) and [Keating \(2022\)](#), sequential HSs could be more accurate and show sensitivity to markedness earlier (faster RTs) than simultaneous HSs; or similar to [Bianchi \(2013\)](#), we could find no difference between the two groups of HSs due to the fact that gender acquisition in Italian is not problematic, given the high degree of transparency of the Italian gender system ([Kupisch et al., 2002](#); [Velnić, 2020](#)), and thus robust to AoO of bilingualism effects.

RQ3: Is HSs' use of markedness information during processing of agreement affected by task modality (offline vs. online)?

We expect to find an effect of markedness in both tasks; however, we leave open the possibility that the degree of this effect will differ across the two modalities.

To answer these questions, we tested a group of HSs of Italian living in Germany and a group of homeland Italian native speakers living in Italy. We used a Self-Paced Reading Task (SPRT) to tap into implicit processing of ungrammaticality and judgments from a Grammaticality Judgment Task (GJT) to examine accuracy, the latter potentially tapping into more explicit factors. The tasks presented complex sentences in Italian where markedness was examined by systematically manipulating the gender specification of the agreeing adjective following the noun.

Materials and methods

Participants

Fifty-four adult HSs of Italian (35 females, $M_{age} = 28.15$; $SD = 6.20$; $range = 18-41$) living in Germany and 40 adult homeland Italian speakers (29 females, $M_{age} = 25.65$; $SD = 3.99$; $range = 18-39$) living in Italy participated in the study. We initially recruited 55 HSs but one participant was excluded because exposed to three languages from birth. All the homeland speakers grew up monolingually in Italy and were living in Italy at the time of testing. All HSs grew up in Germany; however, six HSs were not born in Germany (five were born in Italy and one was born in Argentina), but even in these cases, each had moved to Germany between the age of 1 and 5 years ($M_{age} = 2.5$; $SD = 1.64$). The heritage group comprised 33 simultaneous bilinguals who were exposed to German from birth and had one Italian and one German-speaking parent and 21 sequential bilinguals who had two Italian-speaking parents and their first intensive contact with German occurred between 3 and 6 years ($M_{age} = 1.5$; $SD = 1.97$)

when they started kindergarten in Germany. They all completed their schooling in Germany and they were living in Germany at the time of testing. To assess the effect of HL use on the processing of gender in Italian as well as to quantify aspects of Italian use across the lifespan, all participants completed the Language and Social Background Questionnaire (LSBQ; [Anderson et al., 2018](#)). The LSBQ aims at capturing participants' second language use from childhood to the present day and across several settings and dimensions. It yields two scores related to the amount of (bilingual) language use within specific communicative settings. Specifically, the social score (in our study referred as "HL in the society") is related to language use in the participant's social life (e.g., at work, when writing emails, watching TV, etc.), the possible range is: -7.5 to 80.304 ; the higher the score, the more frequently the second language is used in social settings. Whereas the home score (in our study referred as "HL in the home") is related to language use in home settings (for instance language used with grandparents, during infancy, proficiency in the second language, etc.); the possible range is: -13.9 to 24.163 , the higher the score, the more the second language is used in home settings. Detailed demographic information about the participants is provided in [Supplementary Table S1](#) in the [Supplementary material](#).

Proficiency

Proficiency in Italian was assessed using an adapted version of the Italian placement test originally created by [Alderson \(2005\)](#), known as DIALANG test battery. The test consists of 50 real words and 25 pseudo-words requiring a YES or NO response. In our adaptation (see [Lloyd-Smith et al., 2021](#)), the items appeared in the center of the screen one at the time, and participants were instructed to press on their keyboard key F if they thought the word existed or key J if they did not. Scoring consisted of simply the sum of all correct answers (i.e., one point for each correctly identified word or non-word). The maximum possible score was 75. As shown in [Figure 1](#), HSs had lower proficiency and their scores displayed a much larger degree of variation ($M = 60.33$; $SD = 6.49$; $range = 44-70$) as compared to the homeland native speakers ($M = 69.80$; $SD = 2.33$; $range = 66-75$).

Materials

The study included two main experimental tasks: a *self-paced reading task* and a *grammaticality judgment task*. The materials for both tasks comprised 80 sentences of eight words each. All sentences presented the same structure: subject + auxiliary verb *have* + past participle + indefinite article + trigger noun + adjective (always in post-nominal position) + preposition + object. Morphological markedness was manipulated in the gender specification of the trigger noun and the agreeing adjective, as shown in [Table 1](#), which provides a sample of each of the four experimental conditions. The trigger noun was feminine in half of

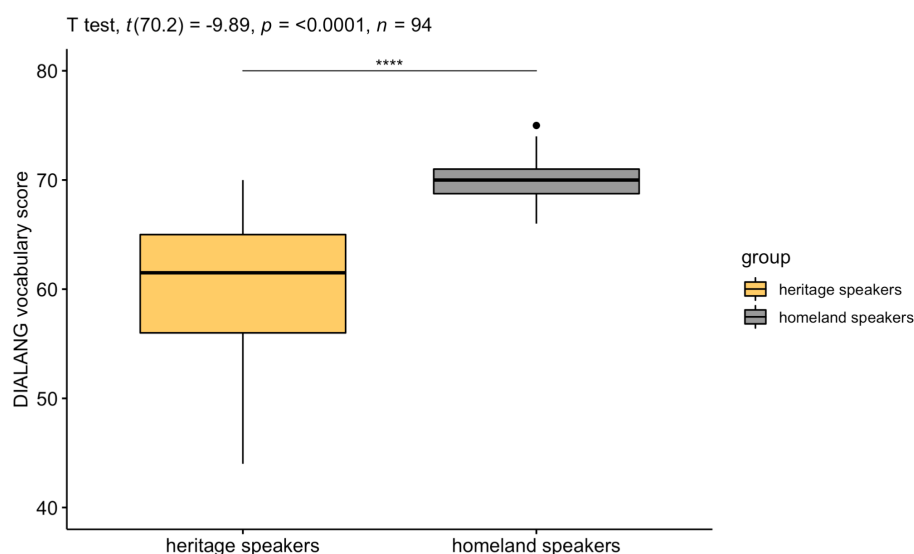


FIGURE 1
Heritage and homeland speakers' scores on the Italian vocabulary test DIALANG (raw scores).

the sentences ($N = 40$) and masculine in the other half ($N = 40$). This can be seen by comparing the sentences in (1–2) to those in (3–4) in Table 1. In (1–2), the noun *Torre* “tower” is feminine and therefore, the agreeing adjective *antica*_{FEM} “old” must also be feminine, as shown in (1). Otherwise, the string would be ungrammatical *antico*_{MASC}, as shown in (2). The opposite pattern is shown in (3–4), where the trigger noun *pesce* “fish” is masculine and, therefore, the agreeing adjective *rosso*_{MASC} “red” must also be masculine, as shown in (3). Otherwise, the string would be ungrammatical, as shown in (4).

The study also included 80 filler sentences (40 grammatical, 40 ungrammatical) which did not manipulate gender agreement and did not include any adjectives. The overall design encompassed an equal amount of grammatical and ungrammatical sentences. These 160 sentences were counterbalanced across four experimental lists where the carrier sentences were the same. Each participant was pseudorandomly assigned to one of the four lists (the same list was used in the SPRT and in the GJT), so that a given participant would see 20 items per each of the two conditions in (1–2; a total of 40) and 20 items per each of the two conditions in (3–4; a total in 40), but no participant saw the same sentence twice.

Properties of the stimuli

None of the trigger nouns exhibited the *-o/-a* canonical endings strongly associated with masculine and feminine genders in Italian to ensure that participants could not resort to a phonological strategy for matching the agreeing elements. Instead, we selected nouns ending in *-e* that in Italian are either completely opaque with respect to their gender or the *-e* forms

part of a derivational suffix which defaults to one or the other gender. In sum, we controlled for noun ending transparency, so that half of the trigger nouns were truly opaque; hence, gender could not be recovered from the surface form (e.g., *pont-e*_M “bridge”), while the other half consisted of nouns ending with *-e* as part of a derivational suffix providing a cue about gender, making them more transparent (e.g., *magli-one*_M “jumper”). Gender congruency was also controlled in a way that half of the trigger nouns in Italian share the same gender in German, while the other half have the opposite gender. We only used nouns with masculine or feminine gender; thus, we did not use nouns that were neuter in German. Furthermore, the trigger nouns were presented in both the singular and the plural form always counterbalanced. A total of 77 trigger nouns were used: 39 masculine nouns (one noun was used twice) and 38 feminine nouns (two nouns were used twice), see Supplementary material 1 for a complete list of all the trigger nouns. A log frequency count for all nouns and adjectives was obtained from the CoLFIS corpus (Corpus e Lessico di Frequenza dell’Italiano Scritto, Bertinetto et al., 2005). Masculine and feminine nouns were matched with respect to frequency, $t(75) = -0.446$, $p > 0.1$, and number of syllables, $t(75) = -0.609$, $p > 0.1$. The masculine and feminine versions of the adjectives were also matched for frequency, $t(78) = 0.803$, $p > 0.1$, and number of syllables, $t(78) = 1.028$, $p > 0.1$.

Tasks

Self paced reading task

The SPRT used a non-cumulative word-by-word center presented design (Marinis, 2010). The 160 sentences were

TABLE 1 Sample stimuli for the experimental conditions.

FEMININE NOUN
Grammatical feminine
1. Daniele ha fotografato una <u>torre antica</u> a Roma.
Daniele took a picture of a tower _{FEM} old _{FEM-marked} in Rome.
Ungrammatical feminine - Default (Unmarked) Error
2. Daniele ha fotografato una <u>torre</u> *antico a Roma.
Daniele took a picture of a tower _{FEM} old _{MASC-unmarked} in Rome.
MASCULINE NOUN
Grammatical masculine
3. Alessandro ha comprato un <u>pesce rosso</u> alla fiera.
Alessandro bought a fish _{MASC} red _{MASC-unmarked} at the fair.
Ungrammatical masculine - Feature Clash (Marked) Error
4. Alessandro ha comprato un <u>pesce</u> *rossa alla fiera.
Alessandro bought a fish _{MASC} red _{FEM-marked} at the fair.

divided into four blocks of 40 sentences each, with 20 correct and 20 incorrect sentences per block separated by short breaks. Detailed instructions and four practice sentences with accuracy feedback preceded the experiment to familiarize participants with the task. None of the practice trials involved agreement errors. In addition, in order to avoid repetition effects, the practice sentences were designed with lexical material that did not appear in the experimental stimuli. Immediately after the practice, the main experiment began. The sentences were presented randomly.

Each trial began with a fixation cross and the first word appeared after 500 ms. Participants used the spacebar to advance through the words. To ensure that the participants were paying attention to the stimuli, a binary yes/no comprehension question appeared after 35% of the sentences on a separate display screen [see (2) for a sentence example]. Participants responded using keys F (YES) or J (NO) on their keyboard. The question stayed on the screen until the participant answered. No feedback was given for correct or incorrect answers. Participants were instructed to read the sentences as fast as possible, and they were told that the task targeted reading comprehension.

- (2) Daniele | ha | fotografato | una_F | torre_F | antica_F | a | Roma.
 R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8.
Daniele è stato a Roma?
 ‘Was Daniele in Rome?’
 a. *Si* “Yes.”
 b. *No* “No.”

Grammaticality judgement task

The GJT stimuli were identical to those of the SPRT, with 160 sentences divided into four blocks of 40 sentences each, with 20 correct and 20 incorrect sentences per block separated by short breaks. Participants read the full sentence on the screen and were asked to judge whether or not the sentence was grammatically correct by pressing keys F (YES) or J (NO) on their keyboard. All sentences were presented in a random order.

Gender assignment task

In order to check whether participants assigned the correct gender to the target nouns used in the main tasks, participants completed a Gender Assignment Task (GAT). Participants were presented with all 77 trigger nouns from the experimental sentences and were instructed to select the appropriate gender-marked determiner from among two options (*il_M* “the” vs. *la_F* “the”) by using the keys F (*il*) and J (*la*) on their keyboard. The trigger nouns were presented one after the other in isolation, and at the end of the task, participants were also asked to indicate whether they knew each word and its meaning.

Procedure

Due to the pandemic, the experimental session was completed online *via* the internet by each participant using their personal computer. All tasks were created using Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine et al., 2020). Data were collected between 28 June 2020 and 30 September 2020. Prior to the experiment, participants filled out the language and social background questionnaires, then they completed the DIALANG in Italian, the SPRT, the GJT, and the GAT. The entire session lasted approximately 45 min and participants were allowed to have breaks in between the tasks. Participants received a compensation for their participation. All participants provided informed consent to take part in the study and all procedures were approved by the research ethics committee of the University of Konstanz, Germany.

Analyses

Trials containing trigger nouns reported as “unknown” by the participants were removed from all the tasks. Homeland speakers reported knowing all the trigger nouns, while HSS’ knowledge of trigger nouns was high with some variability ($M = 72$ out of 77; $SD = 4.65$; $range = 57-77$). Furthermore, accuracy on the GAT was used for data cleaning in the SPRT and in the GJT for both groups; thus, we only included trials with nouns for which the participants assigned the correct target gender in the GAT. Moreover, raw RTs were screened for extreme values and outliers (Keating and Jegerski, 2015; Marsden et al., 2018). We excluded all segments with RTs below 150 ms and above 6,000 ms on the basis of histograms. For the remaining data, we trimmed all raw RTs that deviated more than 2.5 SDs below and above from the participants’ mean per position and per condition. Percentages of removed data and final data pool are provided in Supplementary Table S2 in the Supplementary material.

Sentences were segmented into eight regions (see Example 2 above) and the analyses for RTs were done on three specific regions of interest: *Region 5* = noun (pre-critical), *Region 6* = adjective (critical region), and *Region 7* = spill-over (post-critical).

Accuracy data from the GJT were analyzed with mixed effects logistic regressions (Jaeger, 2008), and RTs from the SPRT were

analyzed with mixed effects linear models (Baayen et al., 2008) in R (R Core Team, 2016). We used the mixed function in the *afex* package (Singmann et al., 2022) to run a likelihood ratio test. The categorical variables were sum-coded and numerical variables were centered around the mean. Pairwise *post-hoc* comparisons with Tukey's contrasts were conducted using the *emmeans* package (Lenth, 2022). Figures were produced using the package *ggplot2* (Wickham, 2016).

The first analysis focused on the comparison between HSs and homeland speakers to establish whether both groups were sensitive to markedness (RQ1) in terms of accuracy in the GJT and RTs in SPRT. The dependent variable was a binary outcome (correct or incorrect) for the GJT and RTs for the SPRT. We included *Group* (heritage vs. homeland speakers), *Grammaticality* (grammatical vs. ungrammatical), and *Gender* (feminine vs. masculine), as well as their interactions (*Group:Grammaticality*, *Group:Gender*, *Grammaticality:Gender*, and *Group:Grammaticality:Gender*) as fixed effects. We included *Grammaticality*Gender* slopes for Subject and Item intercepts and simplified the model following Bates et al. (2015) until there were no convergence issues.

The second analysis was restricted to the heritage group in order to investigate to what extent proficiency and extra-linguistic factors predicted the likelihood of accuracy as well as RTs (RQ2). *Grammaticality* (grammatical vs. ungrammatical), *Gender* (feminine vs. masculine), the DIALANG proficiency scores, and type of bilingualism (simultaneous vs. sequential), HL use in the home (HL home) and in the society (HL society) as well as their interactions (*Grammaticality:Gender*, *Grammaticality:Proficiency*, *Grammaticality:Bilingualism*, *Grammaticality:HL_home*, *Grammaticality:HL_society*, *Gender:Proficiency*, *Gender:Bilingualism*, *Gender:HL_home*, *Gender:HL_society*, *Grammaticality:Gender:Proficiency*, *Grammaticality:Gender:Bilingualism*, *Grammaticality:Gender:HL_home*, and *Grammaticality:Gender:HL_society*) were included as fixed effects in the model. The proficiency scores as well as HL home and HL Society scores were centered prior to statistical analyses. We included *Grammaticality*Gender* slopes for Subject and Item intercepts and simplified the model until there were no convergence issues.

Results

Figures and averages of the SPRT are shown in raw measures for ease of interpretation, but the models were fit to log-transformed RTs, to remove skew, and to normalize model residuals (Vasishth and Nicenboim, 2016). Accuracy results from the GAT are provided in Supplementary material 4.

Self paced reading task

Before analyzing participants' RT data, we examined accuracy rates for the comprehension question responses. Both groups demonstrated a high mean accuracy rate: 93.0% ($SD = 0.25$) in the

heritage and 95.2% ($SD = 0.21$) in the homeland speaker group. Thus, participants were reading for meaning and were attentive during the task. All participants scored above 50% accuracy, so no participant was excluded. For the analysis of RTs, we only included trials that received correct answers. As shown in Figures 2, 3 illustrating overall reading patterns (non-log-transformed RTs), HSs had longer RTs than homeland speakers.

In Region 5 (pre-critical) containing the noun, the significant effect of *Group* ($\text{Chisq} = 19.74$, $p < 0.001$) indicates that overall HSs had longer RTs as compared to homeland speakers. We found no effect of *Grammaticality* nor *Gender*, indicating that the effects observed in the critical region did not start earlier. No further analyses were conducted on this region.

In Region 6 (critical), the between-group analysis revealed a significant effect of *Grammaticality* ($\text{Chisq} = 5.76$, $p = 0.016$), indicating that both groups had shorter RTs for grammatical sentences compared to ungrammatical ones. The effect of *Group* ($\text{Chisq} = 27.67$, $p < 0.001$) indicates significantly longer RTs for HSs as compared to homeland speakers. The significant interaction between *Group:Gender* ($\text{Chisq} = 5.22$, $p = 0.022$) and *post-hoc* pairwise comparisons indicates that the difference in RTs between feminine and masculine nouns was significantly different between groups ($\beta = -0.040$, $SE = 0.017$, $z = -2.327$, $p = 0.019$), indicating that HSs had shorter RTs for feminine nouns compared to masculine ones in comparison to homeland speakers. The three-way interaction between *Group:Grammaticality:Gender* was not significant ($\text{Chisq} = 2.78$, $p = 0.096$). However, since this was our *a priori* comparison, we ran *post-hoc* pairwise comparisons, showing that the difference in RTs between feminine and masculine nouns in the grammatical conditions was not different between HSs and homeland speakers ($\beta = -0.014$, $SE = 0.023$, $z = -0.607$, $p = 0.544$). In the ungrammatical condition, however, the difference in RTs between feminine and masculine nouns was significantly different between groups ($\beta = -0.067$, $SE = 0.023$, $z = -2.843$, $p = 0.005$), indicating that HSs had shorter RTs with feminine nouns (where the ungrammaticality was caused by an unmarked masculine adjective) compared to masculine nouns (where the ungrammaticality was caused by a marked feminine adjective).

In Region 7 (spill-over), a significant effect of *Grammaticality* ($\text{Chisq} = 26.36$, $p < 0.001$) indicates shorter RTs for the grammatical conditions compared to the ungrammatical ones in both groups. The significant main effect of *Group* ($\text{Chisq} = 21.98$, $p < 0.001$) reflects overall longer RTs for HSs as compared to homeland speakers. The significant interaction between *Group:Grammaticality:Gender* ($\text{Chisq} = 5.94$, $p = 0.015$) indicates that grammaticality affected RTs in HSs and homeland speakers in a different way. Subsequent *post-hoc* pairwise comparisons showed that in the grammatical conditions, the difference in RTs between feminine and masculine nouns was not different between HSs and homeland speakers ($\beta = -0.014$, $SE = 0.023$, $z = -0.60$, $p = 0.543$). However, the difference in RTs between feminine and masculine nouns in the ungrammatical conditions was significantly different between groups ($\beta = -0.066$, $SE = 0.023$, $z = -2.843$, $p = 0.004$),

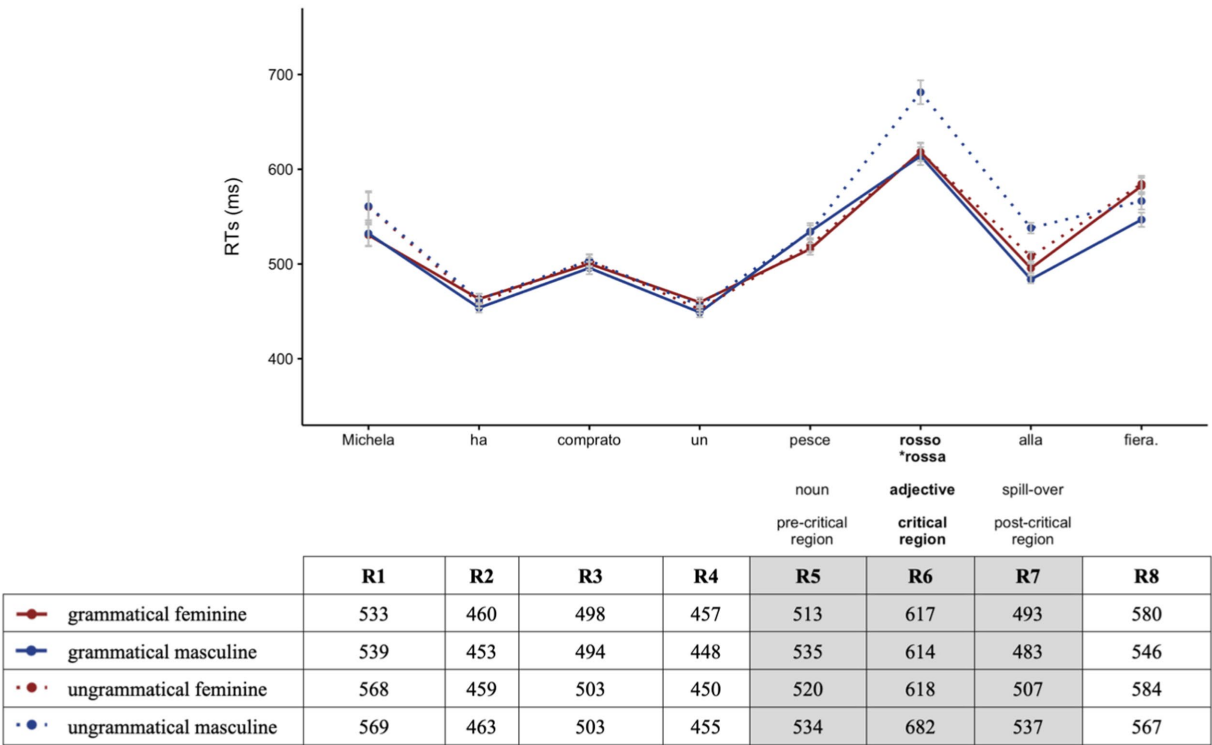


FIGURE 2
Heritage speakers (HSs) mean reading times (RTs) by region for grammatical (solid lines) vs. ungrammatical (dotted lines) sentences for feminine (red) and masculine (blue).

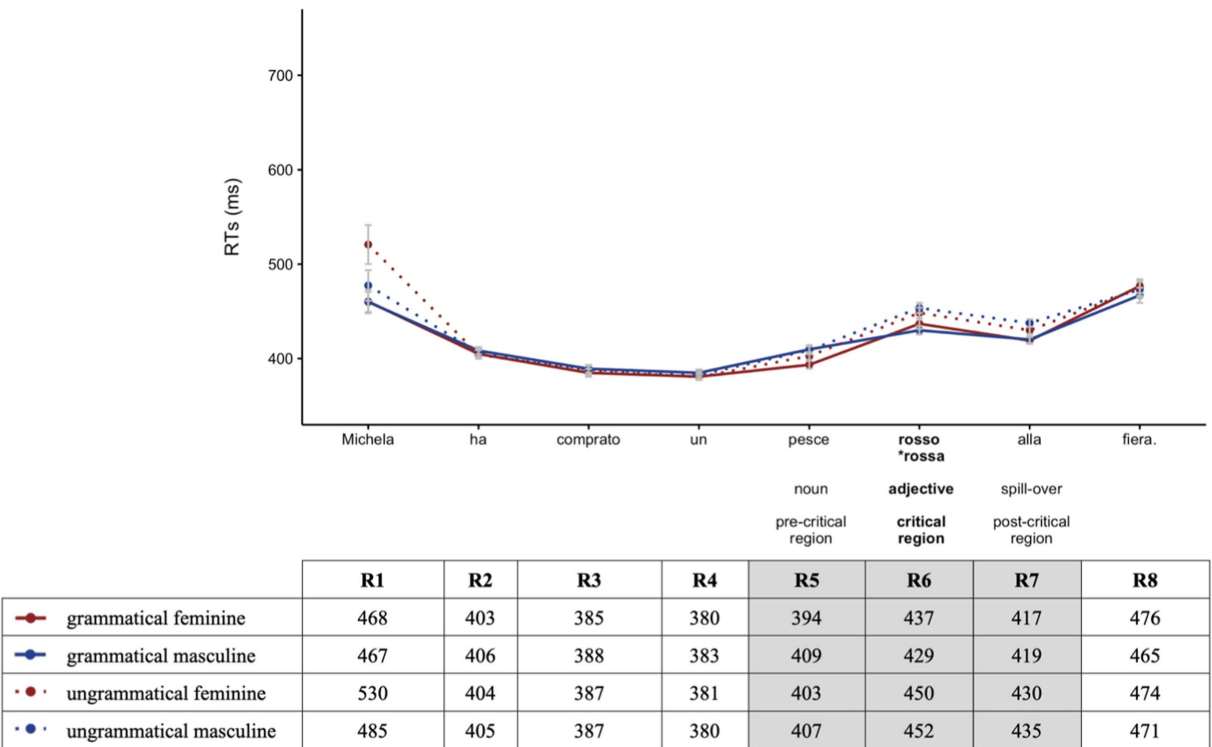


FIGURE 3
Homeland speakers' mean RTs by region for grammatical (solid lines) vs. ungrammatical (dotted lines) sentences for feminine (red) and masculine (blue).

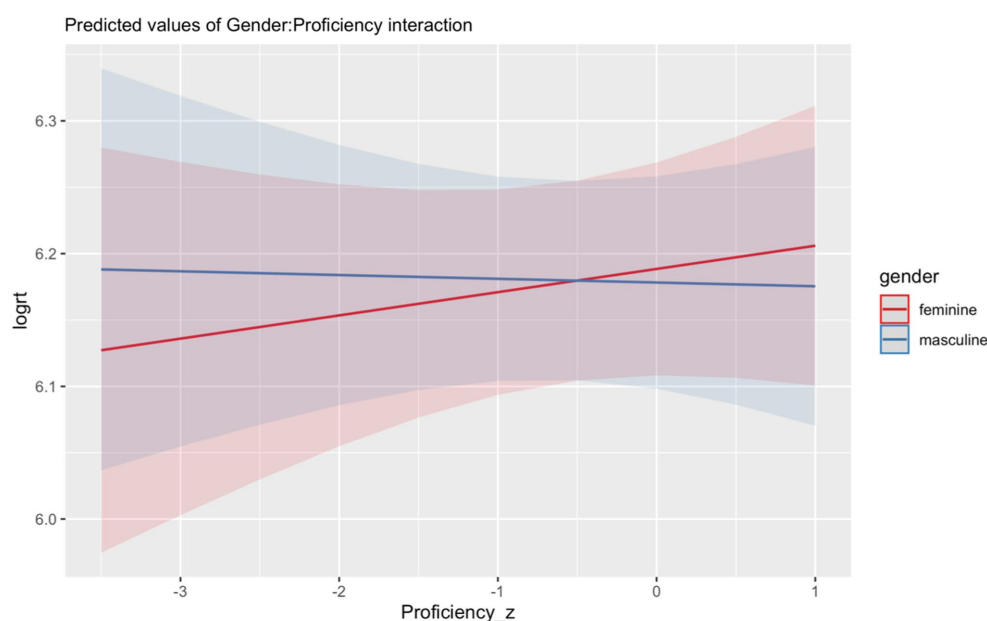


FIGURE 4

Illustration of the interaction between *Gender* (feminine, masculine) and Proficiency scores for Heritage Speakers.

indicating that for HSs ungrammaticality with an unmarked adjective (feminine ungrammatical) led to shorter RTs than ungrammaticality with a marked adjective (masculine ungrammatical).

To investigate whether proficiency and HL use may have affected RTs in HSs, we fitted a model to the heritage group data for Region 6 and Region 7. In Region 6, there were no significant interactions for any of the predictors (p 's > 0.05). In Region 7, the model revealed a main effect of *Grammaticality* ($\text{Chisq} = 6.20$, $p = 0.013$), indicating that HSs were sensitive to ungrammaticalities which were reflected in shorter RTs in the grammatical compared to the ungrammatical condition. The significant interaction between *Grammaticality:Gender* ($\text{Chisq} = 4.81$, $p = 0.028$) confirmed that in the grammatical conditions, there was no difference in RTs between feminine and masculine ($\beta = 0.006$, $SE = 0.020$, $z = 0.334$, $p = 0.987$); however, in the ungrammatical conditions, HSs showed shorter RTs for ungrammatical feminine as compared to ungrammatical masculine ($\beta = -0.060$, $SE = 0.020$, $z = -3.039$, $p = 0.013$). The significant interaction between *Gender:Proficiency* ($\text{Chisq} = 5.21$, $p = 0.023$) indicates that HSs had shorter RTs for sentences with masculine compared to feminine nouns as their proficiency increased (Figure 4).

In summarizing, the online RT data revealed that overall HSs had longer RTs compared to homeland speakers. In the critical and post-critical regions, both groups showed longer RTs for ungrammatical sentences; however, HSs showed significantly longer RTs for feature clash errors when ungrammatical sentences were realized on feminine marked adjectives compared to default errors realized on masculine unmarked adjectives. Finally, proficiency was the only extra-linguistic predictor of HSs' RT

performance, while other extra-linguistic factors were not significant.

Grammaticality judgement task

The results for the GJT are presented in Figure 5 and Table 2. Homeland speakers performed above 90% in all conditions. HSs' performance was above 90% in the grammatical conditions; however, HSs were less likely to judge accurately ungrammatical feminine sentences (default errors) where the violation was realized on masculine unmarked adjectives (52%) compared to ungrammatical masculine sentences (feature clash error) where the violation was realized on feminine marked adjectives (72%).

The model revealed a significant main effect of *Group* ($\text{Chisq} = 31.91$, $p < 0.001$) showing that HSs were overall significantly less accurate as compared to homeland speakers. The significant effect of *Grammaticality* ($\text{Chisq} = 62.99$, $p < 0.001$) shows that both groups were overall significantly more accurate with the grammatical conditions as compared to the ungrammatical ones. The effect of *Gender* ($\text{Chisq} = 21.45$, $p < 0.001$) was also significant and indicates that both HSs and homeland speakers were less accurate with sentences containing feminine marked nouns. The significant interaction between *Group:Grammaticality* ($\text{Chisq} = 24.12$, $p < 0.001$) and subsequent *post-hoc* pairwise comparisons showed that in the grammatical conditions, there was no difference in accuracy between HSs and homeland speakers ($\beta = -0.396$, $SE = 0.304$, $z = -1.305$, $p = 0.192$). However, in the ungrammatical conditions, HSs performed significantly less accurately than homeland speakers ($\beta = -2.585$,

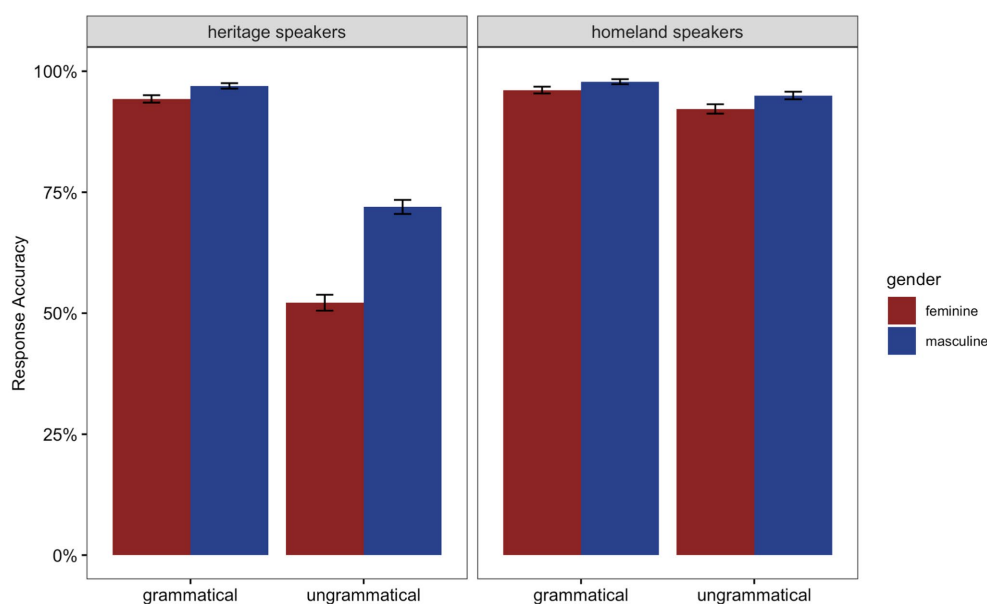


FIGURE 5

Mean response accuracy in percentage for the grammatical and ungrammatical conditions per group in the Grammaticality Judgment Task (GJT). The bars represent the standard error to the mean.

TABLE 2 Mean accuracy scores (%) and SDs per condition for HSs and homeland speakers in the GJT.

Condition	Heritage speakers	Homeland speakers
	M (SD)	M (SD)
Grammatical feminine	94 (0.23)	96 (0.19)
Grammatical masculine	97 (0.17)	98 (0.14)
Ungrammatical feminine	52 (0.50)	92 (0.27)
Ungrammatical masculine	72 (0.45)	95 (0.22)

$SE = 0.337$, $z = -7.668$, $p < 0.001$). There was no significant three-way interaction between *Group:Grammaticality:Gender* ($\text{Chisq} = 0.97$, $p = 0.325$). However, since this was our *a priori* comparison, we run *post-hoc* pairwise comparisons that showed that in the grammatical conditions, the difference in accuracy between feminine and masculine nouns was not different between HSs and homeland speakers ($\beta = -0.112$, $SE = 0.439$, $z = -0.256$, $p = 0.798$). In the ungrammatical conditions, however, the difference in accuracy between feminine and masculine nouns was statistically different between HSs and homeland speakers ($\beta = -0.603$, $SE = 0.297$, $z = -2.030$, $p = 0.042$), indicating that HSs were less accurate with the feminine nouns (where the ungrammaticality was caused by a masculine unmarked adjective—default error) compared to the masculine ones (where the ungrammaticality was caused by a feminine marked adjective—feature clash error).

The model fitted to the heritage group data revealed a main effect of *Grammaticality* ($\text{Chisq} = 227.10$, $p < 0.001$), indicating that HSs were more accurate with grammatical conditions

compared to ungrammatical ones. The main effect of *Gender* ($\text{Chisq} = 26.43$, $p < 0.001$) indicates that HSs were less accurate with sentences containing feminine marked nouns. Furthermore, the model revealed that *Proficiency* ($\text{Chisq} = 15.62$, $p < 0.001$) was a significant predictor of accuracy in the GJT: the higher the score in the vocabulary test, the better the performance in the task. There was also a significant two-way interaction between *Gender* and *HL_home* ($\text{Chisq} = 4.34$, $p = 0.037$), as well as *Grammaticality* and *HL_home* ($\text{Chisq} = 28.20$, $p < 0.001$) as illustrated in [Figures 6C,D](#). These interactions suggest that with more exposure to the HL at home, the smaller the differences are in accuracy between feminine and masculine nouns ([Figure 6C](#)) as well as grammatical and ungrammatical items ([Figure 6D](#)). There was also a significant three-way interaction between *Grammaticality*, *Gender*, and *HL_social* ($\text{Chisq} = 4.35$, $p = 0.037$), and *Grammaticality*, *Gender*, and *Proficiency* ($\text{Chisq} = 4.51$, $p = 0.034$), as illustrated in [Figures 6A,B](#). It appears to be the case that with increasing proficiency, HSs are more sensitive to the distinction between feature clash vs. default error patterns, as shown in [Figure 6A](#) (i.e., the interaction is mainly driven from the difference between feminine and masculine nouns in the ungrammatical condition). Moreover, more exposure to the HL in social contexts seems to modulate the difference in accuracy between feminine and masculine nouns in the grammatical condition, suggesting that with more exposure, HSs have higher accuracy for masculine nouns.

To summarize, the accuracy data show that both groups were more accurate with grammatical conditions as compared

to ungrammatical ones. Furthermore, both HSs and homeland speakers were more accurate with sentences containing masculine unmarked nouns as compared to sentences with feminine marked nouns. There was no difference between groups in the grammatical conditions; however, in the ungrammatical conditions, HSs were more accurate in detecting violations realized on feminine marked adjectives (feature clash errors) compared to violations realized on the masculine unmarked adjectives (default errors). Finally, HSs' proficiency as well language use was predictors of accuracy, whereas the type of bilingualism was not.

Discussion

The present study investigated whether and how morphological markedness influences gender agreement processing in Italian, comparing how this manifests in both native homeland and HSs. Focusing on the HSs, we examined whether RTs and accuracy were calibrated to proficiency, individual HL use, and/or task modality (online vs. offline). To this aim, participants completed an online SPRT and an offline GJT where markedness of the trigger nouns was systematically

manipulated, isolating the potentially unique contribution of markedness to noun-adjective agreement resolution.

To summarize the entirety of the data, RTs revealed that overall HSs were reading at a slower speed compared to homeland speakers. This is not surprising, considering that HSs have less experience with reading in Italian and lower proficiency. In the critical and post-critical regions, both groups showed overall longer RTs during processing of ungrammaticality indicating that the method was successful. In the grammatical conditions, we found no difference in RTs between feminine and masculine nouns between HSs and homeland speakers, whereas in the ungrammatical conditions, only the HSs showed signs that markedness played a specific role in processing agreement. Interestingly, HSs displayed significantly longer RTs when the ungrammatical sentences were realized on feminine marked adjectives (feature clash error) compared to masculine unmarked adjectives (default error), this was not true of the homeland speakers.

The offline data show that both groups were more accurate with grammatical conditions as compared to ungrammatical conditions. Markedness mattered here for both, given higher accuracy with sentences containing masculine unmarked nouns as compared to feminine marked nouns. In the grammatical

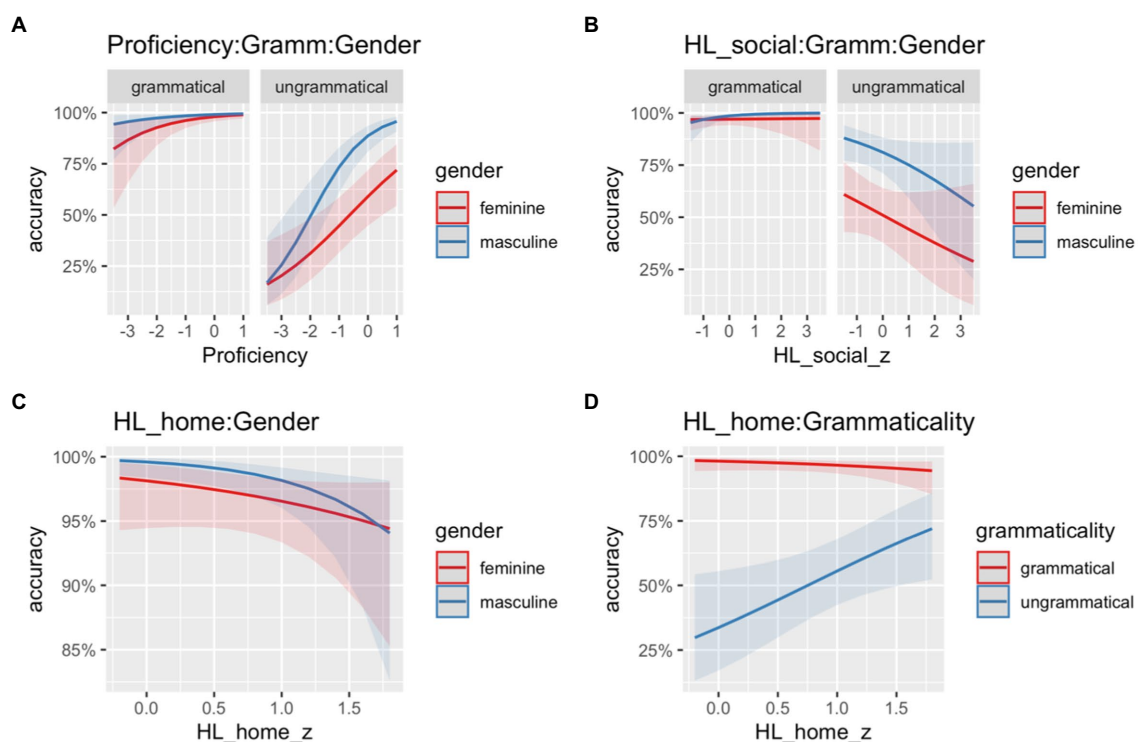


FIGURE 6

(A) Illustration of the three-way interaction between *Proficiency*, *Grammaticality* (grammatical, ungrammatical), and *Gender* (feminine, masculine); (B) Illustration of the three-way interaction between *heritage exposure outside of home context* (HL_social), *Grammaticality* (grammatical, ungrammatical), and *Gender* (feminine, masculine); (C) Illustration of the two-way interaction between *heritage exposure at home* (HL_home) and *Gender* (feminine, masculine); and (D) Illustration of the two-way interaction between *heritage exposure at home* (HL_home) and *Grammaticality* (grammatical, ungrammatical).

conditions, there was no difference between the groups; however, in the ungrammatical conditions while the homeland speakers displayed very high (above 90%) accuracy, HSs were less accurate in detecting violations realized on the masculine unmarked adjectives (ungrammatical feminine—default error) compared to violations realized on feminine marked adjectives (ungrammatical masculine—feature clash error)—the same condition for which they were slower in RTs. That is, there seems to be an apparent *speed-accuracy tradeoff* for HSs conditioned by markedness, a point to which we return below. In both tasks, proficiency was a significant extra-linguistic predictor—speed and accuracy increased as HSs' proficiency increased—while HL use only mattered for accuracy. In the remainder of this section, we discuss how these results fit into the larger context of the relevant literature more generally.

With respect to our first research question concerning whether or not there is an effect of markedness at all, we see clear evidence that markedness matters for HSs and for homeland natives; however, this plays out differently depending on the group and the task. The offline evidence from both groups converges on the fact that masculine is the default gender in Italian (D'Achille, 2003; Corbett and Fraser, 2011). In this respect, the overall present data are in line with other studies reporting higher accuracy with masculine nouns compared to feminine ones for HSs of Romance languages (i.e., Montrul et al., 2008; Alarcón, 2011; Bianchi, 2013; Kupisch et al., 2013; Irizarri van Suchtelen, 2016; Hur et al., 2020) as well as for homeland native speakers and L2 speakers (for Spanish: McCarthy, 2008; for Italian and French: Vigliocco and Franck, 1999, 2001).

However, as it pertains to gender agreement violations—how markedness affects the processing of ungrammaticality—we see a difference between the two groups that at first glance might seem counterintuitive. The results showed that HSs were more sensitive to ungrammaticality when realized on feminine marked adjectives, that is, an error denoting a feature clash running counter to morphological markedness. One might have expected such an error to be as salient, potentially more so, for homeland Italian speakers considering that a feature clash error is argued to be more costly for processing in general and/or attributable to the status of gender specification. Neuroimaging studies, for example Alemán Bañón and Rothman (2016), have shown that homeland natives of Spanish should increase amplitude for these types of errors relative to default agreement ones. Our results are also highly reminiscent of what Fuchs et al. (2015) showed for Spanish homeland speakers, leading them to argue that such patterns provide empirical evidence for the position that masculine in Romance (at least in Spanish) should be evaluated as the absence of a gender specification. Indeed, our results for the HSs speak to the same argumentation for Italian gender, but the fact that the homeland Italian speakers do not align with the Spanish ones in Fuchs et al. (2015) to the same degree leaves us a bit reluctant to make the same conclusions definitively. As alluded to above, our Italian homeland speakers were quite fast in reading these sentences. While the homeland

speakers do show a distinction in overall reading time between grammatical and ungrammatical agreement sentences, the lack of markedness might be attributable to the granularity of what a RT study might be able to reveal as compared to an EEG study as in Alemán Bañón and Rothman (2016) or somehow what could be shown in an auditory acceptability task as in Fuchs and colleagues. Considering the fact that the homeland speakers had above 90% accuracy in both grammatical and ungrammatical conditions and very fast in reading no matter the sentence type, it could be the case that this method is simply unable to tease out any fine-grained effects that markedness might have otherwise conditioned. The case of HSs is distinct precisely because they are not (as) quick readers of Italian and they are not universally at ceiling with all sentence types. Given that they are slower overall and not at ceiling with ungrammatical sentences, an online/RT method had a better chance at revealing an underlying effect for markedness *a priori*. As we noted above, there was a tradeoff between accuracy and speed in the HSs only; the slower the reading, the greater the accuracy with HSs. This tradeoff afforded the HSs the opportunity to process what they were reading, and thus, the effect of markedness had time to reveal itself. Alternatively, or perhaps working in tandem, the fact that HSs might rely more on morphological defaults generally could be significant here. In the present case, a heightened reliance of defaults—a tendency that has been reported also for monolingual (i.e., Pérez-Pereira, 1991) and bilingual (i.e., Eichler et al., 2012) children as well as for L2 learners (i.e., Franceschina, 2001; McCarthy, 2008)—might make feature clash errors even more disruptive for HS processing than for homeland speakers. Coupling a potential HS heightened sensitivity to defaults with the slower overall reading times of HSs and lack of ceiling effects with ungrammatical agreement could have all combined to give rise to the differences we noted in the groups.

Our second research question explored the effects of proficiency and extra-linguistic factors on the processing of gender agreement in HSs. We found that proficiency was a significant extra-linguistic predictor of accuracy and RTs. This is in line with previous studies (i.e., Bianchi, 2013; Kupisch et al., 2013) reporting accuracy for HSs to be modulated by proficiency in the HL. Specifically, the effect of markedness is modulated by proficiency, as shown in Figure 6A, such that the higher the proficiency, the stronger the effect of markedness; thus, HSs are more sensitive to the distinction between feature clash vs. default error patterns. It is worth noting that the proficiency test we used was a measure of lexical knowledge, which is only one dimension of proficiency. However, lexical proficiency has been shown to be a reliable measure for overall language proficiency (i.e., Alderson, 2005) and positive correlations have been found between HSs' lexical knowledge and overall HL proficiency (i.e., Daller et al., 2003; Lloyd-Smith et al., 2019).

Based on previous studies on adult HSs, we were expecting an effect of HL use (i.e., Bianchi, 2013) and AoO of bilingualism (i.e., Montrul, 2008; Keating, 2022); however, we did not find any effect of AoO, whereas HL use affected only accuracy. Specifically, our

results showed that the more use and exposure to the HL at home and in social contexts, the smaller the differences are in accuracy between feminine and masculine nouns. This finding suggests that gender is sensitive to input effects. Our results, supported by HS performance in the GJT, are in line with Bianchi (2013), showing that HL use in adulthood plays a major role during processing of gender agreement, regardless of AoO of bilingualism. Gender in Italian is acquired early (Chini, 1995; Belletti and Guasti, 2015) and it is a considerably transparent system (Kupisch et al., 2002; Padovani and Cacciari, 2003); thus, it is less challenging to acquire and maintain gender in Italian. The reader may recall that in our study, we tested adult HSs who have relatively high proficiency, perhaps significantly so in this European context as compared to studies in other contexts, such as North America, where gender in seemingly linguistically comparable HLs (e.g., Spanish) has been depicted as being vulnerable.

Finally, our third research question investigated whether HSs' use of markedness was affected by task modality (online vs. offline). We hypothesized to find a markedness effect in both tasks; however, we left open the possibility that the degree of the effect would differ between the tasks. Our results showed that HSs were sensitive to markedness in both online and offline comprehension, however in a different way, indicating as discussed above a *speed-accuracy tradeoff* effect. In any task that requires control over both accuracy and responses, participants can optimize either speed or accuracy, or a compromise between the two. Such a compromise leads to increasing speed at the cost of accuracy or increasing accuracy at the cost of speed. This trade-off is widely attested as evidence for development in cognitive control—as children get older and are faced with more challenging tasks, they tend to preserve their accuracy by sacrificing their speed (Davidson et al., 2006; Best et al., 2011). Furthermore, Struys et al. (2018) found that bilinguals tend to show a stronger relationship between speed and accuracy in their performance on cognitive tasks than the monolinguals, suggesting that bilinguals tend to rely more on this optimization strategy to boost their performance. Indeed, bilinguals develop specific strategies to resolve various linguistic conflicts (Muysken, 2013), and our results also corroborate this by demonstrating that HSs chose a strategy to boost accuracy at cost of slower response times on detecting violations conditioned by markedness—a linguistic feature that has been shown to be costly to process for bilinguals.

To conclude, our results indicate that both homeland speakers and HSs access and make use of markedness information during processing of agreement in online and offline sentence comprehension. Most importantly, only HSs showed greater sensitivity to feature clash errors which resulted in slower RTs and higher accuracy on ungrammaticality realized on feminine marked adjectives. Future studies examining gender processing in HSs should consider the effect of markedness both at the level of gender feature on the noun (marked vs. unmarked gender) and errors that involve mismatching marked features (feature clash errors vs. default errors).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

Ethics statement

This study was reviewed by the University of Konstanz Research Ethics Committee and was given a favorable ethical opinion for conduct (IRB 29/2019). The participants provided their written informed consent to participate in this study.

Author contributions

GP, TM, and JR designed and conducted the study. GP and MK analyzed the data. GP, MK, JR, and TM wrote the paper. JR and TM commented and revised the paper. All authors contributed to the article and approved the submitted version.

Funding

This article was supported by generous funding to GP by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska Curie grant agreement No. 765556. JR was funded by the Tromsø Forskningsstiftelse (Tromsø Research Foundation) starting grant No. A43484 and the Heritage-bilingual Linguistic Proficiency in their Native Grammar (HeLPiNG; 2019–2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.965885/full#supplementary-material>

References

- Alarcón, I. V. (2011). Spanish gender agreement under complete and incomplete acquisition: early and late bilinguals' linguistic behavior within the noun phrase. *Biling. Lang. Cogn.* 14, 332–350. doi: 10.1017/S1366728910000222
- Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.
- Alemán Bañón, J., Miller, D., and Rothman, J. (2017). Morphological variability in second language learners: an examination of electrophysiological and production data. *J. Exp. Psychol. Learn. Mem. Cogn.* 43, 1509–1536. doi: 10.1037/xlm0000394
- Alemán Bañón, J., and Rothman, J. (2016). The role of morphological markedness in the processing of number and gender agreement in Spanish: an event-related potential investigation. *Lang. Cogn. Neurosci.* 31, 1273–1298. doi: 10.1080/23273798.2016.1218032
- Anderson, J. A., Mak, L., Chahi, A. K., and Bialystok, E. (2018). The language and social background questionnaire: assessing degree of bilingualism in a diverse population. *Behav. Res. Methods* 50, 250–263. doi: 10.3758/s13428-017-0867-9
- Antón-Méndez, I., Nicol, J. L., and Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax* 5, 1–25. doi: 10.1111/1467-9612.00045
- Anwyl-Irvine, A. L., Massoné, J., Flitton, A., Kirkham, N. Z., and Evershed, J. K. (2020). Gorilla in our midst: an online behavioural experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Battistella, E. L. (1990). *Markedness: The Evaluative Superstructure of Language*. Albany, NY: SUNY Press.
- Bayram, F., Di Pisa, G., Rothman, J., and Slabakova, R. (2021). “Current trends and emerging methodologies in charting heritage language bilingual grammars,” in *The Cambridge Handbook of Heritage Languages*. eds. S. A. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 545–577.
- Bayram, F., Rothman, J., Iverson, M., Kupisch, T., Miller, D., Puig-Mayenco, E., et al. (2019). Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *Int. J. Biling. Educ. Biling.* 22, 919–939. doi: 10.1080/13670050.2017.1324403
- Belletti, A., and Guasti, M. T. (2015). *The Acquisition of Italian: Morphosyntax and Its Interfaces in Different Modes of Acquisition*, Vol. 57. Amsterdam: John Benjamins Publishing Company.
- Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., et al. (2005). Colfis-Corpus e Lessico di Frequenza dell'Italiano Scritto. Available at: <https://www.istc.cnr.it/en/grouppage/colfis>
- Best, J. R., Miller, P. H., and Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learn. Individ. Differ.* 21, 327–336. doi: 10.1016/j.lindif.2011.01.007
- Bianchi, G. (2013). Gender in Italian–German bilinguals: a comparison with German L2 learners of Italian. *Biling. Lang. Cogn.* 16, 538–557. doi: 10.1017/S1366728911000745
- Cacciari, C. (2011). *Psicologia del Linguaggio. Psychology of Language*. Bologna: Il Mulino.
- Chini, M. (1995). *Genere Grammaticale e Acquisizione. Aspetti Della Morfologia Nominale Dell'Italiano L2*. Milano: Franco Angeli.
- Corbett, G. (1991). *Gender*. Cambridge: Cambridge University Press.
- Corbett, G., and Fraser, N. (2011). “Default genders,” in *Gender in Grammar and Cognition: I: Approaches to Gender. II: Manifestations of Gender*. eds. B. Unterbeck, M. Rissanen, T. Nevalainen and M. Saari (Berlin, New York: De Gruyter Mouton), 55–98.
- Costa, A., Kovačić, D., Franck, J., and Caramazza, A. (2003). On the autonomy of the grammatical gender systems of the two languages of a bilingual. *Biling. Lang. Cogn.* 6, 181–200. doi: 10.1017/S1366728903001123
- D'Achille, P. (2003). *L'italiano contemporaneo*. Bologna: Il Mulino.
- Daller, H., van Hout, R., and Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Appl. Linguis.* 24, 197–222. doi: 10.1093/applin/24.2.197
- Davidson, M. C., Amso, D., Anderson, L. C., and Diamond, A. (2006). Development of cognitive control and executive functions from 4 to 13 years: evidence from manipulations of memory, inhibition, and task switching. *Neuropsychologia* 44, 2037–2078. doi: 10.1016/j.neuropsychologia.2006.02.006
- DeLuca, V., Rothman, J., and Pliatsikas, C. (2019). Linguistic immersion and structural effects on the bilingual brain: a longitudinal study. *Biling. Lang. Cogn.* 22, 1160–1175. doi: 10.1017/S1366728918000883
- Deutsch, A., and Bentin, S. (2001). Syntactic and semantic factors in processing gender agreement in Hebrew: evidence from ERPs and eye movements. *J. Mem. Lang.* 45, 200–224. doi: 10.1006/jmla.2000.2768
- Durrell, M. (2011). *Hammer's German Grammar and Usage*. London: Routledge.
- Eichler, N., Hager, M., and Müller, N. (2012). Code-switching within determiner phrases in bilingual children: French, Italian, Spanish and German. *Zeitschrift für französische Sprache und Literatur* 122, 227–258.
- Franceschina, F. (2001). Morphological or syntactic deficits in near-native speakers? An assessment of some current proposals. *Second. Lang. Res.* 17, 213–247. doi: 10.1177/026765830101700301
- Fuchs, Z. (2019). Gender in the nominal domain: Evidence from bilingualism and eye-tracking. Doctoral dissertation. Harvard University.
- Fuchs, Z. (2021). Facilitative use of grammatical gender in heritage Spanish. *Linguistic Approaches to Bilingualism*. doi: 10.1075/lab.20024.fuc [Epub ahead of print].
- Fuchs, Z., Polinsky, M., and Scontras, G. (2015). The differential representation of number and gender in Spanish. *Linguist. Rev.* 32, 703–737. doi: 10.1515/trl-2015-0008
- Gagarina, N., and Klassert, A. (2018). Input dominance and development of home language in Russian-German bilinguals. *Front. Commun.* 3:40. doi: 10.3389/fcomm.2018.00040
- Giancaspro, D. (2017). “Heritage speakers' production and comprehension of lexically- and contextually-selected subjunctive mood morphology,” *Doctoral dissertation*. New Brunswick, NJ: Rutgers-The State University of New Jersey.
- Goebel-Mahrle, T., and Shin, N. L. (2020). A corpus study of child heritage speakers' Spanish gender agreement. *Int. J. Biling.* 24, 1088–1104. doi: 10.1177/1367006920935510
- Hur, E., Lopez Otero, J. C., and Sanchez, L. (2020). Gender agreement and assignment in Spanish heritage speakers: does frequency matter? *Language* 5:48. doi: 10.3390/languages5040048
- Irizarri van Suchtelen, P. (2016). Spanish as a heritage language in the Netherlands. A cognitive linguistic exploration. Utrecht: LOT.
- Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007
- Keating, G. D. (2022). The effect of age of onset of bilingualism on gender agreement processing in Spanish as a heritage language. *Lang. Learn.* 1–39. doi: 10.1111/lang.12510 [Epub ahead of print]
- Keating, G. D., and Jegerski, J. (2015). Experimental designs in sentence processing research: a methodological review and user's guide. *Stud. Second. Lang. Acquis.* 37, 1–32. doi: 10.1017/S0272263114000187
- Köpcke, K. M. (1982). Untersuchungen zum Genusystem der deutschen Gegenwartssprache Research on the gender system of present-day German. Niemeyer.
- Kramer, R. (2015). *The Morphosyntax of Gender*. Vol. 58. Oxford: Oxford University Press.
- Kunkel-Razum, K., Münzberg, F., Gallmann, P., Nübling, D., Barz, I., Fabricius-Hansen, C., et al. (2009). *Duden-Die Grammatik*. Mannheim: Dudenverlag
- Kupisch, T., Akpınar, D., and Stöhr, A. (2013). Gender assignment and gender agreement in adult bilinguals and second language learners of French. *Linguist. Approach. Bilingual.* 3, 150–179. doi: 10.1075/lab.3.2.02kup
- Kupisch, T., Müller, N., and Cantone, K. F. (2002). Gender in monolingual and bilingual first language acquisition: comparing Italian and French. *Lingue e linguaggio* 1, 107–150.
- Laleko, O. (2018). What is difficult about grammatical gender? Evidence from heritage Russian. *J. Lang. Contact* 11, 233–267. doi: 10.1163/19552629-01102004
- Lenth, R. V. (2022). Emmeans: estimated marginal means, aka least-squares means. R package version 1.7.5. Available at: <https://CRAN.R-project.org/package=emmeans>
- Lloyd-Smith, A., Bayram, F., and Iverson, M. (2020). The effects of heritage language experience on lexical and morphosyntactic outcomes. *Stud. Turkish Herit. Lang.* 60:63. doi: 10.1075/sibil.60.04llo
- Lloyd-Smith, A., Einfeldt, M., and Kupisch, T. (2019). Italian-German bilinguals: the effects of heritage language use on accent in early-acquired languages. *Int. J. Biling.* 24, 289–304. doi: 10.1177/1367006919826867

- Lloyd-Smith, A., Gyllstad, H., Kupisch, T., and Quaglia, S. (2021). Heritage language proficiency does not predict syntactic CLI into L3 English. *Int. J. Biling. Educ. Biling.* 24, 435–451. doi: 10.1080/13670050.2018.1472208
- Marinis, T. (2010). “Using on-line processing methods in language acquisition research,” in *Experimental Methods in Language Acquisition Research*. eds. E. Blom and S. Unsworth (Amsterdam: John Benjamins Publishing Company), 139–162.
- Marsden, E., Thompson, S., and Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Appl. Psycholinguist.* 39, 861–904. doi: 10.1017/S01421716418000036
- McCarthy, C. (2008). Morphological variability in the comprehension of agreement: an argument for representation over computation. *Second. Lang. Res.* 24, 459–486. doi: 10.1177/0267658308095737
- Montrul, S. (2008). *Incomplete Acquisition in Bilingualism: Re-Examining the Age Factor*. Vol. 39. Amsterdam: John Benjamins Publishing Company.
- Montrul, S., Davidson, J., De La Fuente, I., and Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Biling. Lang. Cogn.* 17, 118–138. doi: 10.1017/S1366728913000114
- Montrul, S., De La Fuente, I., Davidson, J., and Foote, R. (2013). The role of experience in the acquisition and production of diminutives and gender in Spanish: evidence from L2 learners and heritage speakers. *Second. Lang. Res.* 29, 87–118. doi: 10.1177/0267658312458268
- Montrul, S., Foote, R., and Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition. *Lang. Learn.* 58, 503–553. doi: 10.1111/j.1467-9922.2008.00449.x
- Müller, N. (1994). “Gender and number agreement within DP,” in *Bilingual First Language Acquisition: French and German grammatical development*. ed. J. M. Meisel (Amsterdam: John Benjamins Publishing Company), 53–88.
- Muysken, P. (2013). Language contact outcomes as the result of bilingual optimization strategies. *Biling. Lang. Cogn.* 16, 709–730. doi: 10.1017/S1366728912000727
- Nevins, A., Dillon, B., Malhotra, S., and Phillips, C. (2007). The role of feature-number and feature-type in processing Hindi verb agreement violations. *Brain Res.* 1164, 81–94. doi: 10.1016/j.brainres.2007.05.058
- Padovani, R., and Cacciari, C. (2003). Il ruolo della trasparenza morfologica nel riconoscimento di parole in Italiano [The role of morphological transparency in recognition of Italian words]. *G. Ital. Psicol.* 30, 749–771. doi: 10.1421/11106
- Pérez-Pereira, M. (1991). The acquisition of gender: what Spanish children tell us. *J. Child Lang.* 18, 571–590. doi: 10.1017/S0305000900011259
- Polinsky, M. (2008). Gender under incomplete acquisition: heritage speakers’ knowledge of noun categorization. *Herit. Lang. J.* 6, 40–71. doi: 10.46538/hlj.6.1.3
- Polinsky, M. (2018). *Heritage Languages and Their Speakers*, Vol. 159. Cambridge: Cambridge University Press.
- R Core Team (2016). R: A language and environment for statistical computing. R Foundation for statistical computing, Vienna, Austria. Available at: <https://www.R-project.org/>
- Rothman, J. (2007). Heritage speaker competence differences, language change, and input type: inflected infinitives in heritage Brazilian Portuguese. *Int. J. Biling.* 11, 359–389. doi: 10.1177/13670069070110040201
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814
- Schwarze, C. (2009). *Grammatica Della Lingua Italiana*. Roma: Carocci.
- Singmann, H., Bolker, B., Westfall, J., Aust, F., and Ben-Shachar, M. S. (2022). Afex: analysis of factorial experiments. R package version 1.1-1. Available at: <https://CRAN.R-project.org/package=afex>
- Steinmetz, D. (2006). Gender shifts in Germanic and Slavic: semantic motivation for neuter? *Lingua* 116, 1418–1440. doi: 10.1016/j.lingua.2004.06.014
- Struys, E., Duyck, W., and Woumans, E. (2018). The role of cognitive development and strategic task tendencies in the bilingual advantage controversy. *Front. Psychol.* 9:1790. doi: 10.3389/fpsyg.2018.01790
- Titone, D. A., and Tiv, M. (2022). Rethinking multilingual experience through a systems framework of bilingualism. *Bilingual. Lang. Cogn.* 1–16. doi: 10.1017/S1366728921001127 [Epub ahead of print]
- Torregrossa, J., Andreou, M., Bongartz, C., and Tsimpli, I. M. (2021). Bilingual acquisition of reference. The role of language experience, executive functions and cross-linguistic effects. *Biling. Lang. Cogn.* 24, 694–706. doi: 10.1017/S1366728920000826
- Van Osch, B., Hulk, A., Sleeman, P., and van Suchtelen, P. I. (2014). Gender agreement in interface contexts in the oral production of heritage speakers of Spanish in the Netherlands. *Linguist. Netherlands* 31, 93–106. doi: 10.1075/avt.31.08osc
- Vasishth, S., and Nicenboim, B. (2016). Statistical methods for linguistic research: foundational ideas—part I. *Lang. Ling. Compass* 10, 349–369. doi: 10.1111/lnc3.12201
- Velnić, M. (2020). Acquisition of a transparent gender system: a comparison of Italian and Croatian. *Front. Psychol.* 11:571674. doi: 10.3389/fpsyg.2020.571674
- Vigliocco, G., and Franck, J. (1999). When sex and syntax go hand in hand: gender agreement in language production. *J. Mem. Lang.* 40, 455–478. doi: 10.1006/jmla.1998.2624
- Vigliocco, G., and Franck, J. (2001). When sex affects syntax: contextual influences in sentence production. *J. Mem. Lang.* 45, 368–390. doi: 10.1006/jmla.2000.2774
- Wickham, H. (2016). *ggplot 2: Elegant Graphics for Data Analysis*. New York: Springer. Available at: <https://ggplot2.tidyverse>.



OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway,
Norway

REVIEWED BY

Ian Cunnings,
University of Reading,
United Kingdom
Marie Thérèse Le Normand,
Institut National de la Santé et de la
Recherche Médicale, France

*CORRESPONDENCE

Serkan Uygun
serkan.uygun@es.bau.edu.tr

SPECIALTY SECTION

This article was submitted to
Cognition,
a section of the journal
Frontiers in Psychology

RECEIVED 07 July 2022

ACCEPTED 27 October 2022

PUBLISHED 15 November 2022

CITATION

Uygun S (2022) Processing pro-drop
features in heritage Turkish.
Front. Psychol. 13:988550.
doi: 10.3389/fpsyg.2022.988550

COPYRIGHT

© 2022 Uygun. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License \(CC
BY\)](#). The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Processing pro-drop features in heritage Turkish

Serkan Uygun*

Faculty of Educational Sciences, Department of English Language Teaching, Bahçeşehir University,
Istanbul, Turkey

Previous studies have reported that null subject is not completely lost in heritage speakers, but there is an increase in the production and acceptance of overt subjects. Turkish is a pro-drop language and as a typical feature of pro-drop languages, it requires obligatory verb agreement marking for sentences with null subjects. However, Turkish subject-verb agreement marking is an example of optional agreement in which the 3rd person plural subject has optionality and can be used with singular verb forms under certain conditions. The current study investigates the reading times (RTs) of plural-marked and unmarked verbs in sentences with overt and null subjects during real time sentence processing in comparison to non-heritage speakers of Turkish via a self-paced reading experiment. Significant differences were observed between the heritage and non-heritage speakers of Turkish indicating both quantitative and qualitative real-time processing differences between the two groups. These differences suggest that Turkish heritage speakers need more time to integrate the information in real time processing.

KEYWORDS

heritage speakers, Turkish, pro-drop, subject-verb agreement, sentence processing, morphosyntactic knowledge

Introduction

The original version of the Interface Hypothesis (IH) predicts increased vulnerability for bilinguals in phenomena, where syntax interacts with other modules of language, known as the interfaces (Sorace and Filiaci, 2006). On the other hand, no problems/difficulties were expected for purely syntactic phenomena. The revised version of the IH (Sorace and Serratrice, 2009) made a linguistically principled distinction between internal and external interfaces. The internal interfaces mainly integrate modules that pertain to formal grammar, such as syntax, semantics, and morphology and their interactions (e.g., morphosyntax). On the other hand, external interfaces combine linguistic modules that are related to general cognition and/or world knowledge, like discourse and pragmatics. In other words, external interfaces involve interactions between linguistic and non-linguistic domains (e.g., syntax-pragmatics). According to this version, only external interfaces are expected to be particularly vulnerable/problematic because external interfaces integrate domains from different language levels, which leads to a higher processing load. Phenomena located at external interfaces are predicted to be vulnerable/problematic in bilingual populations either because of their less detailed knowledge or less automatic access to computational constraints within the language

module, or because they have fewer cognitive resources available (Sorace, 2011). According to Tsimplici and Sorace (2006), the main difference between internal and external interfaces is that only the latter requires a higher level of language use because of integrating domains outside of the formal grammar. Sorace (2011) also claims that the IH makes explicit claims for the heritage speakers (HS) at the level of ultimate attainment. Therefore, the claims of IH can be applied to HS as they are an important testing ground for the IH (Montrul and Polinsky, 2011). HS are defined as individuals who were raised in homes where a language other than the dominant community language was spoken, resulting in some degree of bilingualism in both the heritage and the community language (Scontras et al., 2018). Recent studies with HS have suggested that they have control over the rules of particular modules (syntax, phonology), but they have difficulty when integrating grammatical and non-grammatical information (Benmamoun et al., 2013). While Sorace and Serratrice (2009) and Sorace (2011) advocate that structures that lie at the external interfaces are vulnerable/problematic for HS with their higher processing load, Benmamoun et al. (2013) questions whether other interfaces are also affected under heritage language conditions without making a distinction between internal and external interfaces. Their conclusions suggest that HS experience difficulties/problems when they have to compute interface properties. They observed that HS struggle with operations that involve computation across more than one grammatical component (e.g., syntax and morphology) because interface operations in general require knowledge of the principles and constraints operating on both components, together with the ways in which they map onto each other (Benmamoun et al., 2013, p. 165).

Psycholinguistic research on bilingualism has mainly focused on the representation and processing of structures that require the integration of knowledge from different linguistic domains for over 20 years. The appropriate use of some structures cannot be merely determined by syntactic rules, but requires the integration of knowledge from other domains. A typical example is the pro-drop phenomenon. Pro-drop is defined as the deletion of the overt subject in a sentence in cases when the subject may be recovered from the pragmatics and the context of the sentence or from the person information on the verb (Altan, 2013). This feature is usually seen in languages with a rich inflectional morphology. Pro-drop languages habitually use overt subjects mainly to mark pragmatic information such as contrast, emphasis or topic shift (e.g., Enç, 1986). In other words, the selection of the overt and null subjects requires both syntactic and discourse-pragmatic knowledge, which makes the pro-drop phenomenon relevant to the syntax-pragmatics interface, which is an external interface.

The majority of the previous studies with HS have examined the contact between the pro-drop and the non-pro-drop languages mainly focusing on the cross-linguistic influence. Studies on a variety of languages that allow both overt and null subjects (e.g., Italian, Greek, Spanish, and Russian) have reported that null subject expression is not completely lost in HS with the

exception of severe cases of attrition, HS display an increase in the production and acceptance of overt subjects, they lose or display variability in the discourse-pragmatic constraints on overt subjects in the pro-drop language and they have a tendency to use overt subjects in pragmatically redundant contexts, for example, when a referent is not marked for contrast, emphasis or topic shift (Silva-Corvalán, 1994; Polinsky, 1995; Montrul, 2004; Tsimplici et al., 2004; Sorace, 2005; Albirini et al., 2011). According to Polinsky (2018), these results are not surprising, because overt subjects do not change much in the structure of a heritage language (HL). Overt subjects are not ungrammatical in the baseline and they allow HS to be clearer in production because HS exhibit a preference for one-to-one mapping between form and function. The overall use of overt subjects is consistent with the general tendency for overmarking observed in HL. In addition, several researchers have observed that the use of null subjects is already diminished in the speech of first-generation immigrants, whose language serves as input to HS (Otheguy et al., 2007; Montrul, 2016). The decreasing use of null subjects in the input also contributes to the erosion of null subjects in the HL. While most of the existing studies that have compared overt and null subjects have tested the use of overt pronouns, the present study focuses on the processing of sentences with overt subject noun phrases (NPs) and null subjects in Turkish HS.

Pro-drop in Turkish

Turkish is a morphologically rich language in which verbs must agree with the subject in person and number, and the subject position of a sentence or a noun phrase does not need to be filled overtly with a phonologically realized noun phrase (Kornfilt, 1984; Özsoy, 1987). As a pro-drop language, Turkish may have clauses without overt subjects and the discourse-pragmatic context determines the choice between overt and null subjects (Enç, 1986; Taylan, 1986; Kerslake, 1987; Özsoy, 1987; Turan, 1995). Speakers of Turkish usually maintain referents with a null subject as in (b) taken from Azar et al. (2020):

- (a) *Murat dün sinema-ya git-ti.*
Murat yesterday cinema-DAT go-PAST.3SG
“Murat went to the cinema yesterday.”
- (b) *Ø film-i beğen-me-miş.*
Ø movie-ACC like-NEG-PAST.EV.3SG
“(He) did not like the movie.”

The example in (b) shows that the empty pronominal (Ø) is the counterpart of the overt subject pronoun (he) and its referent is determined by the subject-verb agreement (SVA) marking. If an overt subject pronoun or NP were used in example (b), this would not affect the truth value of the sentence because both versions (with null subject and overt subject pronoun or NP) of example (b) carry the same truth value. This illustrates that as long as the

referent can be recovered from the context, the speaker may use a null subject in the sentence.

The 3rd person pronouns in Turkish (singular: *o* → he/she/it; plural *onlar* → they) do not encode gender or animacy (Azar et al., 2020). When a null subject is used for a 3rd person plural pronoun (e.g., as for *çocuklar* “children” from example c), the verb must always be marked with the 3rd person plural marker (*-lar/ler*) to avoid any ambiguity on the subject referent.

- (c) *Çocuk-lar okul-dan çık-tı-(lar)*
Child-PL school-ABL leave-PAST-(3PL)
“Children left the school.”
- (d) *Ø ev-e git-ti-ler.*
Ø home-DAT go-PAST-3PL
“(They) went home.”
- (e) *Ø ev-e git-ti.*
Ø home-DAT go-PAST.3SG
“(He/She/It) went home.”

The empty pronominal in example (d) refers to the 3rd person plural pronoun *çocuklar* “children” in (c) because the verb is marked with the 3rd person plural marker. In contrast, the empty pronominal in example (e) does not refer to the 3rd person plural pronoun *çocuklar* “children” in (c) because the verb is not marked correctly, leading the subject referent to be infelicitous. While the correct interpretation of the empty pronominal must be *they* with the verb being marked correctly as in example (d), when the verb is not marked with the 3rd person plural suffix, the empty pronominal would be interpreted as *he/she/it* as in example (e).

On the other hand, when a referent is pragmatically marked for emphasis, contrast or topic shift, the overt subject is usually preferred over the null subject (Enç, 1986), as in (g) where the subject referent is marked for contrast and is expressed with an overt subject pronoun, *o* “she” instead of a null subject, which is taken from Azar et al. (2020):

- (f) *Aynı film-i Suzan da izle-miş.*
same movie-ACC Suzan too watch-PAST.EV.3SG
“Suzan also watched the same movie.”
- (g) *Ama o çok beğen-miş.*
but she a lot like-PAST.EV.3SG
“But she liked it a lot.”

Previous studies in Turkish

The initial analyses on the use of overt and null subjects, which mainly focused on the overt and null pronouns in Turkish, were either theoretical (Enç, 1986; Taylan, 1986; Özsoy, 1987) or collected data from fiction novels (Kerslake, 1987; Turan, 1995). A few studies investigated the acquisition of null subjects in monolingual Turkish children. Slobin and Talay (1986) examined the speech transcripts of nine children between the ages of 24 to 56 months. They claimed that the subject in Turkish can

be encoded by SVA marking alone or by an overt subject. Based on the analyses of all child utterances, they proposed that by the age of 24 months, SVA is correctly marked on verbs across a range of tenses, and both overt and null subjects are used by the children. They also found that both the morphosyntactic (SVA marking) and pragmatic (overt vs. null subjects) knowledge for marking the subject of a sentence are well established at early ages. Altan (2006, 2013) explored the use of null subjects in Turkish children grouped into three different age groups: age 2, age 3 and age 4 and replicated the results of Slobin and Talay (1986). She also observed that when children produce null subject sentences with 3rd person plural pronouns, they are adding the 3rd person plural marker on the verb correctly. Example (i) shows how a 2,8-year-old child participant marked the verb correctly although he was not expected to drop the subject pronoun because the experimenter was specifically asking about the subject.

- (h) *O-nu kim al-dı san-a?*
that-ACC who buy-PAST.3SG you-DAT
“Who bought that for you?”
- (i) *Ø al-dı-lar.*
Ø buy-PAST-3PL
“(They) bought.”

There are also studies that focused on the use of null subjects in bilingual Turkish-speaking children who also speak a non-pro-drop language. Some of these studies found that Turkish children in contact situations are similar to monolingual Turkish children because they not only use the overt and null subjects to the same extent (Verhoeven, 1990; Aarssen, 1996) but also benefit effectively from the pragmatic context that requires the use of overt subjects (Özcan et al., 2000). Conversely, there are also studies that revealed differences between monolingual and bilingual Turkish-speaking children. In one of those studies, Haznedar (2010) collected spontaneous Turkish data from one Turkish-English bilingual child and one Turkish monolingual child. The results of the data comparison revealed that the Turkish-English bilingual child’s production of overt subjects in Turkish is more than the control child. This finding could be interpreted as due to cross-linguistic influence from English regarding the suppliance of overt subjects in the context of Turkish-English bilingual acquisition. In another study, Sağın Şimşek (2010) compared the use of overt and null subjects in four Turkish-German bilingual and four Turkish monolingual children who were between the ages of four to eight. The researcher found that the bilingual children use overt subjects more than the monolingual controls and attributed this result to the influence of German, which is a non-pro-drop language.

There are also studies that explore the use of overt and null subjects in adult non-heritage Turkish speakers. For example, Azar and Özyürek (2015) used two silent videos to elicit narration from non-heritage Turkish speakers (non-HS). The only personal pronoun the researchers observed in the experiment was 3rd person singular pronoun (*o* → he/she/it) and they found that Turkish non-HS prefer overt subjects significantly more than null subjects

to reintroduce subject referents. However, in the maintenance context, non-HS used null subjects significantly more. Similar findings were obtained in Azar et al. (2016, 2017) that used the same methodology. In another narration elicitation study, Azar et al. (2019) also observed sentences with 3rd person plural pronoun (*onlar* → they) and found that when the 3rd person plural pronoun is reintroduced, they are reintroduced with a null subject and the verb is marked with 3rd person plural marker to avoid any ambiguities regarding the subject referent as in example (k).

- (j) *Üç tane bayan var.*
three piece woman exist
“There are three women.”
- (k) *Ø yemek yap-ıyor-lar.*
Ø food cook-PRES.CONT-3PL
“(They) are cooking.”

As a result of the studies with adult non-heritage Turkish speakers, the researchers concluded that the use of subject in Turkish is primarily limited by pragmatic purposes; that is, overt subjects to mark emphasis and null subjects in the context of maintenance. These results clearly indicate that Turkish is a pro-drop language and in line with previous research in other pro-drop languages, null subject is the default form to maintain reference (e.g., Carminati, 2002).

Research has also been conducted with Turkish HS to examine the contact between pro-drop Turkish and a non-pro-drop language such as Dutch or English. Doğruöz (2007) analyzed the spoken corpora of Turkish in the Netherlands and Turkish in Turkey and found no effect of contact between the languages when the quantity of the subject pronouns in informal interviews were compared. Doğruöz and Backus (2009) analyzed the use of subject pronouns in informal interviews with Turkish HS living in the Netherlands and could not find any cross-linguistic effects when the amount of overt subjects is taken into consideration. They only observed a 2% of redundant overt subject use in the heritage data. In another study, Koban Koç (2016) interviewed HS living in New York City and the results showed that HS use overt subjects significantly higher than the non-HS living in Turkey. By using narrative elicitation of two silent videos, Azar et al. (2017, 2020) concluded that Turkish HS living in the Netherlands were similar to non-HS because HS perform similar to non-HS in terms of the discourse status or pragmatic constraints in the use of pronouns during speech. However, they also found a difference between the groups because HS use overt subjects slightly more than their non-HS peers. In addition, they observed that the reintroduction of a 3rd person plural pronoun is done with a null subject and the verb being correctly marked (example m).

- (l) *İki kız masa-da sebze doğru-yor.*
two girl table-LOC vegetable slice-PRES.CONT.3SG
“Two girls are slicing vegetables on the table.”
- (m) *Ø bir kavanoz aç-ma-ya çalış-ıyor-lar.*
Ø a jar open-VN-DAT try-PRES.CONT-3PL

“(They) are trying to open a jar.”

As can be seen, previous studies in Turkish have employed offline methods such as spoken corpora analysis, spontaneous speech, narrative elicitation and informal interviews and these studies have mainly focused on the conditions of the contextual and discourse/pragmatic appropriateness on the use of overt and null subjects because these conditions (i.e., external interface) regulate the choice of overt vs. null subjects in all pro-drop languages including Turkish. The HS studies have also investigated the contextual and discourse/pragmatic appropriateness *via* offline methods. Overall, the results are inconclusive and do not provide further evidence for the vulnerability/difficulty of external interfaces observed in HS. However, none of these studies have focused on the use of optional SVA marking (i.e., internal interface) with 3rd person plural subjects in sentences with overt subject NPs and null subjects. Previous research has shown that HS have difficulties with SVA marking (Benmamoun et al., 2013; Polinsky, 2018) making this phenomena of Turkish, that displays optionality, an interesting testing ground.

Optional SVA marking in Turkish

Another important aspect in the use of null subjects is the SVA marking. Pro-drop languages typically display a rich inflectional morphology which allows for subjects to be dropped because agreement governs the empty *pro* category and helps to recover the dropped subjects (Cherici, 2021). Turkish verbs agree with the subjects in person and number (Enç, 1986) and Turkish marks subject agreement on the verbal element by means of a person suffix (Taylan, 1986). Like most agreement-marking languages, singular subjects require singular verb forms and plural subjects require plural verb forms. However, Turkish SVA marking is an example of optional agreement in which the 3rd person plural subject has optionality and can be used with singular (unmarked) verb forms under certain conditions. Example (n) illustrates this optionality, where the omission of plural suffix *-lar/ler* renders the verb form indistinguishable from the 3rd person singular form.

- (n) *Öğrenci-ler okul-dan gel-di-(ler).*
student-PL school-ABL come-PAST-(3PL)
“Students came from school.”

Turkish non-HS usually have a tendency to avoid using the same morphological marker within the same clause or phrase. According to Johanson (1998), this is a general tendency in Turkic languages with an attempt to use morphological devices economically and avoid redundancy. Previous research with Turkish non-HS has shown that for sentences with 3rd person plural subjects, the acceptability of a plural marker on the verb is affected by semantic and pragmatic factors such as the subject's degree of animacy (e.g., Schroeder, 1999; Bamyacı et al., 2014).

While animate plural subjects may take either a plural or an unmarked verb (example o), inanimate plural subjects usually take an unmarked verb (example p) (Sezer, 1978). Using plural marked or unmarked verb forms with animate plural subjects depends on the speaker's stylistic preferences without a difference regarding the meaning (Sezer, 1978; Kornfilt, 1997).

- (o) *Çilingir-ler kapı-lar-ı aç-tı-(lar)*.
locksmith-PL door-PL-ACC open-PAST-(3PL)
“Locksmiths opened the door.”
- (p) *Anahtar-lar kapı-lar-ı aç-tı-(*lar)*
Key-PL door-PL-ACC open-PAST-(*3PL)
“Keys opened the door.”

As for the HS, several acceptability judgment studies have been conducted. For example, Bamyacı (2016) found that they are similar to non-HS in the way they treat SVA with 3rd person plural subjects, but she also observed that HS have a greater likelihood of accepting plural-marked verbs. Lago et al. (2019) found that while non-HS prefer unmarked verb forms with animate 3rd person plural subjects, HS accept both plural-marked and unmarked forms to a similar extent. Recently, Uygun and Felser (2021) reported that HS rate plural verb forms better when the subject was animate, but did not find a difference in the overall acceptance of plural-marked vs. unmarked verbs between HS and non-HS.

To summarize, previous studies on the optional SVA marking with HS have displayed several differences when compared to the non-HS. These results are in line with the predictions that phenomena displaying optionality are more affected under HL conditions resulting in differences when compared to non-HS (Benmamoun et al., 2013). All of the above-mentioned studies provide us information about the metalinguistic judgment of the HS when there was no time constraint, yet it is not known how they would perform in sentences with overt subject NPs and null subjects when their reaction times are measured.

The study

Since most existing studies that have compared overt and null subjects have tested the use of overt pronouns, the main purpose of the current study was to investigate the reading times (RTs) of plural-marked and unmarked verbs in sentences with overt subject NPs and null subjects during real time sentence processing. While previous studies have always focused on the contextual and discourse/pragmatic appropriateness of using overt vs. null subjects (i.e., external interface), this is the first study to explore the optional SVA marking (i.e., internal interface) in a pro-drop language. While the verb can be either plural-marked or unmarked in sentences with 3rd person animate plural subjects, the verb must be always plural-marked for sentences when a null pronoun replaces the 3rd person animate plural subject. By employing a self-paced reading experiment, it was expected to gain more insights into implicit

processing preferences for the optional SVA marking in sentences with both overt subject NPs and null subjects and make a comparison between heritage and non-heritage speakers of Turkish. Since offline tasks do not offer direct access to participants' mental processes as they unfold in real time, it was decided to use an online task, which measures participants' automatic responses to language stimuli, providing a more direct access to how language processing unfolds in real time (Bayram et al., 2021). In a self-paced reading task, participants read sentences presented to them one word or phrase at a time on the computer screen. According to Bayram et al. (2021), the main goal of this task is not to make a quantitative comparison by focusing on the reading times of HS and non-HS group but to understand whether HS process their HL qualitatively different from non-HS of the same language.

The following research questions were sought to investigate:

1. Is there a difference between Turkish HS and non-HS in their RTs of plural marked and unmarked verbs in sentences with overt subject NPs and null subjects?
2. Is the optional SVA marking in Turkish, which is considered as an internal interface, vulnerable/problematic to acquire and cause a processing load for HS?

Based on the previous results, non-HS are expected to show no RT differences for plural marked vs. unmarked verbs in sentences with overt subject NPs because they have no problems with the optional SVA marking in Turkish. Conversely, for null subject sentences, non-HS are expected to display longer RTs for unmarked verbs because unmarked verbs cause a mismatch between the subject and the verb. As for the HS, they are expected to show shorter RTs for plural marked verbs in sentences with overt subject NPs which is indicative of their problems with the optional agreement marking, providing evidence for their struggle with operations that involve computation across more than one grammatical component (e.g., syntax and morphology) and that internal interfaces are also vulnerable for them (Benmamoun et al., 2013). In addition, they are expected to perform similar to the non-HS in sentences with null subjects as SVA marking is compulsory. If no difference is observed between HS and non-HS in sentences with overt subject NPs and null subjects, this would support the revised version of the IH (Sorace and Serratrice, 2009) by showing that internal interfaces are not difficult to acquire and process.

Materials and methods

Participants

Forty non-heritage Turkish speakers (non-HS) were recruited and tested in Istanbul. All non-HS participants were born and raised in Turkey and they had never lived abroad. One non-HS participant was excluded due to high error rates in the filler condition (> 30%). The remaining 39 non-HS participants

(mean age = 36.87, SD = 9.21, age range = 19–60, 29 females) were either university graduates or were studying at the university at the time of testing and they all spoke the standard dialect of Turkish. 60 Turkish heritage speakers (HS) residing in Berlin and Potsdam were recruited and tested. All of the HS in the study were exposed to Turkish from birth and spoke both Turkish and German in their daily lives. One participant was excluded due to low Turkish proficiency score (below 12 out of 20 indicating a proficiency level lower than B2 level) from the Turkish TELC (The European Language Certificates) test which is designed for B2 level based on the Common European Framework of Reference (CEFR). The language structure part of the TELC test consists of two cloze tests with 20 questions in total. The remaining data of 59 HS (mean age = 27.78, SD = 6.06, age range = 19–50, 42 females) were analyzed. All HS had an early age of acquisition of German (mean age = 3.01, SD = 1.85, age range = 0–6) and a high score from the Turkish TELC (mean score = 18.44, SD = 1.62, score range = 13–20). In addition, the HS group self-rated their weekly use of Turkish and the results show a predominant use of Turkish in a normal week covering reading, writing, speaking, and listening, with a mean rate of 61.61% (SD = 21.82, range = 15–90). The HS group also self-rated their Turkish proficiency level out of 10 across four language skills and the results revealed a high proficiency level for the HS (Speaking: mean = 7.91, SD = 1.63; Listening: mean = 8.84, SD = 1.17; Writing: mean = 7.17, SD = 2.04; Reading: mean = 8.21, SD = 1.70). Both the TELC scores and the self-ratings indicate a high level of Turkish proficiency in the HS. All participants received a small fee for their participation.

Materials

The present experiment had a factorial design with two within-participant factors and group as the between-participant factor. By manipulating the existence of the subject (null subject vs. overt subject NP) and verb marking (plural-marked vs. unmarked), 24 experimental sentence sets were created with four different conditions as illustrated in (q–t). All experimental sentences had a context sentence and the subject was always a 3rd person plural animate subject to investigate the optional SVA marking. For null subject sentences, the context sentence is in the SOV order and the verb is always unmarked. The context sentences in the present experiment do not aim to explore the contextual and discourse/pragmatic appropriateness on the use of overt and null subjects. The context sentence below for examples (q) and (r) actually facilitates the use of null subjects. The NS-PL condition in example (q) is correct because the subject in the context sentence *polisler* “policemen” is a 3rd person plural animate subject and the verb in example (q) has the plural marker *-ler* in the end (*götürdüler* “took”) and therefore the subject referent is unambiguous. In contrast, in the NS-SG condition in

example (r), the verb is unmarked (*götürdü* “took”) causing a mismatch between the subject referent and the verb and making the subject referent infelicitous.

Context sentence for null subject sentences:

Polis-ler dün genç hırsız-ı yakala-dı.

police-PL yesterday young thief-ACC catch-PAST.3SG.

“The policemen caught the young thief yesterday.”

(q) Null subject – Plural (NS-PL):

Hırsız-ı karakol-a götür-dü-ler ama hırsız kaç-tı.

thief-ACC police station-DAT take-PAST-3PL but thief run away-PAST.3SG

“(The policemen) took the thief to the police station, but the thief ran away.”

(r) Null subject – Singular (NS-SG):

Hırsız-ı karakol-a götür-dü ama hırsız kaç-tı.

thief-ACC police station-DAT take-PAST.3SG but thief run away-PAST.3SG

“(The policemen) took the thief to the police station, but the thief ran away.”

The same context sentence for overt subject NP sentences is transformed into the passive voice without providing the agent, which facilitates the use of an overt subject NP for sentences in examples (s) and (t). In both OS-PL and OS-SG conditions, the subject is a 3rd person plural animate subject (*polisler* “policemen”). This means that the verb can be used as both plural-marked as in example (s) (*götürdüler* “took”) or unmarked as in example (t) (*götürdü* “took”). Both versions are grammatically correct and they do not differ in terms of meaning.

Context sentence for overt subject sentences:

Dün genç hırsız yakala-n-dı.

yesterday young thief catch-PSV-PAST.3SG.

“The young thief was caught yesterday.”

(s) Overt subject – Plural (OS-PL):

Polis-ler hırsız-ı karakol-a götür-dü-ler ama hırsız kaç-tı.

police-PL thief-ACC police station-DAT take-PAST-3PL but thief run away-PAST.3SG

“The policemen took the thief to the police station, but the thief ran away.”

(t) Overt subject – Singular (OS-SG):

Polis-ler hırsız-ı karakol-a götür-dü ama hırsız kaç-tı.

police-PL thief-ACC police station-DAT take-PAST.3SG but thief run away-PAST.3SG

“The policemen took the thief to the police station, but the thief ran away.”

Four different presentation lists were created in a Latin-square design and the items in each version were pseudo-randomized and mixed with 48 filler sentences, resulting in a total of 72 items per list.

Design and procedure

The experiment was designed on the web-based platform Ibox Farm (Drummond, 2013) and the sentences were presented word-by-word using the noncumulative moving window paradigm (Just et al., 1982). Each trial began with a screen presenting a sentence in which the words were masked by underscores. When the participant pressed the space bar button, a word was revealed and the previous word was re-masked. After reaching the final word of the sentence which appeared with a full stop, the participants pressed the space bar button again to decide if the second sentence was a grammatically and semantically good continuation of the context sentence by pressing “f” for “yes” and “j” for “no” response. After their response, they had to press the space bar button for the next trial.

The experiment began by requesting participants to complete a demographic questionnaire and give their consent. Then, they were instructed to read the sentences carefully and answer the questions as quickly as possible. Five practice items were presented to familiarize the participants with the procedure. Participants received a link to the experiment and completed the test on their personal computers. A progress bar shown above the sentences allowed them to keep track of their progress. The experiment took approximately 20 min, and the HS group additionally completed the Turkish proficiency test afterwards.

Results

The dependent measures were word-by-word RTs of different regions in the experimental sentences. The main interest was in obtaining significant group differences in these regions. For word-by-word reading data, RTs exceeding 2.5 standard deviations above and below a participant's mean log reading time were deemed outliers and removed (HS group = 2.76%; non-HS group = 3.11%). To counter the problems of word length and individual differences in reading times, residual reading times (RRTs) were calculated on the remaining data with linear modelling on the log transformed RTs. Positive values indicate that a reading time is slower than expected whereas negative values indicate a faster reading time. RRTs were analysed for five regions of interest: the critical region of “The verb,” the pre-critical region of “Before the verb” and the three “Spillover” regions following the critical region (see Table 1 below for the regions, analyses were conducted starting from Region 3).

Statistical analyses were conducted with R, an open-source programming language and environment for statistical computing (R Development Core Team, 2017). The RRTs data were analyzed with linear mixed-effects regression models with crossed random effects for items and subjects (Baayen et al., 2008). The models were fitted using the package *lme4* (Bates et al., 2015). The models included the subject-level variable “Group” (HS vs. non-HS) and the item-level variables “The Existence of Subject” (null subject vs. overt subject NP) and “Verb Marking” (plural marked vs.

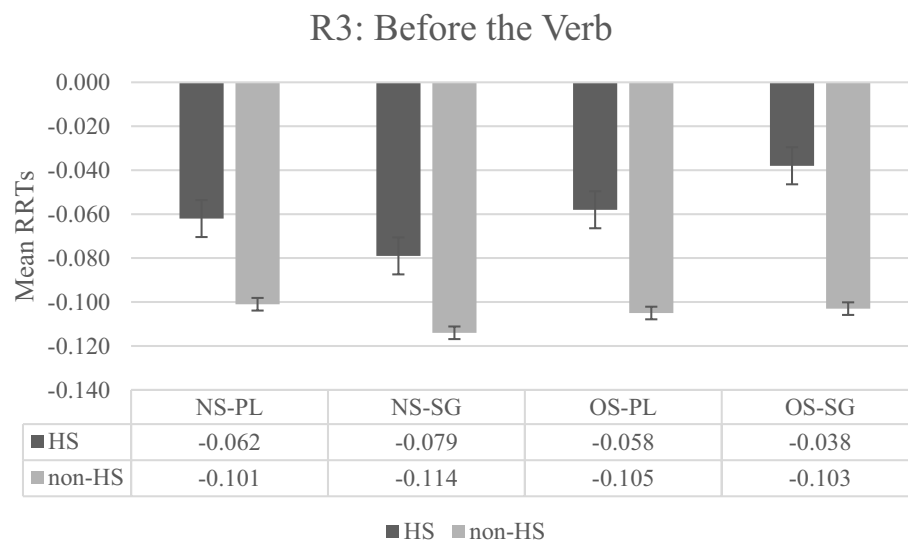
TABLE 1 Regions of interest in the experimental sentences.

Regions	Null subject	Overt subject	Example
Region 1	<i>Not applicable</i>	Subject	Polisler (<i>The policemen</i>)
Region 2	Object	Object	hırsız (<i>the thief</i>)
Region 3	Before the verb	Before the verb	karakola (<i>to the police station</i>)
Region 4	The verb	The verb	götürdü(ler) (<i>took-SG or PL</i>)
Region 5	Spillover 1	Spillover 1	ama (<i>but</i>)
Region 6	Spillover 2	Spillover 2	hırsız (<i>the thief</i>)
Region 7	Spillover 3	Spillover 3	kaçtı (<i>ran away</i>)

unmarked) as fixed effects. The model also included random intercepts for item and subject. For the main effects and overall interactions, sum-coded contrasts (−0.5, 0.5) were employed to the factors Group, The Existence of the Subject and Verb Marking. For single comparisons, treatment contrasts were applied. Initially, a model with random intercepts and slopes for all fixed effects and their interactions was constructed and when this maximal model failed to converge, it was gradually simplified until convergence was reached (Barr et al., 2013). The Akaike Information Criterion (AIC) was used for model comparison because it provides a measure that penalizes complexity and leads to predictors being kept only when they substantially contribute to explaining variance in the data (Venables and Ripley, 2002). The model with the lower AIC value was selected and this procedure was repeated until the simplification process did not produce a model with a lower AIC. The final version of the model included by item and by subject random slopes for the interaction of the existence of the subject and verb marking. The effect sizes are reported by using model coefficients in log odds (β), standard errors (SE), *t*-statistics and *p* values. *P*-values were computed by using the *lmerTest* package and the Satterthwaite's approximation for denominator degrees of freedom (Kuznetsova et al., 2014).

The first region of interest in the experiment is region 3 (Before the verb), the region prior to the critical (see Figure 1) region. The RRTs analysis of this region indicate a significant main effect of group only (see Table 2). The non-HS group had significantly faster RRTs than the HS group (β : 0.046, SE: 0.016, $t = 2.958$, $p < 0.005$) in the pre-critical region “Before the verb.”

The second region of interest is region 4 (The verb), which is the critical region in the experiment (see Figure 2). In this region, significant main effects of verb marking (β : −0.068, SE: 0.020, $t = -3.368$, $p < 0.001$) and group (β : −0.079, SE: 0.023, $t = -3.510$, $p < 0.001$) and a significant three-way interaction of the existence of subject, verb marking, and group (β : 0.140, SE: 0.055, $t = 2.528$, $p < 0.013$) were obtained (see Table 3). The effect of verb marking indicates that plural-marked verbs receive significantly shorter RTs than the unmarked verbs and HS had significantly faster RTs than the non-HS group. To resolve the significant existence of subject, verb marking and group interaction, each group was

**FIGURE 1**

Mean RRTs of both groups for Region 3 (Before the verb). RRTs, Residual reading times; HS, Heritage speakers; non-HS, Non-heritage speakers; NS, Null subject; OS, Overt subject; PL, Verb is plural-marked; SG, Verb is unmarked.

TABLE 2 Linear mixed effects model output for Region 3 (Before the Verb).

	β	SE	<i>t</i>	<i>P</i>
Intercept	-0.082	0.011	-7.226	0.000*
Subject (Null subject vs. Overt subject)	-0.014	0.012	-1.223	0.222
Verb marking (Plural-marked vs. Unmarked)	0.001	0.013	0.091	0.928
Group (HS vs. non-HS)	0.046	0.016	2.958	0.004*
Subject*Verb marking	0.026	0.029	0.898	0.370
Subject*Group	-0.016	0.023	-0.725	0.468
Verb marking*Group	-0.005	0.025	-0.190	0.850
Subject*Verb marking*Group	0.022	0.048	0.457	0.648

Formula in R: RTresidual ~ Subject*Verb marking*Group + (1 + Subject*Verb marking | item) + (1 + Subject*Verb marking | subject). Bold values indicate the significant results.

analyzed separately. For the non-HS group, a significant main effect of verb marking (β : -0.072, *SE*: 0.032, t = -2.243, p < 0.027) and a significant interaction between the existence of subject and verb marking (β : 0.097, *SE*: 0.048, t = 2.009, p < 0.045) were obtained. Plural-marked verbs receive significantly shorter RTs than unmarked verbs in general. The significant interaction reveals that in null subject sentences, unmarked verbs take significantly longer to read than plural-marked verbs (β : -0.119, *SE*: 0.044, t = -2.753, p < 0.007); however, in overt subject NP sentences, unmarked verbs take numerically longer to read (β : -0.022, *SE*: 0.039, t = -0.562, p < 0.575). For the HS group, there was only a significant main effect of verb marking (β : -0.064, *SE*: 0.021, t = -3.091, p < 0.003) indicating that plural-marked verbs receive significantly shorter RTs than unmarked verbs.

Figure 3 illustrates the RRTs for region 5 (Spillover 1), which comes right after the critical region. A significant main effect of the existence of subject (β : 0.033, *SE*: 0.015, t = 2.197, p < 0.029)

and a significant interaction between the existence of subject and verb marking (β : -0.091, *SE*: 0.032, t = -2.864, p < 0.005) were obtained in this region (see Table 4). In general, null subject sentences receive significantly longer RTs than overt subject NP sentences. Regarding the significant interaction, in null subject sentences, unmarked verbs take significantly longer than plural-marked verbs (β : -0.068, *SE*: 0.022, t = -3.109, p < 0.003) while in overt subject NP sentences, plural-marked verbs take numerically longer to read (β : 0.023, *SE*: 0.019, t = 1.215, p < 0.225). In addition, when the verb was unmarked, null subject sentences take significantly longer to read (β : 0.079, *SE*: 0.023, t = 3.427, p < 0.001). However, there is no significant difference when the verb was plural-marked (β : -0.012, *SE*: 0.017, t = -0.723, p < 0.471).

Finally, as illustrated by Figure 4 and 5, there were no significant main effects or interactions in the last two regions, namely region 6 and 7 (Spillover 2 and 3).

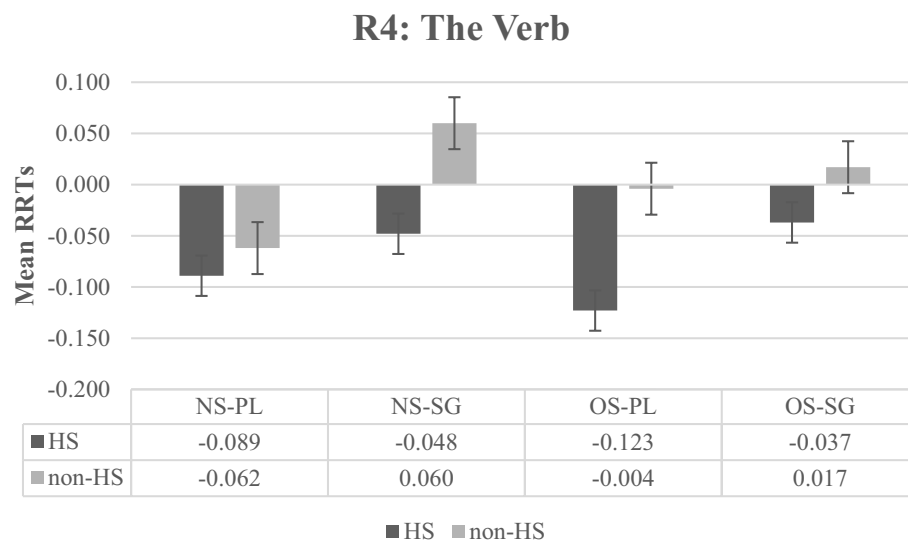


FIGURE 2

Mean RRTs of both groups for Region 4 (The verb). RRTs, Residual reading times; HS, Heritage speakers; non-HS, Non-heritage speakers; NS, Null subject; OS, Overt subject; PL, Verb is plural-marked; SG, Verb is unmarked.

TABLE 3 Linear mixed effects model output for Region 4 (The Verb).

	β	SE	t	P
Intercept	-0.033	0.019	-1.756	0.080
Subject (Null subject vs. Overt subject)	0.004	0.015	0.265	0.792
Verb marking (Plural-marked vs. Unmarked)	-0.068	0.020	-3.368	0.000*
Group (HS vs. non-HS)	-0.079	0.023	-3.510	0.000*
Subject*Verb marking	-0.026	0.035	-0.729	0.466
Subject*Group	0.017	0.028	0.628	0.530
Verb marking*Group	0.008	0.030	0.281	0.778
Subject*Verb marking*Group	0.140	0.055	2.528	0.012*

Formula in R: $R^2_{\text{residual}} \sim \text{Subject*Verb marking*Group} + (1 + \text{Subject*Verb marking} \mid \text{item}) + (1 + \text{Subject*Verb marking} \mid \text{subject})$.

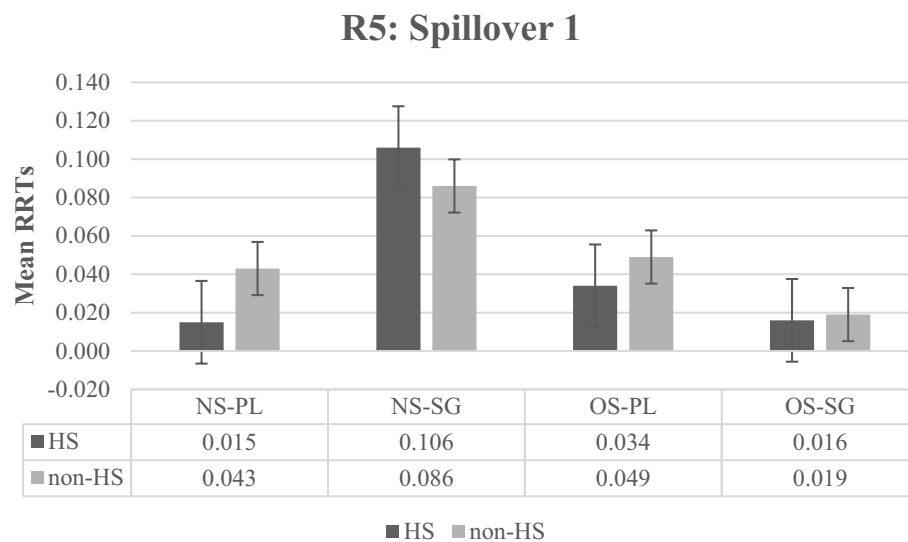
Discussion

By carrying out an online experiment that measures the RTs, the present study tried to explore the RRTs for plural-marked and unmarked verbs in order to investigate the optional SVA marking with heritage and non-heritage speakers of Turkish when they read sentences with overt subject NPs and null subjects.

The results suggest both quantitative and qualitative differences between the HS and non-HS groups in the critical region “The verb.” First of all, HS are significantly faster than non-HS in this region. Secondly, the HS group has significantly faster RTs for the plural marked verbs in general, indicating that their RTs are not affected by the manipulation of the existence of the subject. This also means that HS prefer plural-marked verbs both in overt subject NP and null subject sentences. While this tendency is correct for the null subject sentences to prevent the subject referent ambiguity, it clearly shows their difficulty in using the optional SVA marking in overt subject NP sentences.

Conversely, the non-HS group displays a significant interaction of the existence of subject and verb marking in this region. Similar to the HS group, non-HS favor plural-marked verbs in sentences with null subject to keep the subject referent unambiguous. However, for sentences with overt subject NP, the non-HS group behaves differently from the HS group because there is no significant RT difference between plural-marked and unmarked verbs, which indicates the use of optional SVA marking with no difficulty.

Yet, in the “Spillover 1” region, which comes right after the critical region “The verb,” both groups behave similarly and there are not any quantitative or qualitative differences. First of all, there is no significant RT difference between the groups. Secondly, in this region, the existence of the subject affects both groups in the same way with significantly longer RTs for sentences with null subject. In addition, the significant interaction of the existence of subject and verb marking reveals that unmarked verbs take significantly longer to read than

**FIGURE 3**

Mean RRTs of both groups for Region 5 (Spillover 1). RRTs, Residual reading times; HS, Heritage speakers; non-HS, Non-heritage speakers; NS, Null subject; OS, Overt subject; PL, Verb is plural-marked; SG, Verb is unmarked.

TABLE 4 Linear mixed effects model output for Region 5 (Spillover 1).

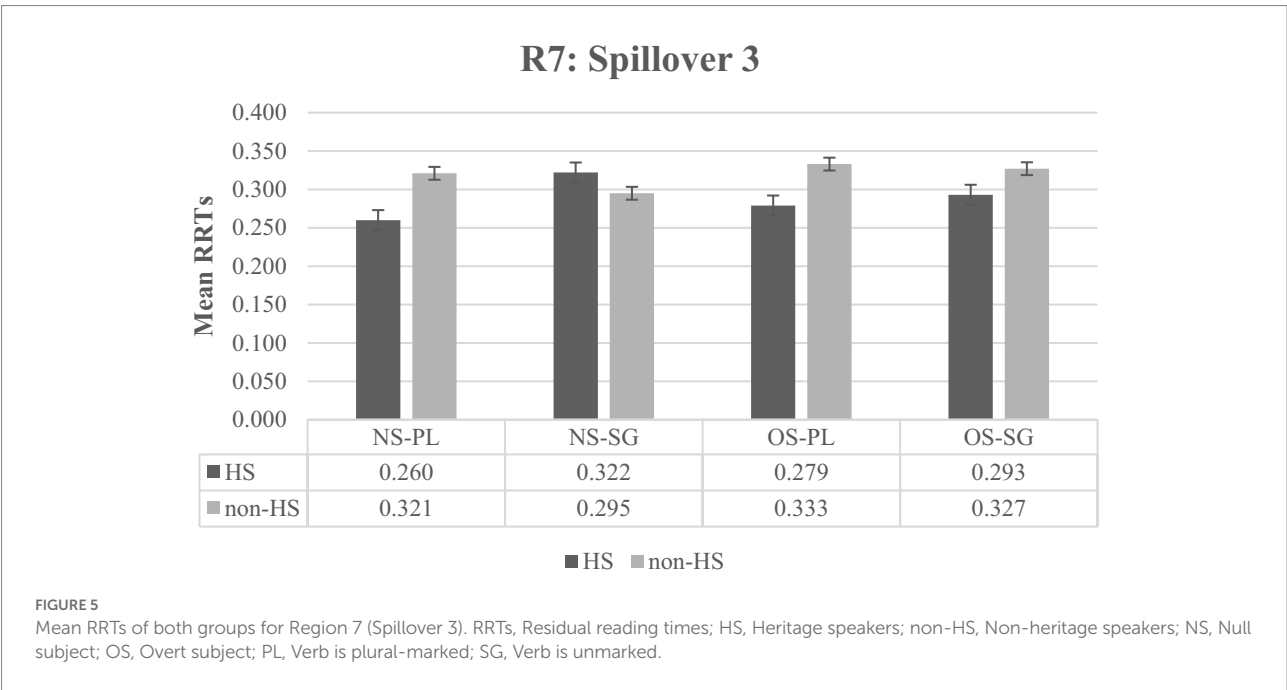
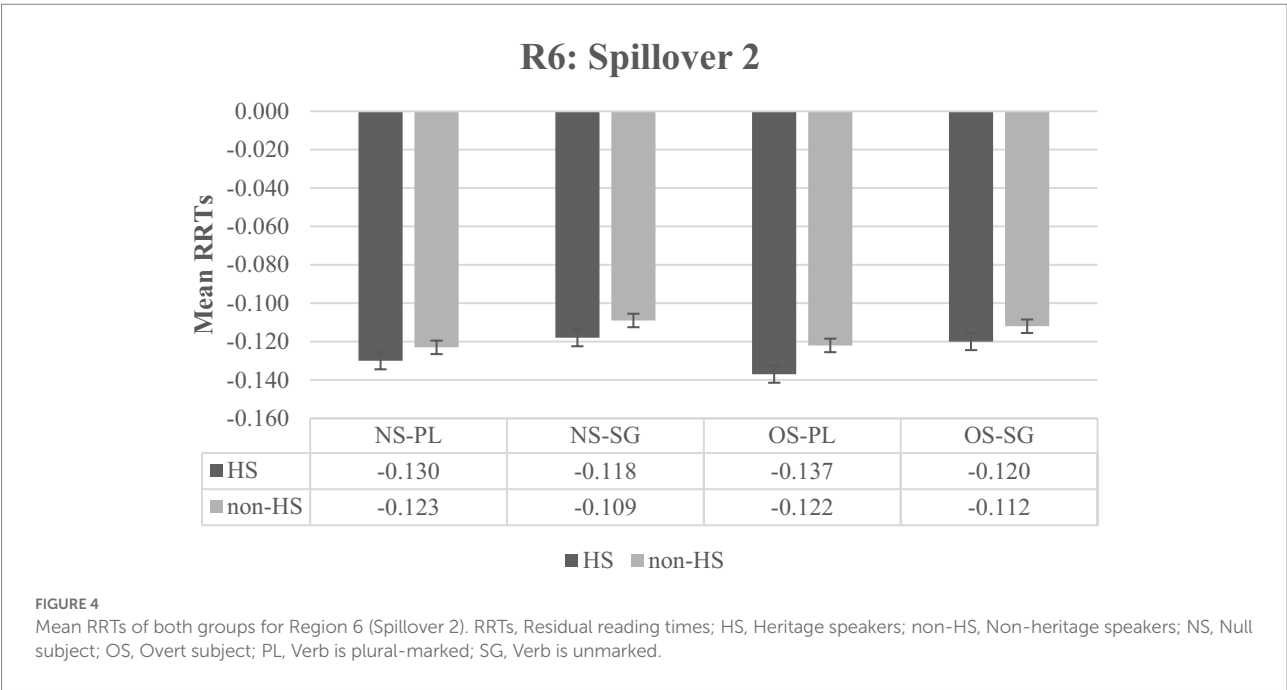
	β	SE	t	P
Intercept	0.046	0.007	6.255	0.000*
Subject (Null subject vs. Overt subject)	0.033	0.015	2.197	0.028*
Verb marking (Plural-marked vs. Unmarked)	−0.022	0.016	−1.367	0.172
Group (HS vs. non-HS)	−0.006	0.012	−0.537	0.592
Subject*Verb marking	−0.091	0.032	−2.864	0.004*
Subject*Group	0.005	0.027	0.186	0.852
Verb marking*Group	−0.027	0.027	−1.014	0.310
Subject*Verb marking*Group	−0.039	0.057	−0.677	0.498

Formula in R: $RTR_{residual} \sim \text{Subject} * \text{Verb marking} * \text{Group} + (1 + \text{Subject} * \text{Verb marking} | \text{item}) + (1 + \text{Subject} * \text{Verb marking} | \text{subject})$. Bold values indicate the significant results.

plural-marked verbs in null subject sentences because they cause a mismatch between the subject and the verb, but there is no significant difference between plural-marked and unmarked verbs in overt subject NP sentences indicating no difficulty for the optional SVA marking.

How can the observed between-group difference in the critical region “The verb” be accounted for? Recall that the IH makes a clear distinction between internal and external interfaces for bilinguals including HS. While internal interfaces involve interactions between language modules (e.g., syntax and morphology), external interfaces have interactions between linguistic and non-linguistic cognitive systems (e.g., syntax and discourse). The IH predicts processing limitations to be affected only in external interfaces because structures that require internal mappings are less taxing than structures that require external mappings (Sorace, 2011, 2012). In contrast to this view, Benmamoun et al. (2013) claims that HS experience problems when they have to compute interface properties without making

a distinction between internal and external interfaces. There are studies that found evidence to support the claims of Benmamoun et al. (2013). For example, Benmamoun (2000) investigated the construct state used to form genitive construction in Arabic and observed that HS do not treat the construct state as a single prosodic unit because they were using double marking. This divergence indicated that HS fail to compute the internal interface between syntax and phonology. Internal interface effects have also been observed by Albirini et al. (2011) in Arabic agreement and coordination, which relies on the interaction between syntax and the morpho-phonological component of the grammar. The authors suggested that HS could no longer control the interface between syntax and the morpho-phonology in their grammars. In another study, Mendez et al. (2015) found that HS perform differently from non-HS with internal interface properties as well, claiming that internal interface properties are also difficult to acquire. In addition, Gondra (2022) concluded that HS are vulnerable to both internal (syntactic-semantic) and external



(pragmatic-discursive) interfaces. Another example of the internal interface is the optional SVA marking, which the current study investigates. According to [Benmamoun et al. \(2013\)](#), HS are expected to have difficulty with interfaces between syntax and morphology and these interfaces are predicted to be more difficult to acquire or more vulnerable to attrite. The data in the critical region “The verb” shows that HS have difficulties with the internal interface of syntax (sentences with overt subject NPs vs. null subjects) and morphology (plural-marked vs. unmarked verbs)

because they behave differently from non-HS when they have to integrate these two language modules.

In addition, it is also known that controlling two languages has significant impacts on linguistic and general cognitive abilities leading to several advantages and disadvantages for bilinguals ([Sorace, 2011](#)). Because both languages are simultaneously activated in the bilingual mind even in cases when one is contextually unnecessary ([Marian et al., 2003](#); [Bialystok, 2009](#)), bilingual processing is predicted to be less

efficient than monolingual processing. Sorace (2011) claims that bilinguals are less efficient than monolinguals because their knowledge of or access to computational constraints within the language module is less detailed and/or less automatic than in monolinguals and because they have fewer general cognitive resources to deploy on the integration of different types of information in online language comprehension and production. According to Sorace (2011), accessing and integrating two types of knowledge is *more costly* than accessing only one type of knowledge and the problem mainly lies in the bilinguals' less optimal ability to consistently and effectively integrate different types of knowledge. Since HS are considered as a subgroup of bilinguals, these claims are directly relevant to HS as well although this problem is expected to be smaller for HS in comparison to L2 speakers. The online integration of different types of knowledge may incur a cost for HS as they may be less efficient in integrating diverse knowledge when compared to the non-HS group. In the current experiment, the online integration of syntactic knowledge (whether the sentence has an overt subject NP vs. null subject) and morphological knowledge (whether the verb is plural-marked or unmarked) is a demanding task that requires a lot of cognitive demands (Rothman and Slabakova, 2011). Because HS fail to integrate these different types of knowledge successfully, they are found to be significantly different from the non-HS group in the critical region "The verb." Polinsky and Scontras (2020) recently proposed that HS are likely to face difficulty with phenomena that impose cognitive demands as a result of their processing resource limitations. They claim that HS restructure their grammar to free up processing resources resulting in a change in their grammar. The limited nature of their processing resources in the non-dominant language forces their grammar to be less ambiguous, more regular and having less structure. For the optional SVA marking, "the restructuring of grammar" means that HS try to regularize the optional SVA system by over-using plural suffixes in contexts in which non-HS prefer to use the unmarked verb forms. This limitation may explain why the integration process of syntactic and morphological knowledge incurs a cost for the HS as they behave differently from the non-HS group only in the critical region "The verb" and prefer the plural-marked verbs more regardless of the existence of the subject.

Another important factor that may lead to difficulties in integrating different types of knowledge is the quality and quantity of input that HS receive. According to Polinsky and Scontras (2020), less time dedicated on a language leads to reduced input, which is considered to be a crucial factor that leads to the observed divergences between heritage and non-heritage speakers. Regarding the quantity of input, they claim that different grammatical phenomena might be sensitive to input quantity. For example, if a phenomenon is rare and not reinforced, HS will never encounter the necessary input to learn the phenomenon successfully. For the input quality, they assert that HS' input is

limited to a small set of speakers and the topics common to the situations in which the HL is used. Mendez et al. (2015) also suggest that any changes in the input quality of HS would result in displaying less sensitivity to appropriate grammatical choices, especially for structures that allow for two or more options which must be inferred from the reduced or suboptimal linguistic input conditions. In addition, researchers have recently agreed on the vital role of both the quality and quantity of input in integrating knowledge from different sources (Chondrogianni and Marinis, 2011; Kupisch et al., 2014; Unsworth et al., 2014; Unsworth, 2016). Therefore, it can be concluded that both the quantity and quality of input play a central role for the observed performance of HS in the present study.

What is more, from a methodological point of view, the online nature of the task provides additional information about the temporal resolution of the processing rather than the metalinguistic knowledge of the participant which is provided by judgment tasks. Online tasks can provide information about the point at which the integration of information from different sources becomes difficult and thus leading to different processing patterns (Sorace, 2011). This is exactly what was observed in the present study. While the HS performed differently from the non-HS group in the critical region "The verb," they behaved similar to the non-HS group in the next region. This shows that HS face difficulties in integrating the syntactic and morphological knowledge in the critical region "The verb" and they need more time to integrate this knowledge compared to the non-HS. Yet, after this critical point, their processing mechanism functions in the same way as the non-HS group as no group differences are observed in the "Spillover" regions. Previous studies with HS on the optional SVA marking in Turkish have used the acceptability judgment task, which is an offline task that mainly measures the metalinguistic knowledge. While two studies have reported an over-acceptance of plural-marked verbs among HS (Bamyacı, 2016; Lago et al., 2019), one study has revealed no difference in the overall acceptance of plural-marked vs. unmarked verbs between HS and non-HS (Uygun and Felser, 2021). The over-use of plural-marked verbs in both offline and online tasks may be attributed to the less robust grammar of HS. According to Putnam (2019), HS develop unstable and unconsolidated grammars as a result of the competition between their (two or) more languages when compared to non-HS.

These processing resource limitations that lead to the restructuring of grammar and the reduced input conditions may explain why the integration process of knowledge from two different sources incurs a cost for the HS group as they display the existence of subject and verb marking interaction only in the "Spillover 1" region but not in the critical region "The verb" while the non-HS group shows this interaction both in "The verb" and "Spillover 1" regions. The HS group experiences problems when they have to integrate the syntactic and morphological knowledge (i.e., internal interfaces) as they are less affected by the existence of subject in the same way as the non-HS group. The non-HS

group was able to contrast the two subject conditions more strongly than the HS in the critical region of ‘The verb’ while the HS group failed to contrast this manipulation. But more real-time processing research in optional SVA marking with HS is needed to assess and compare HS and non-HS groups’ linguistic behavior and performance to be able to reach more generalizable conclusions regarding internal interfaces.

Conclusion

Since the acquisition of the phenomena displaying optionality is known to be difficult as a result of the suboptimal input and acquisition conditions, the optional SVA marking of Turkish has been investigated in HS and non-HS by employing a self-paced reading experiment that measures the reading times of the words. SVA marking is an internal interface involving the combination of syntactic and morphological knowledge and is not expected to be difficult to acquire and vulnerable to attrite. The results indicate that HS behave differently from non-HS even in internal interfaces and the nature of the experiment enables us to see at which point(s) there are quantitative and qualitative differences between the groups and whether HS restructure their grammar to compensate for their processing problems under time pressure. It is very important to test different phenomena in HS *via* online and offline measures with an attempt to understand the factors that make HS and their native language different from the non-HS and to obtain a comprehensive picture of theories about bilingualism and heritage language.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the author, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the University of Potsdam. The patients/participants provided their written informed consent to participate in this study.

References

- Aarssen, J. (1996). *Relating Two Events in Two Languages: Acquisition of Cohesive Devices by Turkish-Dutch Bilingual Children at School Age*. [dissertation]. Tilburg: Tilburg University.
- Albirini, A., Benmamoun, E., and Saadah, E. (2011). Grammatical features of Egyptian and Palestinian Arabic heritage speakers’ oral production. *Stud. Sec. Lang. Acquisit.* 33, 273–303. doi: 10.1017/S0272263110000768
- Altan, A. (2006). “Acquisition of a null subject language” in *Unpublished Paper Presented at the 13th International Conference on Turkish Linguistics, August 2006* (Sweden: Uppsala University)
- Altan, A. (2013). Acquisition of a null subject language. *Dil Dergisi* 161, 5–22. doi: 10.1501/Dilder_0000000194

Author contributions

The author designed the experiment, analysed the data, wrote the manuscript and has approved it for publication.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project Number: 317633480 – SFB 1287 Project B04.

Acknowledgments

I thank Çilem Çiçek for her help with participant recruitment and data collection and Anna Jessen for her help with the data analysis.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2022.988550/full#supplementary-material>

Azar, Z., Backus, A., and Özyürek, A. (2016). “Pragmatic relativity: gender and context affect the use of personal pronouns in discourse differently across languages” in *Proceedings of the 38th Annual Meeting of Cognitive Science Society*. eds. A. Papafragou, D. Grodner, D. Mirman and J. Trueswell (Austin, TX: Cognitive Science Society), 1295–1300.

Azar, Z., Backus, A., and Özyürek, A. (2017). “Highly proficient bilinguals maintain language-specific pragmatic constraints on pronouns: evidence from speech and gesture” in *Proceedings of the 39th Annual Meeting of Cognitive Science Society*. eds. G. Gunzelmann, A. Howes, T. Tenbrink and E. Davelaar (Austin, TX: Cognitive Science Society), 81–86.

Azar, Z., and Özyürek, A. (2015). Discourse management: reference tracking in speech and gesture in Turkish narratives. *Dutch J. Appl. Linguist.* 4, 222–240. doi: 10.1075/dujal.4.2.06aza

- Azar, Z., Özyürek, A., and Backus, A. (2019). General and language specific factors influence reference tracking in speech and gesture in discourse. *Discourse Process.* 56, 553–574. doi: 10.1080/0163853X.2018.1519368
- Azar, Z., Özyürek, A., and Backus, A. (2020). Turkish-Dutch bilinguals maintain language-specific reference tracking strategies in elicited narratives. *Int. J. Bilingual.* 24, 376–409. doi: 10.1177/1367006919838375
- Baayen, R. H., Davidson, D. J., and Bates, D. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Bamyacı, E. (2016). *Competing structures in the bilingual mind: A psycholinguistic investigation of optional verb number agreement*. Berlin: Springer
- Bamyacı, E., Häussler, J., and Kabak, B. (2014). The interaction of animacy and number agreement: an experimental investigation. *Lingua* 148, 254–277. doi: 10.1016/j.lingua.2014.06.005
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. (2013). Random-effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bayram, F., Pisa, G., Rothman, J., and Slabakova, R. (2021). “Current trends and emerging methodologies in charting heritage language grammars” in *The Cambridge Handbook of Heritage Languages and Linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 545–578. doi: 10.1017/9781108766340.025
- Benmamoun, A. (2000). *The feature structure of functional categories*. Oxford: Oxford University Press.
- Benmamoun, A., Montrul, S., and Polinsky, M. (2013). Heritage languages and their speakers: opportunities and challenges for linguistics. *Theor. Linguist.* 39, 129–181. doi: 10.1515/tl-2013-0009
- Bialystok, E. (2009). Bilingualism: the good, the bad, and the indifferent. *Bilingual. Lang. Cogn.* 12, 3–11. doi: 10.1017/S1366728908003477
- Carminati, M. N. (2002). *The Processing of Italian Subject Pronouns*. [dissertation]. Amherst, MA: University of Massachusetts.
- Cherici, A. (2021). Topic-drop vs. pro-drop: Null subjects in Chinese native speakers’ L3 Italian. IULC Work. Pap. 20. Available at: <https://scholarworks.iu.edu/journals/index.php/iulcwp/article/view/32209/36954>
- Chondrogianni, V., and Marinis, T. (2011). Differential effects of internal and external factors on the development of vocabulary, tense morphology and morpho-syntax in successive bilingual children. *Linguist. Approach. Bilingual.* 1, 318–345. doi: 10.1075/lab.1.3.05cho
- Doğruöz, A. S. (2007). *Synchronic Variation and Diachronic Change in Dutch Turkish: A Corpus Based Analysis*. [dissertation]. Tilburg: Tilburg University.
- Doğruöz, A. S., and Backus, A. (2009). Innovative constructions in Dutch Turkish: an assessment of ongoing contact-induced change. *Bilingual. Lang. Cogn.* 12, 41–63. doi: 10.1017/S1366728908003441
- Drummond, A. (2013). Ibox Farm. Available at: <http://spellout.net/ibexfarm/>
- Enç, M. (1986). “Topic switching and pronominal subjects in Turkish” in *Studies in Turkish Linguistics*. eds. D. I. Slobin and K. Zimmer (Amsterdam: Benjamins), 195–208.
- Gondra, A. (2022). Testing the interface hypothesis: heritage speakers’ perception and production of Spanish subject position with unergative and unaccusative verbs. *Int. J. Bil. Educ. Bilingual.* 25, 1730–1764. doi: 10.1080/13670050.2020.1799322
- Haznedar, B. (2010). Transfer at the syntax-pragmatics interface: pronominal subjects in bilingual Turkish. *Sec. Lang. Res.* 26, 355–378. doi: 10.1177/0267658310365780
- Johanson, L. (1998). “The structure of Turkic” in *The Turkic Languages*. eds. L. Johanson and É. Á. Csátó (London: Routledge), 30–66.
- Just, M., Carpenter, P., and Woolley, J. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol.* 111, 228–238. doi: 10.1037//0096-3445.111.2.228
- Kerslake, C. (1987). “Noun phrase deletion and pronominalization in Turkish” in *Studies on Modern Turkish*. eds. H. E. Boeschoten and L. T. Verhoeven (Tilburg: Tilburg University Press), 91–104.
- Koban Koç, D. (2016). Social variables and Turkish subject pronoun use in New York City: the effect of language effect. *Poznan Stud. Contemp. Linguist.* 52, 431–453. doi: 10.1515/pscl-2016-0018
- Kornfilt, J. (1984). *Case Marking, Agreement and Empty Categories in Turkish*. [dissertation]. Cambridge, MA: Harvard University.
- Kornfilt, J. (1997). *Turkish*. London, New York: Routledge
- Kupisch, T., Lein, T., Barton, D., Schröder, D. J., Stangen, I., and Stoehr, A. (2014). Acquisition outcomes across domains in adult simultaneous bilinguals with French as weaker and stronger language. *J. French Lang. Stud.* 24, 347–376. doi: 10.1017/S0959269513000197
- Kuznetsova, A., Bruun Brockhoff, P., and Haubo Bojesen Christensen, R. (2014). lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). R package version 2.0-11. Available at: <https://cran.r-project.org/web/packages/lmerTest/index.html>
- Lago, S., Gracaniin-Yukse, M., Şafak, D. F., Demir, O., Kırkıci, B., and Felsler, C. (2019). Straight from the horse’s mouth: agreement attraction effects with Turkish possessors. *Linguist. Approach. Bilingual.* 9, 398–426. doi: 10.1075/lab.17019.lag
- Marian, V., Spivey, M., and Hirsch, J. (2003). Shared and separate systems in bilingual language processing: converging evidence from eye-tracking and brain imaging. *Brain Lang.* 86, 70–82. doi: 10.1016/s0093-934x(02)00535-7
- Mendez, T. L., Rothman, J., and Slabakova, R. (2015). Discourse-sensitive clitic-doubled dislocations in heritage Spanish. *Lingua* 155, 85–97. doi: 10.1016/j.lingua.2014.01.002
- Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: a case of morphosyntactic convergence. *Bilingual. Lang. Cogn.* 7, 125–142. doi: 10.1017/S1366728904001464
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press
- Montrul, S., and Polinsky, M. (2011). Why not heritage speakers? *Linguist. Approach. Bilingual.* 1, 58–62. doi: 10.1075/lab.1.1.07mon
- Otheguy, R., Zentella, A. C., and Livert, D. (2007). Language and dialect contact in Spanish in New York: towards the formation of a speech community. *Language* 83, 770–802. doi: 10.1353/lan.2008.0019
- Özcan, F. H., Keçik, I., Topbaş, S., and Konrat, A. (2000). “A comparative study in pronominal use in the discourse of monolingual Turkish-speaking and bilingual Turkish-Danish speaking children” in *Det er Conversation 801 değil mi? Perspectives on the Bilingualism of Turkish Speaking Children and Adolescents*. eds. A. Holmen and J. N. Jørgensen (Copenhagen, Denmark: LÆ rerhojskoles), 121–136.
- Özsoy, S. (1987). “Null subject parameter and Turkish” in *Studies on Modern Turkish*. eds. H. E. Boeschoten and L. T. Verhoeven (Tilburg: Tilburg University Press), 82–90.
- Polinsky, M. (1995). Cross-linguistic parallels in language loss. *Southwest J. Ling.* 14, 87–123.
- Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.
- Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Bilingual. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245
- Putnam, M. T. (2019). The (in)stability of grammars. *Stud. Sec. Lang. Acquisit.* 41, 275–278. doi: 10.1017/S0272263119000299
- R Development Core Team. (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria
- Rothman, J., and Slabakova, R. (2011). The mind-context divide: on acquisition at the linguistic interfaces. *Lingua* 121, 568–576. doi: 10.1016/j.lingua.2011.01.003
- Sağın Şimşek, S. Ç. (2010). *Adıl düşürme değiştireninin Türkçe tekdilli ve Türkçe-Almanca ikidilliler tarafından işletimi*. Mersin: Ulusal Dilbilim Kurultayı.
- Schroeder, C. (1999). *The Turkish Nominal Phrase in Spoken Discourse*. Wiesbaden: Harrassowitz Verlag.
- Scontras, G., Polinsky, M., and Fuchs, Z. (2018). In support of representational economy: agreement in heritage Spanish. *Glossa A J. Gen. Ling.* 3, 1–29. doi: 10.5334/gjgl.164
- Sezer, E. (1978). Eylemlerin çoğul öznelerle uyumu. *Genel Dilbilim Dergisi* 1, 25–32.
- Silva-Corvalán, C. (1994). *Language Contact and Change*. Oxford: Clarendon Press.
- Slobin, D. I., and Talay, A. (1986). “Development of pragmatic use of subject pronouns in Turkish child language” in *Proceedings of the Turkish Linguistics Conference*. eds. A. Aksu-Koç and E. E. Taylan (Istanbul: Boğaziçi University Publications), 207–228.
- Sorace, A. (2005). “Selective optionality in language development” in *Syntax and Variation: Reconciling the Biological and the Social*. eds. I. Cornips and K. Corrigan (Amsterdam: Benjamins), 55–80. doi: 10.1075/cilt.265
- Sorace, A. (2011). “Cognitive advantages in bilingualism: is there a bilingual paradox?” in *Multilingualism: Language, Power and Knowledge*. ed. P. Valore (Pisa: Edistudio), 335–358.
- Sorace, A. (2012). Pinning down the concept of interface in bilingual development: A reply to peer commentaries. *Linguist. Approach. Bilingual.* 2, 209–217. doi: 10.1075/lab.2.2.04sor
- Sorace, A., and Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Sec. Lang. Res.* 22, 339–368. doi: 10.1191/0267658306sr2710a
- Sorace, A., and Serratrice, L. (2009). Internal and external interfaces in bilingual language development: beyond structural overlap. *Int. J. Bilingual.* 13, 195–210. doi: 10.1177/1367006909339810

Taylan, E. E. (1986). "Pronominal vs. zero representation of anaphora in Turkish" in *Studies in Turkish Linguistics*. eds. D. I. Slobin and K. Zimmer (Amsterdam: Benjamins), 209–232.

Tsimpli, I., and Sorace, A. (2006). "Differentiating interfaces: L2 performance in syntax-semantics and syntax-discourse phenomena" in *Proceedings of the 30th Annual Boston University Conference on Language Development*. eds. D. Bamman, T. Magnitskaia and C. Zaller (Somerville: Cascadia Press), 653–664.

Tsimpli, I., Sorace, A., Heycock, C., and Filiaci, F. (2004). First language attrition and syntactic subjects: a study of Greek and Italian near-native speakers of English. *Int. J. Bilingual.* 8, 257–277. doi: 10.1177/13670069040080030601

Turan, D. (1995). *Null vs. Overt Subjects in Turkish Discourse: A Centering Analysis*. [dissertation]. Philadelphia, PA: University of Pennsylvania.

Unsworth, S. (2016). Early child L2 acquisition: age or input effects? Neither, or both? *J. Child Lang.* 43, 608–634. doi: 10.1017/S030500091500080X

Unsworth, S., Argyri, F., Cornips, L., Hulk, A., Sorace, A., and Tsimpli, I. (2014). The role of age of onset and input in early child bilingualism in Greek and Dutch. *Appl. Psycholinguist.* 35, 765–805. doi: 10.1017/S0142716412000574

Uygun, S., and Felser, C. (2021). *Constraints on subject-verb agreement marking in Turkish-German bilingual speakers*. *Linguist. Approach. Bilingual*

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics With S*. Springer: New York.

Verhoeven, L. (1990). "Acquisition of discourse cohesion in Turkish" in *Studies on Turkish Linguistics*. ed. S. Koç (Ankara: Middle East Technical University), 437–452.



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY

Yi Liu,
Hong Kong Polytechnic University,
Hong Kong SAR, China
Jessica L. Montag,
University of Illinois at Urbana-Champaign,
United States

*CORRESPONDENCE

Anamaria Bentea
anamaria.bentea@uni-konstanz.de

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 12 August 2022

ACCEPTED 10 October 2022

PUBLISHED 22 November 2022

CITATION

Bentea A and Marinis T (2022) Multiple
wh-interrogatives in child heritage
Romanian: On-line comprehension and
production.
Front. Psychol. 13:1018225.
doi: 10.3389/fpsyg.2022.1018225

COPYRIGHT

© 2022 Bentea and Marinis. This is an
open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that
the original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Multiple wh-interrogatives in child heritage Romanian: On-line comprehension and production

Anamaria Bentea^{1*} and Theodoros Marinis^{1,2}

¹Department of Linguistics, University of Konstanz, Konstanz, Germany, ²School of Psychology and
Clinical Language Sciences, University of Reading, Reading, United Kingdom

This study compared the online comprehension and the production of multiple interrogatives in 18 Romanian-English bilingual children aged 6;0–9;2 ($M_{AGE}=8;0$) living in the UK who have Romanian as heritage language (L1) and English as majority language (L2) and 32 Romanian monolingual children aged 6;11 to 9;8 ($M_{AGE}=8;3$). We examined whether differences emerge between heritage and monolingual children in the online comprehension and in the production of multiple interrogatives in Romanian, which requires fronting of all wh-phrases, contrary to English. The main aim was to uncover to which extent similarities or differences in morphosyntactic properties between the L1 and the L2 systems affect the acquisition and processing of the heritage language/L1. Online comprehension was assessed in a self-paced listening task, while production was assessed using an elicitation task. The results reveal that Romanian heritage children show similar online comprehension patterns to monolingual children for multiple interrogatives in Romanian. A different pattern emerges for production as heritage children produce less complex multiple questions in Romanian and avoid movement of two wh-phrases in all elicited structures. Given that their predominant responses for multiple interrogatives only make use of the structural option present in English, namely one fronted wh-phrase and one *in-situ*, we take this to show that there is transfer from the majority language to the heritage language. Thus, language production in the children's L1 seems to be affected by properties of the dominant L2, under cross-linguistic influence. Taken together, the results for both comprehension and production suggest that heritage children are able to establish the underlying representation of multiple wh-movement structures, similarly to monolinguals, but have difficulties activating the more complex structure in production.

KEYWORDS

heritage language, child bilingualism, multiple interrogatives, Romanian, self-paced listening, elicited production

Introduction

Various studies on heritage language (HL) acquisition have investigated the end-state grammars of adult HL speakers (Montrul, 2016; Polinsky, 2018; Polinsky and Scontras, 2020a,b) and have shown that they are highly heterogeneous in terms of first language (L1) acquisition outcomes and typically diverge from monolinguals in their L1 when tested in offline comprehension and production (Benmamoun et al., 2013; Montrul, 2016; Polinsky and Scontras, 2020a). This variability resembles that often found among second language (L2) learners, although L1 exposure starts from birth (Kupisch and Rothman, 2018). In contrast, few studies have focused on exploring how HL grammatical knowledge is accessed and implemented during on-line language processing (see Bayram et al., 2021; Jegerski and Sekerina, 2021 for a review) and even less is known about online language processing in HL children, children who speak a language that is different from the dominant societal language (Kupisch and Rothman, 2018).

The present study aims to bring further insights into HL development in child heritage speakers by comparing the performance of Romanian heritage children with L2 English to L1 Romanian-speaking children raised monolingually using both on-line comprehension and production tasks. In order to better understand how differences in surface syntactic structure between the heritage and the dominant societal language affect HL development, we examined whether heritage children pattern similarly to monolingual children on the real-time processing and the production of various types of multiple wh-questions, which display different syntactic properties in the heritage language, Romanian, and in the societal language, English: while Romanian fronts both wh-words, English, only fronts one wh-word, the second one remaining *in situ*. By investigating performance under different modalities, we aimed to get a more straightforward glimpse at the nature of the differences between child heritage speakers and child monolingual L1 speakers and how this relates to cross-linguistic influence (Serratrice, 2013; Meir and Janssen, 2021; van Dijk et al., 2021).

The paper is organized as follows. We first review previous studies on cross-linguistic influence in bilingual children with a focus on HL development. Then we present the properties of multiple wh-questions in Romanian and the findings for the acquisition of these structures in monolingual children. We conclude the introductory section with the research questions and predictions of the current study. We proceed with the presentation of participants, methods, and procedure. We then present the results, followed by discussion and conclusion.

Cross-linguistic influence in early bilingual acquisition

The topic of cross-linguistic influence at the level of morphosyntax has been extensively investigated in child

bilingualism (see Serratrice (2013) for an overview and van Dijk et al. (2021) for a recent meta-analysis evaluating cross-linguistic influence across 26 experimental studies). Research has shown that one language can have an effect on the other language at a morphosyntactic level (Hulk and Müller, 2000) and can lead to differences between monolingual and bilingual children which can be either quantitative, qualitative, or both. Quantitative differences stem from the frequency with which a certain structure is accepted or used by bilingual compared to monolingual children (Serratrice et al., 2004; Argyri and Sorace, 2007; Nicoladis and Gavrila, 2015). In other words, a phenomenon also present in monolingual development is reinforced in bilingual development under the influence of one language over the other. Qualitative differences stem from the presence of different language patterns in bilingual children's production and comprehension relative to monolinguals (Nicoladis, 2006, 2012; Strik and Pérez-Leroux, 2011). Recently, Bosch and Unsworth (2020) investigated cross-linguistic influence in the production and acceptability of V2 word orders in English-Dutch bilingual children and found both quantitative and qualitative differences. Bilinguals accepted V2 orders with auxiliary verbs significantly more than monolingual children, but also accepted V2 with main verbs, contrary to monolinguals.

According to Hulk and Müller (2000) and Müller and Hulk (2001), cross-linguistic influence holds when the child's two languages overlap at the surface level. If one language (language A) displays two structural options and the other language (language B) only makes one of these options available, then the option shared by the two languages may be reinforced in language A under influence from language B. In other words, "*there has to be a certain overlap of the two systems at the surface level*" (Hulk and Müller, 2000, p. 228–229). However, there is mixed evidence from the literature showing that cross-linguistic influence does not hold even in the presence of such structural overlap (Argyri and Sorace, 2007) or that cross-linguistic influence occurs in the absence of structural overlap (Nicoladis, 2006, 2012). Importantly, cross-linguistic influence does not seem to occur all the time and one of the factors that has been proposed to influence cross-linguistic influence is language dominance, which refers to the language that the child uses more frequently or the language in which the child has higher proficiency (Yip and Matthews, 2006). Here the prediction is that cross-linguistic influence goes from children's dominant language into their weaker language (van Dijk et al., 2021), although there are also studies which found no relation between cross-linguistic influence and language dominance (Blom, 2010; Serratrice et al., 2012), showing that cross-linguistic influence can occur independently of language dominance.

While the majority of studies on early bilingual acquisition has investigated children's offline comprehension, judgements, and production, only a few have examined real-time sentence processing in bilingual children. These have mainly focused on early L2 learners and compared children's real-time processing of L2 morphosyntactic properties to that of their monolingual peers

(Marinis, 2007; Chondrogianni and Marinis, 2012; Marinis and Saddy, 2013; Chondrogianni et al., 2015a,b) and generally report qualitatively similar processing patterns in bilinguals and monolinguals. Lemmerth and Hopp (2019) and van Dijk et al. (2022) specifically tested the effects of cross-linguistic influence on bilingual children's on-line sentence processing. van Dijk et al. (2022), for example, tested English-Dutch and German-Dutch bilinguals aged 5 to 9 on a self-paced listening task assessing processing of word order in Dutch sentences. They found similar listening patterns in the V2 and V3 condition in Dutch in both monolinguals and bilinguals, but also report effects of cross-linguistic influence in the German-Dutch group in the condition instantiating a structural overlap between the two languages. In other words, the German-Dutch bilinguals slowed down when listening to V2 structures in Dutch and this slowdown was more pronounced in children who were more German dominant.

In contrast to the substantial literature on L2 acquisition, comparatively fewer studies investigated the acquisition of morphosyntax in HL development and how this is affected by cross-linguistic influence from the societal language. Some studies found no effects of cross-linguistic, suggesting that language-external factors shape child HL development (Daskalaki et al., 2019; Rodina et al., 2020). Other studies linked the differences in performance between child heritage speakers and monolinguals to the properties of the societal language (Meir et al., 2017; Meir and Janssen, 2021).

The acquisition of wh-dependencies in the HL has also received little attention. Cuza (2016) used an elicited production task to assess subject-verb inversion in matrix and embedded questions in Spanish heritage children aged 5;0 to 13;3 born and raised in the US. The results showed that Spanish-English bilingual children produce subject-verb inversion in Spanish to a significantly lower rate than their monolingual peers and that they also use subject-verb inversion less in embedded compared to matrix questions. Cuza (2016) argues that this pattern of performance arises from the interplay between cross-linguistic influence from English, the societal language, language dominance and issues of structural complexity. In a similar vein, Strik and Pérez-Leroux (2011) assessed Dutch-French bilinguals aged 5 to 7 and living in France on the production of wh-questions in Dutch, their L1. Although Strik and Pérez-Leroux (2011) do not use the label Dutch heritage speaker for their bilingual group, the children included in their study match the criteria used to define HSs (see Kupisch and Rothman (2018) for a discussion on HL terminology and early child bilingualism). Strik and Pérez-Leroux (2011) found that some of the wh-questions that bilingual children produced in Dutch differed qualitatively from those produced by Dutch monolingual children and followed a French-like structure. These were questions with a fronted wh-phrase and without subject-verb inversion, like **Wat jij doe giraffe?* (lit. What you do giraffe?), and also wh-*in-situ* questions as in **Jij doe wat giraffe?* (lit. You do what giraffe?). According to Strik and Pérez-Leroux, complexity is a trigger for cross-linguistic influence such that structures involving less derivational complexity in one language

(e.g., *in-situ* questions) may impact structures which are derivationally more complex in the other language (e.g., wh-fronting with subject-verb inversion).

These previous works reporting different performance patterns in heritage compared to monolingual children assessed only children's productive skills in their heritage language/L1. Various studies with monolinguals and bilinguals have revealed asymmetries between comprehension and production (Hendriks and Koster, 2010; Grimm et al., 2011) and although there are studies showing that production outpaces comprehension (see Hendriks (2014), Martinez-Nieto and Restrepo (2022) for the acquisition of pronouns), other studies report better performance in comprehension compared to production. Chondrogianni and Marinis (2012), for example, examined the on-line processing and production of tense and non-tense morphemes in L2 English children and children with Developmental Language Disorder (DLD). While the DLD children manifested difficulties with both comprehension and production, the typically-developing L2 children showed on-line sensitivity to the omission of tense morphemes, similarly to the L1 English children, despite variable production rates. Haiden et al. (2009) compared the comprehension and production of wh-questions in French by English-speaking children with L2 French and found high accuracy rates for their comprehension of questions with wh-fronting, on a par with those. In this study we compare HL children's production to their real-time comprehension of multiple wh-questions and use both off-line and on-line methods. This can reveal whether HL children show qualitatively similar processing patterns to monolinguals but also whether asymmetries appear in the comprehension and production of questions with multiple wh-movement.

Multiple wh-interrogatives in (child) Romanian

Full acquisition of multiple wh-questions involves various aspects that are subject to cross-linguistic variation. We will briefly outline the properties of multiple wh-questions that Romanian-speaking children need to acquire, by putting emphasis on differences with English. (1) illustrates multiple *who*-questions and (2) exemplifies *which*-questions in Romanian.

1. a. Cine pe cine acoperă?
who.Nom PE who covers
'Who is covering whom?'
b. *Pe cine cine acoperă?
PE who who covers
*'Whom is who covering?'
2. a. Care fată pe care băiat îl acoperă?
which girl PE which boy him covers
'Which girl is covering which boy?'
b. Pe care băiat care fată îl acoperă?
PE which boy which girl him covers
'Which boy is which girl covering?'

In terms of **lexical properties** of wh-words, wh-objects in Romanian are marked with a differential object marker *pe*, similar to *a* in Spanish. Although in the prescriptive use of English, *who* shows overt case-assignment in the form of *whom*, increasingly native English-speakers use *who* instead in informal spoken contexts (Aarts, 1994). Additionally, *care* ('which')-phrases in Romanian are doubled by a co-indexed clitic pronoun *il* ('him') for masculine and *o* ('her') for feminine.

In terms of **movement properties**, wh-words move overtly, the difference with respect to English being that multiple wh-words in Romanian move together to a clause-initial position, as shown by (1) and (2) above. This is a property that Romanian shares with Bulgarian and other Slavic languages. According to Alboiu (2002), multiple wh-constructions in Romanian are derived by first moving the closest candidate (the subject), defined in terms of c-command, to a Spec,XP position. The remaining phrases then move *via* a 'tucking in' mechanism (see Richards, 1997) below the specifier created by the moved subject and this 'tucking in' movement of the following wh-phrases can take place in any order. On the other hand, fronting a *who*-object over a *who*-subject is ungrammatical (in both Romanian and English), as indicated by the asterisk in example (1b). Movement of wh-words in both languages obeys Superiority (Chomsky, 1973), a condition that limits the ordering of wh-words and blocks one wh-word from moving over another wh-word occupying a hierarchically higher position in the structure. Alboiu (2002) suggests that Superiority is observed in Romanian under her proposed analysis. Given that the subject occupies a structurally higher position and is the closest candidate, it should move first. This requirement does not hold for *which*-questions, as evidenced by the grammaticality of the example in (2b; see Pesetsky (2000) for an explanation). Laenzlinger and Soare (2005) and Soare (2009) convincingly argue for Romanian that *which*-expressions always appear clause-initially, preceding *who*-phrases. By adopting a split-CP analysis (Rizzi, 1997) and a cartographic approach to syntactic structures (Rizzi, 2004), Soare (2009) shows that *which*-phrases target the specifier position of a Topic head above the specifier Focus position which they postulate as the landing site of *who*-phrases.

The **semantic properties** of multiple wh-questions require establishing a pairing relation between the wh-phrases: a felicitous answer for a question like (1a) is "The girl covers the dog and the boy covers the cat." in which the exhaustive sets of *who* and *which* are pairwise linked.

Children's experience with such sentences is extremely reduced. Grebenyova (2005, 2011) showed that there are only five instances of such questions in the English CHILDES database. A search through the two corpora on Romanian in CHILDES (MacWhinney, 2000) yielded no instances of multiple *wh*-questions. The acquisition of multiple wh-questions has received relatively little attention in the literature. Grebenyova (2011) elicited multiple interrogatives from 20 monolingual English-speaking children (aged 3;07–6;02), 20 monolingual Russian-speaking children (aged 3;05–6;05) and 18 Malayalam-speaking children (aged 4;05–5;04). The three languages differ

with respect to the movement properties of wh-words. Russian allows multiple wh-fronting, while English fronts one wh-phrase and Malayalam is a *wh-in-situ* language. Grebenyova's findings demonstrate that English- and Malayalam-acquiring children have adult-like knowledge of the syntax of multiple wh-questions, whereas Russian-speaking children allow fronting of only one of the wh-phrases, following an English-like structure.

To our knowledge, three studies so far investigated the acquisition of multiple wh-questions in Romanian and they all looked at how Romanian-speaking monolingual children ranging in age from 4 to 9 years old comprehend this type of question. Bentea (2010) examined how 4- to 6-year-old English, French and Romanian children (24 in total) interpret multiple *wh*-questions (i.e., whether they assign pair-list readings to multiple interrogatives). Bentea (2010) was also interested in whether children assign an adult-like structure to multiple interrogatives in their language and whether cross-linguistic differences appear between English, French and Romanian children regarding the interpretation and structure of multiple questions. Bentea's (2010) results showed similar performance in the English and French groups, while Romanian-speaking children were more likely to answer only the lower *wh*-element present in the question. In the same vein, Măniță (2017) addressed the question of exhaustivity in the comprehension of Romanian multiple interrogatives. Măniță (2017) tested 42 monolingual Romanian-speaking children (age range 4;0–6;10) and found that the rate with which children give exhaustive answers increases with age, although it does not reach ceiling performance at the age of 6. Furthermore, her results show that children preferentially answered the highest *wh*-word, which was also the subject.

In a recent study on the processing of Romanian multiple *who* and *which*-questions, Bentea and Marinis (2021) show that both monolingual children (6 to 9-year-olds) and adults slow down when processing *who*- compared to *which*-phrases, as measured by reaction times (RTs) in a self-paced listening task. However, only adults seem to show an online sensitivity to the ordering constraints in *who*-questions illustrated in (1b) above. Bentea & Marinis also report higher accuracy scores with multiple *who*- than *which*-questions and show that the latter pose more difficulties for comprehension, particularly in the object-subject order (1,2b), where participants (especially children) show a preference to interpret the first wh-element as agent, along the lines of what has been reported for the processing and comprehension of simple *which*-questions. Bentea & Marinis also found that children even at the age of 6 and 7 answered only one of the wh-phrases, similarly to Bentea (2010) and Măniță (2017), but provided exhaustive lists of referents either for the wh-subject or the wh-object. This suggests that Romanian children have difficulties with pairing the two wh-elements and that this difficulty persists until around the age of 8 when they are able to *exhaust* the question domain and also *pair* the two wh-elements. Therefore, the question that arises is whether bilinguals, who receive less input than monolinguals and are often not tutored in the L1, converge on the correct syntactic structure for multiple wh-questions and attain knowledge of the grammaticality

distinctions among *who* and *which*-multiple questions, especially when such wh-dependencies display a different structure in the L2, the societal language.

To sum it up, multiple wh-interrogatives allow to explore the extent to which bilingual children's language comprehension and production are affected by cross-linguistic influence, as they vary across languages and display language-specific syntactic and semantic properties that children need to acquire although these structures are not frequent in the parental input. In this study, we compare for the first time both the on-line/off-line comprehension and the production of these structures in Romanian heritage children in order to get a clearer picture of the way in which the societal language (here English) influences the acquisition of morphosyntax in the HL.

Research questions and predictions of the current study

The present study investigates the early stages in the acquisition of the HL to examine whether the differences that emerge between HL children and monolinguals hold not only for production, as has been shown by the previous studies examining the acquisition of simple wh-questions in HL children, but also for comprehension. We postulated that the use of a more sensitive and implicit on-line comprehension task, like the self-paced listening task used in this study, might offer a more straightforward glimpse into underlying language representations that are accessed for real-time processing. Together with production tasks, on-line comprehension might help to better understand what differentiates between HL children and monolingual children. The study focused on Romanian as heritage language and addressed the following research questions:

1. Do Romanian HL children and Romanian monolingual children differ when processing questions with multiple wh-fronting in an on-line processing task?

Previous studies with L2 children looking at real-time sentence processing report qualitatively similar processing patterns in bilinguals and monolinguals for tense (Chondrogianni and Marinis, 2012), articles (Chondrogianni et al., 2015a), articles and clitics (Chondrogianni et al., 2015b), word order (van Dijk et al., 2022). Therefore, we expected Romanian HL children to show similar processing patterns to Romanian monolingual children. On the other hand, if there is cross-linguistic influence from English, the societal language, on Romanian HL children's processing of multiple wh-questions, regardless of surface overlap (Nicoladis, 2006, 2012), then heritage children should slow down when they hear the second wh-phrase immediately following the first wh-word.

2. Do Romanian HL children and Romanian monolingual children differ with respect to the production of interrogatives with multiple wh-movement and how does this compare to comprehension?

This is the first study to examine the production of multiple wh-interrogatives in Romanian, as previous studies have only looked into how Romanian-speaking children comprehend this type of questions (Bentea, 2010; Măniță, 2017; Bentea and Marinis, 2021). If the Romanian-speaking children tested in Romania have fully acquired the syntax of multiple interrogatives, they should mainly produce questions with multiple wh-fronting. As far as the heritage group is concerned, we base our predictions on the previous studies on the production of wh-dependencies in child HL (Strik and Pérez-Leroux, 2011; Cuza, 2016) which show qualitative differences between HL and monolingual children in the production of wh-questions. We thus expected Romanian heritage children to be more likely to produce multiple wh-questions with one fronted wh-phrase and one *in-situ*, under cross-linguistic from English, the majority language. Moreover, if asymmetries arise between production and comprehension, then we expect the results to show a similar pattern to that reported for other bilingual populations in which comprehension of multiple interrogatives outpaces their use in production (Haiden et al., 2009; Chondrogianni and Marinis, 2012; Chondrogianni et al., 2015a,b).

Materials and methods

Participants

Eighteen 6- to 9-year-old Romanian heritage children with English as L2 (6 boys; age range = 6;0–9;3; mean age = 96.6 months; $SD = 13.7$ months) living in the United Kingdom (Greater London area and South-East England) and 30 Romanian monolingual children aged six to nine (15 boys; age range = 6;11–9;8; mean age = 99.1; $SD = 11.2$) living in Romania, participated in the study¹. None of the monolingual children had a history of speech and/or language delay or impairment, while one bilingual child had mild expressive language delay diagnosed at the age of three and for which she underwent Speech and Language Therapy until the age of six. As this participant's results at the time of testing did not differ from those of other children, they were included in all subsequent analyses.

Details regarding the bilingual children's language history, including information about their current use of and exposure to both Romanian and English, were collected using a modified version of the Questionnaire for Parents of Bilingual Children (PABIQ; Tuller, 2015). Three parents did not complete the questionnaire. The language background data obtained show that all children were exposed to Romanian from birth, but had a different age of onset (AoO) of English: one child was a simultaneous bilingual, nine children were exposed to English

¹ The monolingual group ($n=30$) overlaps with the sample of children ($n=32$) reported in Bentea and Marinis (2021) which compared the online processing of multiple wh-questions in Romanian monolingual children and adults. In the current study, we investigated in addition the production of multiple interrogatives.

before the age of two (between 5 and 18 months), and five children were exposed to English after the age of three (between 3 and 5 years). The mean age of onset (AoO) of English was 1;10 years ($SD=18$ months, range = 0–5;0 years) and the mean length of exposure (LoE) to English was 6;0 years ($SD=21$ months, range = 2;1–8;0 years). The language background data for the heritage children illustrated in Table 1. Also show that Romanian is used more at home than English, as determined by a paired t -test [$t(14)=3.61$, $p=0.002$], whereas English is the dominant language outside the home [$t(14)=3.62$, $p=0.003$], as well as when it comes to children's current expressive language skills, as reported by the parents [$t(14)=4.45$, $p<0.001$].

Tasks

Self-paced listening task

Children's ability to comprehend Romanian multiple wh-questions in real-time was assessed with an on-line SPL task with picture verification [see Marinis and Saddy (2013)]. In this task, participants reaction times (RTs) are measured every time they press a key on the keyboard in order to listen to sentences presented word-by-word or phrase-by-phrase. The advantage of using such a task is that children administer the auditory stimuli at their own pace and this gives an indication of how fast they process each word/phrase in the sentence.

The self-paced listening task in the current study was presented as a computer game with a puppet (Paddington the Bear). The children were told that they have to listen very carefully to Paddington's questions in order to be able to identify the correct characters in the picture that followed each question. The experimental items contained embedded questions with two fronted wh-phrases in which we crossed two factors: the order of the wh-constituents [either the wh-subject preceded the wh-object (SO) or *vice-versa* (OS)] and the type of wh-phrase (*who* vs. *which*). There were 40 test items in total, with 10 items per condition. The sentences

TABLE 1 Language background information for the heritage Romanian group.

		Romanian	English
Amount of	$M (SD)$	0.82 (0.20)	0.45 (0.29)
exposure before	$MIN-MAX$	0.50–1.00	0–1.00
4yo			
Parental ratings of	$M (SD)$	0.51 (0.23)	0.85 (0.18)
current skills	$MIN-MAX$	0.20–1.00	0.40–1.00
Language use at	$M (SD)$	0.64 (0.23)	0.40 (0.18)
home	$MIN-MAX$	0.17–1.00	0.20–1.00
Current exposure	$M (SD)$	0.44 (0.18)	0.70 (0.19)
outside the home	$MIN-MAX$	0.14–0.71	0.36–1.00

included stories about superheroes (Superman, Batman, and Spiderman) and princesses (Anna, Elsa, and Jasmine) engaging in imaginary activities with different people or animals. At the beginning of every trial, children listened to a lead-in sentence, introducing the characters. This was then followed by the test sentence (i.e., an embedded question) segmented into 8 parts, each part containing one to three words. No images appeared on the screen while children listened to the lead-in and the test sentence. The slashes in the examples signal the end of each segment when children had to press the Space bar to hear the next segment. After the final segment of each test sentence, a picture with three pairs of characters appeared on the screen. All the visual stimuli used in the study can be provided upon request. The position of the pairs and the direction of the actions varied between pictures. Children had to verbally identify all the pairs of characters performing the action described in the sentence. An answer was coded as "correct" when all the relevant pairs were identified. After hearing the sentence in (3), for example, the expected answer was *The fireman is splashing Superman and the elephant is splashing Batman*. This is because Romanian only allows exhaustive pair-list interpretations for multiple wh-questions like the ones exemplified in (3) to (6) below.

Lead-in: This is an image of Spiderman, Superman, Batman, a boy, a fireman and an elephant.

Test sentence: Paddington/ wants to know/

3. SO *who* **cine/ pe cine/** stropește/ în joacă/ seara/ la circ/
 who PE who splashes playfully in the evening at the circus
 "who is splashing whom playfully at the circus in the evening."

4. OS *who* ***pe cine/ cine/** stropește/ în joacă/ seara/ la circ/
 PE who who splashes playfully in the evening at the circus
 "whom is who splashing playfully at the circus in the evening."

Lead-in This is an image of Anna, Elsa, Jasmine, and three monkeys.

Test sentence: Paddington/ wants to know/

5. SO *which* **care prințesă/ pe care maimuță/** o/ acoperă/ dimineața/ la zoo/
 which princess PE which monkey her covers in the morning at the zoo
 "which princess is covering which monkey at the zoo in the morning."

6. OS *which* **pe care prințesă/ care maimuță/** o/ acoperă/ dimineața/ la zoo/
 PE which princess which monkey her covers in the morning at the zoo
 "which princess is which monkey covering at the zoo in the morning."

Apart from the 40 test items, 10 fillers were also included as distractors. Half of the fillers were simple subject *who*-questions and the other half were simple subject *which*-questions. They included both transitive actions (*eat, hold, read, cut, smell*) in which the agent was always animate and the patient inanimate, as well as intransitive actions (*fly, jump, run, sleep*).

Elicited production task

For the elicited production task, children played a guessing game with Paddington the Bear in which they were

7.	SO <i>who</i>	Cine who	pe PE	cine who	a has	mângâiat? patted	
		“Who patted whom?”					
8.	SO <i>which</i>	Care which	fată girl	pe PE	care which	pisică cat	a has mângâiat-o? patted-her
		“Which girl patted which cat?”					
9.	OS <i>which</i>	Pe PE	care which	pisică cat	care which	fată girl	a has mângâiat-o? patted-her
		“Which cat did which girl pat?”					
10.	OS <i>which-who</i>	Pe PE	care which	pisică cat	cine who	a has	mângâiat-o? patted-her
		“Who patted which cat?”					

Children interacted with a Paddington the Bear puppet for the whole duration of the task (see details in the Procedure section below).

The structure of each scenario was as follows. First, both the child and Paddington saw an image of Lego figures and heard *Here are two girls, a boy, two cats and a monkey*, which helped them to familiarise themselves with the characters in the image. The experimenter then covered Paddington's eyes and ears so that the puppet could not see or hear anymore what was happening next in the image. At this point, the child saw another image and heard *Look! This girl is patting the black cat and this girl is patting the white cat. The boy is taking a picture of the monkey*. Then the image presented at the beginning of the trial appeared again on the screen. Paddington's eyes and ears were uncovered and afterwards the experimenter would tell the puppet *Paddington, we can tell you that the boy did not pat anyone, but each girl patted a different cat*. The experimenter then prompted the child to ask Paddington a question about this (e.g., *Which girl patted which cat?*) and then Paddington made his guess. In order to make sure that children produced questions with two wh-phrases, they were told that they need to ask questions about two things at the same time. The task also included six filler items. These prompted children to produce simple wh-questions (three argument questions with a mismatch in animacy like *What did the father wash?* and three adjunct questions such as *Where did the queen sit?*). Children's responses were coded as “felicitous” when they produced a target question with multiple wh-movement like in examples (7) to (10) above.

prompted to produce 24 multiple wh-questions with a subject-object and an object-subject order. Six items were designed to elicit multiple *who*-questions with a SO order (7), 12 items elicited multiple *which*-questions (six with a SO order as in (8) and six with an OS order as in (9)), while six more items were designed to elicit multiple wh-questions with an OS order (10), in which the object was a *which*-phrase and the subject a *who*-phrase, a grammatical option in Romanian.

Other responses were coded as “infelicitous,” alongside the type of error produced.

Procedure

The study was approved by the Ethics Committee of the University of Reading. Informed parental consent was obtained for each child prior to the testing sessions, as well as the oral consent of the child. Each participant was tested individually in a quiet room either at their school or in their home. The experimenter gave oral instructions for both tasks. These were administered at least 1 week apart and in different orders, such that half of the children first saw the SPL task, followed by the elicited production task, while for the other half we first assessed production and then comprehension. Each testing session lasted around 30–40 min, but children could take breaks whenever they felt tired. They received stickers and certificates of participation after each task and at the end of the study they received as well a voucher which they could use to buy books at local bookstores.

The SPL task was programmed and administered using PsychoPy (Peirce et al., 2019). The task started with an introduction in which an image of Paddington first appeared on the screen, telling children what they need to do in the task. Children were also familiarized with the images and names of each of the three superheroes and princesses, although almost all of them already knew these characters from cartoon movies. Before starting the test phase,

the children were presented with four practice items, two of which contained simple *what*-questions (e.g., *What* is Batman reading?) and two multiple *wh*-questions with an animate subject and an inanimate object (e.g., *Who* is drinking *what*?). Each participant then listened to a total of 50 sentences during the test phase and these were randomized in PsychoPy. The task instructions and all the sentences were pre-recorded by a native speaker of Romanian.

The elicited production task was administered as a PowerPoint presentation. The task started by introducing each child to the Paddington puppet. The experimenter explained that Paddington wanted to become a magician and for this he had to improve his guessing skills. The child's task was to help Paddington by playing a game with him in which the child asked Paddington questions about various images and the puppet had to guess the correct answer. For each correct guess, the child gave Paddington a smiley face sticker and if Paddington had at least 20 correct guesses by the end of the task, the child gave Paddington his Magician Diploma. Given the complexity of the target questions, these were presented in four blocks and interspersed with fillers. Each block contained 6 items of the same type (e.g., *SO who* questions) and the order of the blocks was randomized across four lists, each containing 30 items. The task began with four practice trials which elicited multiple *wh*-questions with a mismatch in animacy (e.g., *Who ate what?* or *Which boy hid where?*). The audio presentation of the items was also pre-recorded by a native speaker of Romanian. The questions that the children produced were recorded on answer sheets and then coded for analysis.

Results

Comprehension of multiple *wh*-dependencies

Figure 1 presents the descriptive results for **comprehension accuracy** of Romanian multiple *wh*-questions with two *who* and two *which*-phrases and with distinct subject-object orders. The results indicate that both the Romanian-speaking monolinguals and the heritage Romanian children comprehend multiple *who*-questions well (above 0.80), but show lower accuracy for multiple *which*-questions.

The two groups also show similarities in the errors they produce, which are (i) over-exhaustive answers, (ii) list answers, and (iii) reversed role answers. For over-exhaustive answers children name all the pairs in the image, even when one pair performs a different action. For example, children would answer a question like in (3) or (4) above with *The fireman is splashing Superman, the elephant is splashing Batman, and the boy is pulling Spiderman* although the question only refers to the action of splashing. List answers are cases in which children only answer one of the two *wh*-words by exhaustively listing all the characters involved in the action. That is, they would answer question (5) with *Jasmine and Elsa*. Role reversals are answers in which the Agent-Patient roles are reversed. For instance, children would

answer the question in (6) with *Anna is covering the black monkey* and thus assign the wrong thematic role to *which princess*. There were also very few instances in which children simply identified the wrong action, for example the pulling action when the question was about splashing. Table 2 summarizes the number and type of errors for monolingual and heritage children and indicates that the most common errors children make with multiple *who*-questions are providing list answers to only one of the *wh*-words, while role reversal errors are the most frequent with multiple *which*-questions. This suggests that children find it harder to assign the correct thematic roles in a patient-before-agent structure when the question contains two *which*-phrases.

Given the binary nature of the data (Correct/Incorrect), we analyzed the accuracy results using a binomial generalized linear mixed model with group (monolingual vs. heritage), *wh*-type (*who* vs. *which*), *wh*-order (*SO* vs. *OS*), and their interaction as fixed factors. The fixed factors were coded using repeated contrast coding which tests the difference between the mean of the dependent variable for one level of the categorical variable and the mean of the dependent variable for the adjacent level (Schad et al., 2020). The random-effects structure included intercepts for both participants and items, as well as random slopes for *wh*-type by participant. Alternative models with a more complex random-effect structure either failed to converge or were not retained as their goodness-of-fit resulted in increased Akaike information criterion (AIC)-value. The analysis was conducted using the *lme4* package (Bates et al., 2015) in R (R Core Team, 2022). Planned comparisons, if justified, were done using the *emmeans* package (Lenth, 2022).

The three-way interaction *group*wh-type*wh-order* did not significantly improve the fit of the model as indicated by model

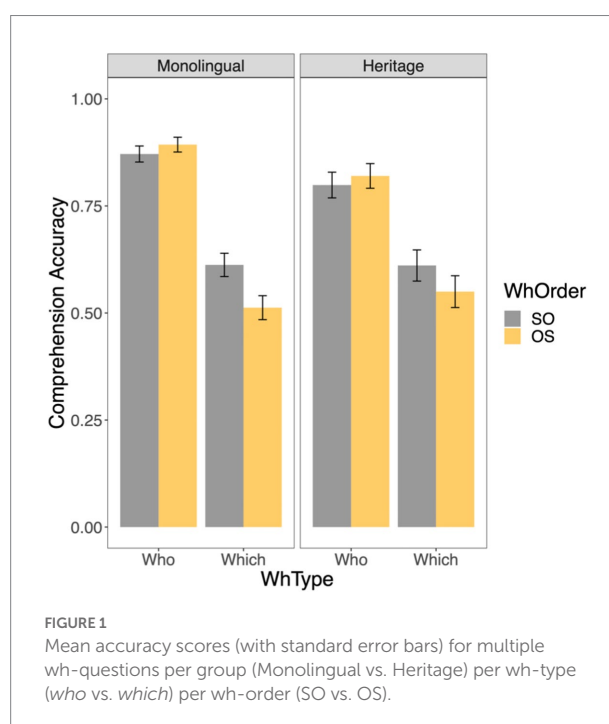


TABLE 2 Type and number of errors by group for comprehension accuracy of multiple wh-questions (total number of errors per condition in parantheses).

Group	Condition	Error types			
		Over-exhaustive	List answers	Reversed role	Wrong action
Monolingual	Who	16 (76)	56 (76)	0 (76)	4 (76)
	Which	32 (280)	63 (280)	185 (280)	0 (280)
Heritage	Who	6 (69)	59 (69)	0 (69)	4 (69)
	Which	25 (151)	53 (151)	75 (151)	0 (151)

TABLE 3 Model summary for comprehension accuracy of multiple wh-questions.

	Estimate	SE	Z	Sig.
(Intercept)	1.78317	0.32066	5.561	<0.001
group(monolingual vs. heritage)	-0.08496	0.60404	-0.141	0.888
wh-type(who vs. which)	2.89120	0.39958	7.236	<0.001
wh-order(SO vs. OS)	0.00179	0.15963	0.011	0.991
group _{monolingual vs. heritage} *	0.06594	0.68398	0.096	0.923
wh-type _{who vs. which}				
group _{monolingual vs. heritage} *	0.12909	0.26980	0.478	0.632
wh-order _{SO vs. OS}				
wh-type _{who vs. which} *	0.80690	0.30057	-2.685	0.007
wh-order _{SO vs. OS}				
Observations	1995 0.220/0.663			
Marginal R ² /				
Conditional R ²				

The values in bold indicate statistical significance.

comparison using the anova function ($p = 0.988$). The summary of the final model is given in Table 2. The results showed a significant effect of wh-type (*who*-questions were comprehended better than *which*-questions) and a significant wh-type*wh-order interaction. No differences in performance emerged between the monolingual and the heritage group (Table 3).

We followed-up on the significant interaction between wh-type and wh-order with a pair-wise comparison. This showed that response accuracy for questions with two *who*-phrases did not differ significantly for the SO and OS orders ($\beta = -0.402$, $SE = 0.271$, $z = -1.480$, $p = 0.138$), while there was a significant difference between the two wh-orders in the case of questions with two *which*-words, with SO *which* questions yielding significantly better comprehension scores than OS *which* questions ($\beta = 0.405$, $SE = 0.150$, $z = 2.703$, $p < 0.01$).

The segment-by-segment residual reaction times (RTs) for the online processing of multiple wh-questions are illustrated in Figure 2 (for the monolingual group) and Figure 3 (for the heritage group). We plot raw residual RTs for readability, but the analyses were done on log-transformed RTs. Only items with correct responses to the comprehension questions were included

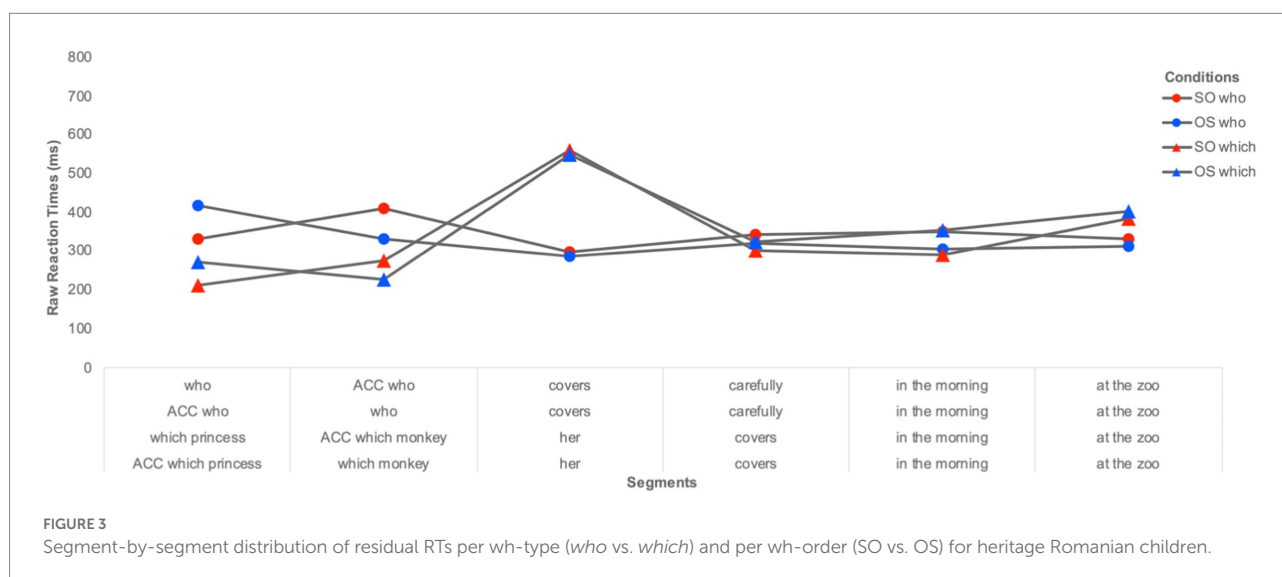
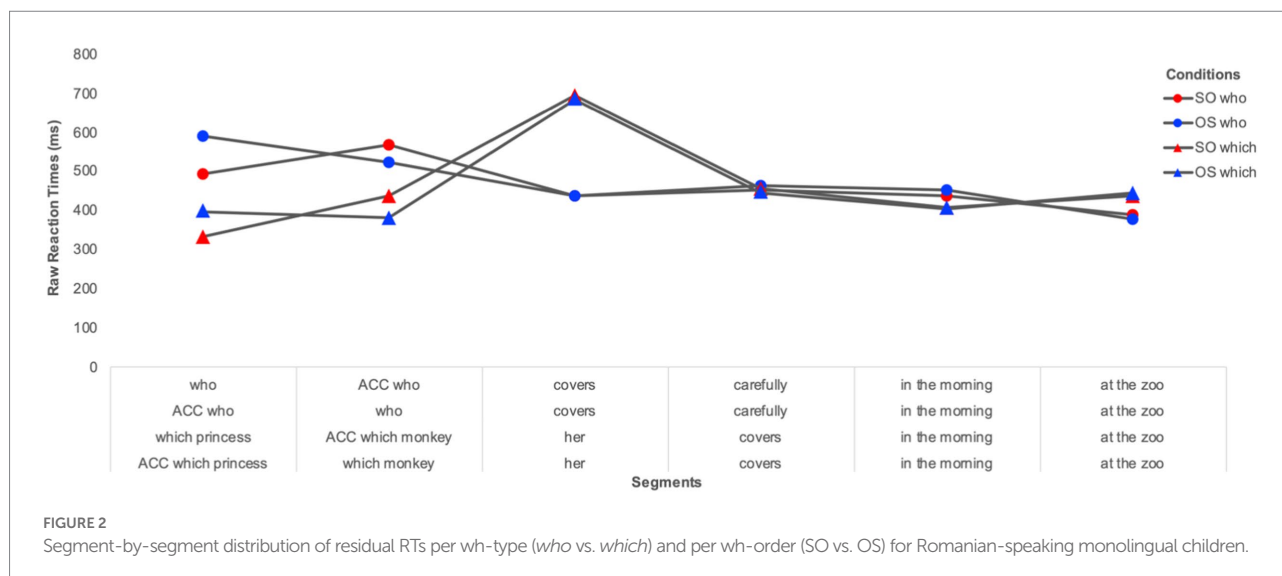
in the RT analyses. Residual RTs were calculated to control for the difference in length between segments. Extreme values were calculated based on boxplots and were excluded from the final analyses. These were residual RTs below -600 ms and above 2000 ms. RTs of 2 *SD* above or below the mean per condition per participant and per item were considered as outliers and therefore replaced with the mean per condition per participant and per item. The total proportion of extreme values and outliers was 3.2% of all data points.

We report the results from six segments, starting with the segment containing the first wh-word and including the segments after the verb, as these can reveal spill-over effects. The log-transformed RTs for each segment were analysed using linear mixed-effects models fit with the maximal random effects structure that converged. This included by-participant and by-item random intercepts and slopes for wh-type by-participant. Group (Monolingual vs. Heritage), wh-type (*who* vs. *which*), wh-order (SO vs. OS), as well as their interaction, were included in the fixed effects structure for each model. All fixed effects were coded using repeated contrast coding. Values of p were calculated by way of Satterthwaite's approximation to degrees of freedom with the *lmerTest* package (Kuznetsova et al., 2017).

Segment 1 (the first wh-word) Results attested to a significant effect of group ($\beta = 0.178$, $SE = 0.057$, $t = 3.121$, $p = 0.003$), indicating that the monolingual children had overall longer RTs than the heritage children. There was also a significant effect of wh-type ($\beta = 0.237$, $SE = 0.042$, $t = 5.522$, $p < 0.003$), with longer RTs for *who*-phrases ($M = 482$ ms) than *which*-phrases (317 ms), as well as a significant effect of wh-order ($\beta = -0.178$, $SE = 0.057$, $t = 3.121$, $p = 0.003$), with shorter RTs for subject ($M = 371$ ms) compared to object wh-words (458 ms). None of the interactions was significant.

Segment 2 (the second wh-word) Results revealed a significant effect of group ($\beta = 0.143$, $SE = 0.055$, $t = 2.564$, $p = 0.013$), with the monolingual children displaying longer RTs than the heritage children, and a significant effect of wh-type ($\beta = 0.167$, $SE = 0.034$, $t = 4.845$, $p < 0.001$), with longer RTs for *who*-phrases ($M = 484$ ms) than *which*-phrases (352 ms). No other effect or interaction were significant.

Segment 3 (the verb in multiple who-questions and the clitic in multiple which-questions) There was a significant effect of group ($\beta = 0.190$, $SE = 0.084$, $t = 2.266$, $p = 0.028$), with longer RTs for the



monolingual group compared to the bilingual group, and a significant effect of wh-type ($\beta = -0.436$, $SE = 0.046$, $t = -9.291$, $p < 0.001$) showing that the verb segment in the multiple *who*-conditions yielded shorter RTs ($M = 386$ ms) than the clitic segment in multiple *which*-conditions ($M = 638$). There was also a significant interaction Group*WhType ($\beta = 0.151$, $SE = 0.051$, $t = 2.910$, $p = 0.003$). As a follow-up on the significant interaction, pair-wise comparisons revealed that the heritage group showed significantly shorter RTs than the monolingual group for the *who*-conditions, ($\beta = -0.266$, $SE = 0.086$, $t = -3.065$, $p = 0.003$), while no significant differences surfaced between the two groups in the *which*-conditions.

Segment 4 (the adverb in multiple *who*-questions and the verb in multiple *which*-questions) No effect was significant.

Segment 5 (in the morning) Results attested to a significant effect of group ($\beta = 0.149$, $SE = 0.072$, $t = 2.057$, $p = 0.045$), as the monolingual children had longer RTs than the heritage children. No other effect was significant.

Segment 6 (at the zoo) No effect was significant.

To summarize, the results of the comprehension task show similar response accuracy and a similar pattern during online comprehension of multiple wh-questions in Romanian-speaking monolingual and heritage Romanian children.

Production of multiple wh-dependencies

The results of the elicitation task showed that children do not only produce questions with multiple wh-fronting, the expected target structure based on the syntax of Romanian multiple wh-questions, but that they often produce other structures as well. To reflect this variability in the children's answers, four scoring categories were used, each corresponding to the four main question types that children produced and which we classified as follows:

- a. MWH_MULTIPLEMOVE: when children produced a question with two wh-words, either *who* or *which*, and with both wh-words fronted, as in (11).
 11. **Cine pe cine** a mângâiat ?
 who PE whom has patted
 “Who patted whom?”
- b. MWH_SINGLEMOVE: when children produced a question containing two wh-words in which only one wh-phrase is fronted and the other one appears in-situ, like in (12).
 12. **Cine** a mângâiat **pe cine** ?
 who has patted PE whom
 “Who patted whom?”
- c. SIMPLE_WH: when participants produced a grammatically correct question but only with one fronted wh-word (13).
 13. **Care fată** a mângâiat pisica?
 which girl has patted cat.the_ESG
 “Which girl patted the cat?”
- d. COORDINATED_WH: when children produced a question with two coordinated wh-words (14).
 14. **Cine și pe cine** a mângâiat ?
 who and PE who has patted?
 “Who and whom patted?”

The distribution of responses differs between the Romanian monolingual and Romanian heritage children, as can be seen in Figure 4. Monolingual children produce three types of questions at similar rates: MWH_MULTIPLEMOVE (0.34), MWH_SINGLEMOVE (0.28) and SIMPLE_WH (0.31). They also produce COORDINATED_WH to a lesser extent (0.07). The most frequent type question that the heritage children produce is MWH_SINGLEMOVE (0.68), followed by SIMPLE_WH (0.24). There are also a few instances of COORDINATED_WH questions (0.04), as well as instances of MWH_MULTIPLEMOVE questions (0.04). However, a closer look at the data reveals that the majority of multiple wh-questions with two fronted wh-phrases are produced by one child and that there are six other children who only produce one question with multiple wh-movement throughout the whole task.

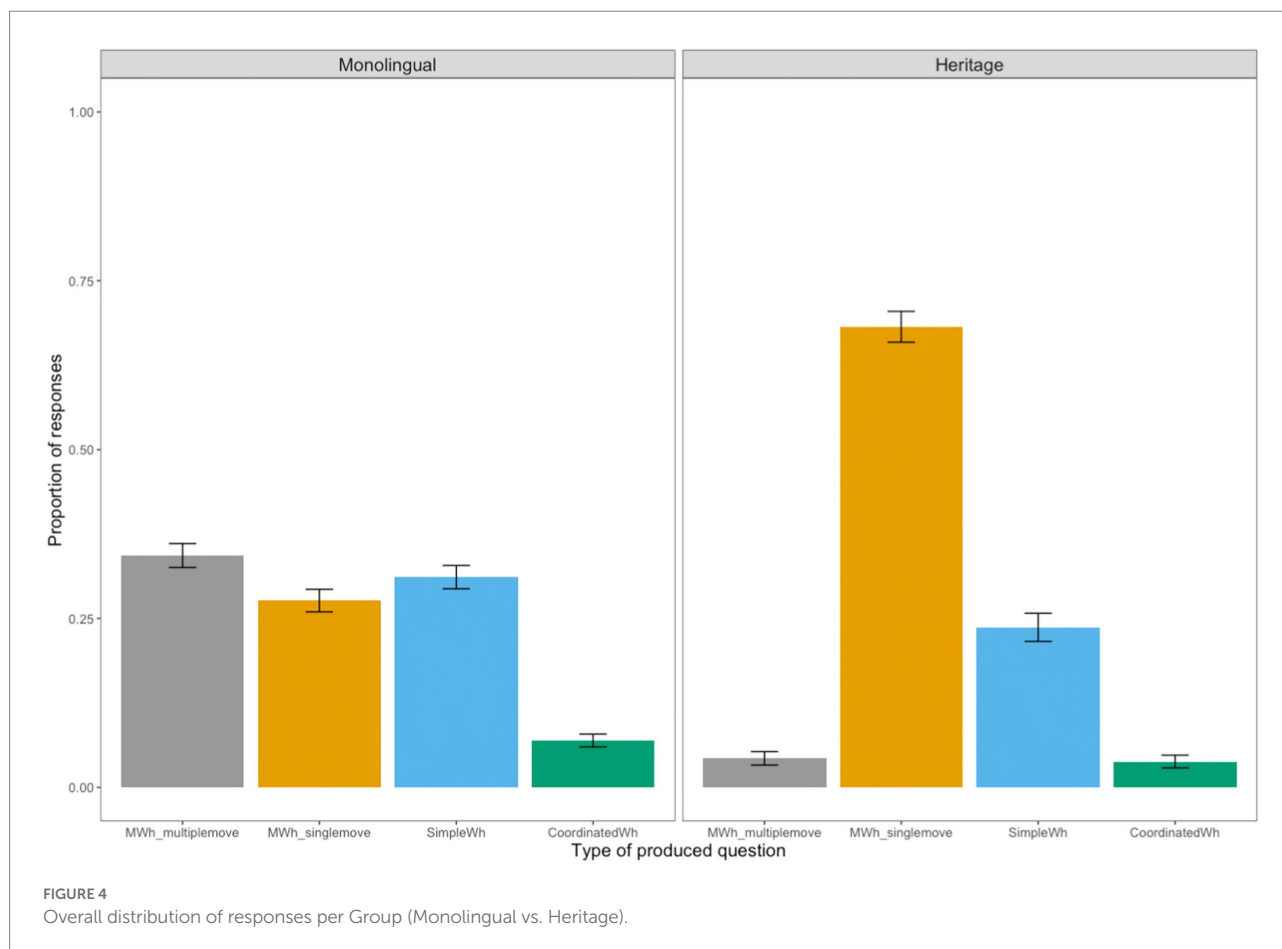
A finer characterization of the results (Figure 5) gives an indication of the response distribution within the two groups of participants for each type of elicited question (see examples 7 to 10 above). As the children mainly produced questions with a subject-object order, we collapsed the results of SO *which* and OS *which* questions and thus report the data for three types of multiple wh-questions with two *who*-phrases (who), with two *which*-phrases (which), and with one *which* and one *who* phrase (which-who). Figure 5 shows that MWH_SINGLEMOVE questions (with one wh-phrase fronted and one *in-situ*) represent the preferred produced structure for the heritage group across all types of elicited questions, irrespective of whether these contained only *who*, only *which*, or both *which* and *who* phrases. The monolingual group

produce more MWH_SINGLEMOVE structures of the type illustrated in (12) with questions containing *which*-phrases, whereas they produce more MWH_MULTIPLEMOVE structures like in (10) in the presence of two *who*-elements.

In order to uncover whether differences emerge between monolingual and heritage children in their productions and whether these differences are modulated by the type of elicited question (*who* vs. *which* vs. *which-who*), we fitted a generalized liner mixed model for each of the four scoring categories outlined above, namely MWH_MULTIPLEMOVE, MWH_SINGLEMOVE, SIMPLE_WH, COORDINATED_WH. The dependent variable in each model was response accuracy, that is, the correct production of questions within each scoring category. We analyzed the production data using the *lme4* package (Bates et al., 2015) in the R environment (R Core Team, 2022), specifying the optimizer ‘bobyqa’ for our models. Each model included the same fixed factors (Group: Monolingual vs. Heritage and QuestionType: *who* vs. *which* vs. *which-who*) as well as their interaction. We used a repeated contrast specification for the fixed factors. The random effect structure included by-participant random slopes. We only report the significant effects and interactions.

MWH_MULTIPLEMOVE

The statistical analysis revealed a main effect of Group ($\beta = -4.659$, $SE = 1.102$, $z = -4.225$, $p < 0.001$). Overall, the heritage children produced significantly fewer questions with multiple wh-fronting compared to the monolingual children. The analysis also revealed a significant difference between *which* and *who* questions ($\beta = -0.957$, $SE = 0.362$, $z = -2.643$,



$p=0.008$). Multiple wh-fronting was most frequent when children had to produce questions with two *who*-phrases compared to two *which*-phrases. The interaction $\text{Group}_{\text{Heritage vs Monolingual}} * \text{QuestionType}_{\text{which-who vs which}}$ was also significant ($\beta = -2.904$, $\text{SE} = 1.220$, $z = -2.381$, $p = 0.017$). As a follow-up on the significant interaction, pair-wise comparisons with an adjusted alpha level using the Tukey method showed that the monolingual group produced significantly fewer questions with multiple wh-movement when the target items only contained *which*-phrases compared to when they contained one *who*-phrase and one *which*-phrase (*which* – *which-who*: $\beta = -0.726$, $\text{SE} = 0.299$, $z = -2.432$, $p = 0.039$).

MWH_SINGLEMOVE

The statistical analysis revealed a main effect of Group ($\beta = 3.667$, $\text{SE} = 1.123$, $z = 3.265$, $p = 0.001$). This means that, overall, heritage children produced significantly more multiple wh-questions with one element fronted and one *in-situ* than the monolingual children. There was also a significant difference between *which* and *who* questions ($\beta = 1.849$, $\text{SE} = 0.305$, $z = 6.054$, $p < 0.001$), indicating that there were overall more multiple wh-questions with single wh-fronting for *which*-questions relative to *who*-questions.

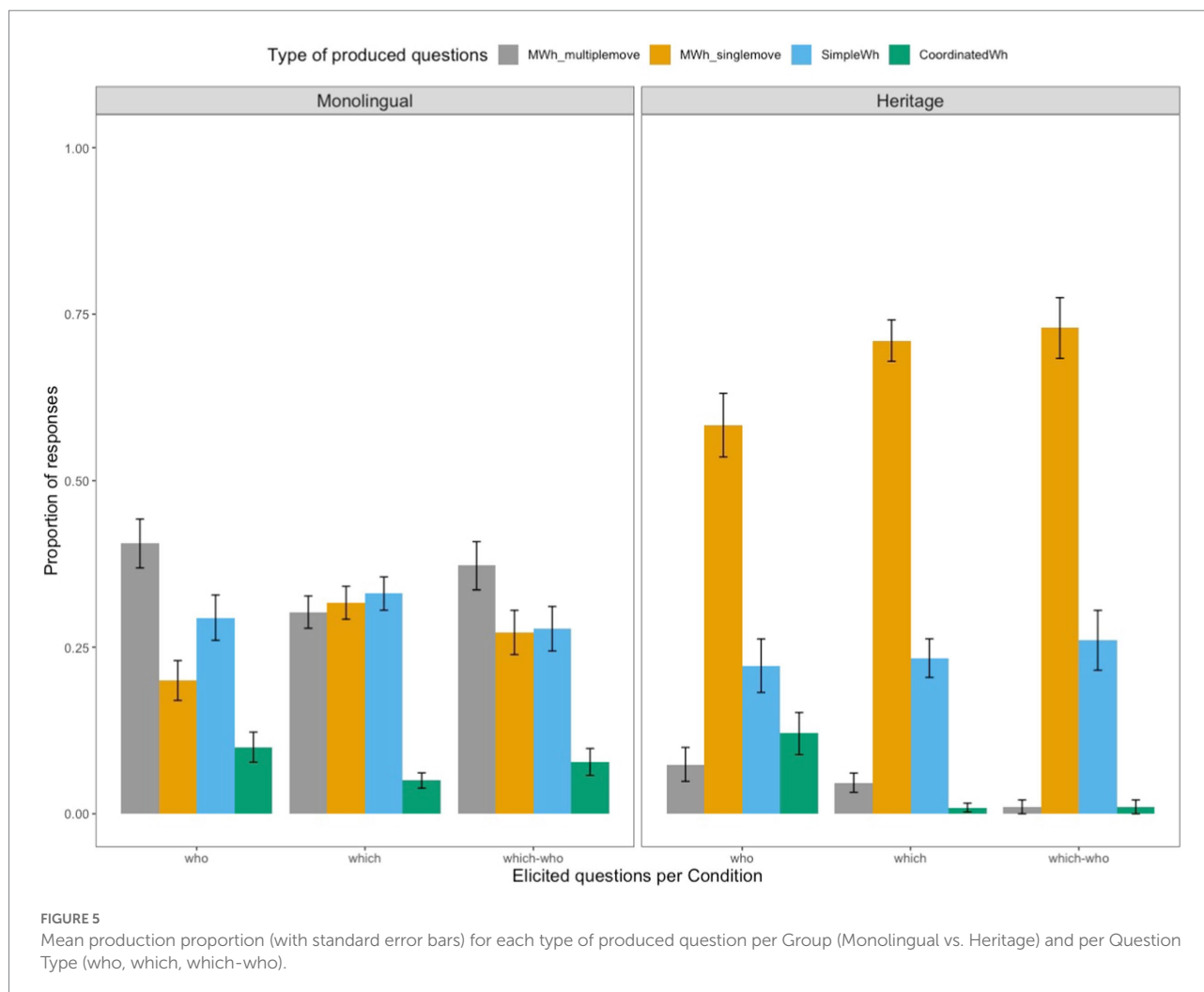
SIMPLE_WH

There was no difference between the two groups on this measure.

COORDINATED_WH

The statistical analysis revealed a significant difference between *which* and *who* questions ($\beta = -2.994$, $\text{SE} = 0.580$, $z = -5.155$, $p < 0.001$). This means that children produced less coordinated wh-questions when both wh-words elements were *which*-phrases. There was also a significant $\text{Group}_{\text{Heritage vs Monolingual}} * \text{QuestionType}_{\text{which vs who}}$ interaction ($\beta = -2.479$, $\text{SE} = 1.157$, $z = -2.141$, $p = 0.032$). Subsequent pairwise comparisons show that the heritage group produces significantly fewer coordinated wh-questions when the target items contained two *which*-phrases compared to when these contained two *who*-phrases ($\beta = -4.234$, $\text{SE} = 1.011$, $z = -3.091$, $p < 0.001$) and this difference is more pronounced in heritage children than in monolinguals.

Thus, heritage children display a different pattern than monolinguals in production, as they produce significantly fewer questions with multiple wh-fronting and significantly more questions with one fronted wh-phrase and one *in-situ* compared to the monolingual children.



Discussion

The main questions guiding the present study were whether differences emerge between Romanian-English bilingual children, for whom Romanian is the heritage language, and Romanian monolingual children in the on-line comprehension and production of multiple wh-interrogatives and whether these differences are due to cross-linguistic influence from the dominant English. To address these questions, we used an on-line comprehension task and an elicitation task targeting multiple *who*- and *which*-questions. The on-line task investigated how Romanian heritage children and Romanian monolingual children process wh-dependencies with two fronted wh-words, while they were listening to sentences on-line for comprehension. This task also aimed to find out whether Romanian-speaking children were sensitive to the asymmetry in object-over-subject movement between multiple *who*- and *which*-questions in real-time comprehension. The combination of production and on-line comprehension tasks can help to get a better picture of heritage language development and better understand the relationship between performance in production and real-time processing.

The results of the comprehension task reveal similar accuracy for the interpretation of multiple wh-questions in Romanian heritage and monolingual children. Both groups comprehend questions with a fronted *who*-subject and *who*-object well, irrespective of the order in which the two wh-phrases occur. In contrast, they have more difficulties with the comprehension of multiple *which*-question, particularly when the wh-object precedes the wh-subject. This is in line with cross-linguistic findings reported for the comprehension of simple *which*-questions showing that children find object *which*-questions harder to comprehend than subject *which*-questions and that the type of wh-element also affects comprehension of wh-dependencies, with object *who*-questions being acquired earlier than object *which*-questions (Friedmann et al., 2009; Bentea and Durrleman, 2013; Contemori et al., 2018).

The self-paced listening data also show a similar pattern during on-line comprehension of multiple interrogatives in Romanian heritage and monolingual children, with shorter RTs when processing *which* vs. *who*-phrases. This finding, coupled with the offline response results, reflects a speed-accuracy trade-off: children are more accurate with *who*- than

which- multiple interrogatives, but they slow down when they process *who*- as compared to *which*-phrases. This difference in processing between *wh*-elements has also been attested for English by Hofmeister et al. (2013) who tested English-speaking adults in a self-paced reading task and reported more efficient processing in English multiple *wh*-questions for *which*-constituents compared to *who*-phrases. Moreover, Romanian heritage-children do not slowdown upon listening to the second *wh*-phrase, as we predicted would be the case if their processing of multiple *wh*-dependencies in Romanian would be affected by cross-linguistic influence from the dominant English. Furthermore, neither group showed an on-line sensitivity to the ungrammatical object-subject order in multiple *who*-questions, contrary to the findings for Romanian monolingual adults in Bentea and Marinis (2021). The authors show that, adults, but not children, are sensitive to the ungrammaticality of multiple *who*-questions in which the *wh*-object precedes the *wh*-subject and one explanation they put forth is that this effect is delayed in children, in other words that it might only surface after the end of the sentence. Other visual world studies have also found young children to be slow in processing *wh*-dependencies (Contemori et al., 2018), with effects occurring after the end of the sentence (Adani and Fritzsche, 2015). Structures like multiple *wh*-interrogatives are more complex than simple *wh*-questions, as they require encoding, integrating, and retrieving two *wh*-elements in the structure. This added complexity may delay processing in children even more.

Taken together, the results of the on-line comprehension task corroborate previous findings that looked at real-time sentence processing in L2 children and found qualitatively similar processing patterns in bilinguals and monolinguals (Chondrogianni and Marinis, 2012; Chondrogianni et al., 2015a,b). Our findings also show that, at the quantitative level, Romanian heritage children process sentences at a faster rate than monolingual children (see van Dijk et al. (2022) for similar results with Dutch bilingual and monolingual children). The faster processing behavior observed in the bilingual group in the self-paced listening task could potentially suggest a general effect of bilingualism on sentence processing in children, which could result in more efficient sentence processing. However, more research is needed to explore this observation further. The key finding remains that heritage children do not differ qualitatively from monolingual children and display a similar on-line comprehension pattern to monolingual children for multiple questions in Romanian. When they encounter multiple fronted *who*- and *which*-questions they are able to parse them incrementally in the same way as monolingual children of the same age. They process *which* *wh*-words faster than the *who* *wh*-words at the beginning of the sentence and subject *wh*-phrases faster than object *wh*-constituents. This suggests that the processing of some syntactic dependencies is preserved in child HL. More studies are however needed to confirm the possible absence of cross-linguistic influence on processing strategies in the HL, as well as to uncover the role that language dominance plays in child HL processing. Our findings seem to be at odds with those of Van Dijk et al. (2022) who report effects of cross-linguistic influence on the processing of V2

structures in Dutch by German-Dutch bilingual children. These effects were more pronounced the more dominant the children were in German. Van Dijk et al.'s results also show that such CLI effects were stronger in instances of partial structural overlap between German and Dutch and were evident as inhibition during listening. In other words, the German-dominant children slowed down when listening to structures in Dutch that had a similar V2 order in German. Although the Romanian-English bilingual children in this study were dominant in English, as measured through their current expressive skills reported by the parents, we did not find evidence of cross-linguistic influence on the processing of multiple interrogatives in Romanian. One potential explanation is the lack of overlap in surface structure between multiple interrogatives in Romanian and English. Future research should thus address the role that structural overlap and language dominance have in modulating online processing in the heritage language.

Let us now turn to the elicitation task which examined whether Romanian heritage and monolingual children are able to produce questions with multiple *wh*-movement and whether they have fully acquired the specific syntax of multiple interrogatives in Romanian, which requires fronting of all *wh*-phrases. Contrary to comprehension, we found differences in the production of multiple *wh*-questions between Romanian heritage and monolingual children. Monolingual children produce questions with multiple *wh*-fronting (mainly SO *who*), but also questions in which only one *wh*-phrase is fronted, the other one remaining *in-situ*. This option exploited in production surfaces mostly in questions with two *which*-phrases. In addition, monolingual children also produce a significant number of simple *wh*-questions, that is, questions with only one *wh*-word. While heritage children also produce simple questions, they produce significantly more questions with one fronted *wh*-phrase, one *in-situ* than monolingual children and, with the exception of one child, avoid multiple *wh*-movement, contrary to monolinguals. Importantly, there were no instances of multiple *who*-questions with an ungrammatical object-subject order in any of the children's productions. This indicates that their grammatical system does not allow this option and further reinforces the idea that their lack of sensitivity to the object-over-subject ungrammaticality in multiple *who*-questions in on-line comprehension is not due to a different grammar than that of adults, but it most likely stems from the processing load associated with such complex structures involving multiple movement dependencies.

The results for production thus suggest that heritage children seem to opt for a less complex structure that involves fronting of only one *wh*-phrase. Monolingual Romanian children also make use of a range of structures when prompted to produce *wh*-questions with multiple fronting, pointing to the fact that they avoid as well the complexity associated with multiple *wh*-interrogatives which require movement of two *wh*-phrases. Importantly, what the results of the monolingual children show is that they also employ two structural options to derive multiple interrogatives, the multiple *wh*-fronting and the one *wh*-moved, one *wh-in situ* option. Given that the predominant response pattern for multiple *wh*-questions in Romanian heritage children

makes use of the only structural option present in English, we take this to show that there is cross-linguistic influence from the dominant societal language to the heritage language. Similarly, other studies have linked the differences in performance between child heritage speakers and monolinguals to the properties of the societal language (Meir et al., 2017; Meir and Janssen, 2021). The use of the one *wh*-fronted, one *in-situ* option is reinforced in Romanian under influence from English, the dominant language for the heritage children (Hulk and Müller, 2000; Serratrice, 2013). Our findings for production also match those reported by Strik and Pérez-Leroux (2011) for the production of *wh*-questions in Dutch by Dutch-French bilingual children, who were French dominant, and who also produced instances of *wh in-situ* in Dutch, a less complex option than *wh*-fronting. The results from the current study, together with those of Strik and Pérez-Leroux (2011) seem to suggest that structures which require less complex structural derivations (such as one *wh*-fronted, one *in-situ*) are acquired earlier and also that they are more likely to influence structures which require more complex derivations (like multiple *wh*-fronting), in line with a complexity-based theory of transfer (Strik and Pérez-Leroux, 2011).

The question that remains is why cross-linguistic influence occurred in children's production but not in their comprehension. When comparing the results for both comprehension and production of multiple *wh*-dependencies in Romanian we observe that heritage children are able to establish the underlying representation of multiple *wh*-movement structures, similarly to monolinguals when they encounter multiple fronted *wh*-movement structures, but have difficulties activating the more complex structure in production. Such comprehension/production asymmetries have been attested in the majority language of bilingual children for tense (Chondrogianni and Marinis, 2012), articles (Chondrogianni et al., 2015a,b), articles and clitics (Chondrogianni et al., 2015a,b) and have been taken as evidence in favour of the claim that underlying syntactic representations are intact in child L2 acquisition even if non-target-like structures appear in production (Haznedar and Schwartz, 1997). This suggests that in comprehension, where the referents and the linguistic structure are given, children can parse and assign an interpretation to the structure. In doing so, they have to keep in working memory information about the *wh*-fronted elements and then retrieve them in order to establish the correct dependencies between the moved *wh*-phrases and the verb. In production, on the other hand, they have to start at the conceptual level, they have to plan and build the structure themselves by deciding about the thematic role, case, grammatical function, and syntactic position of the *wh*-phrases as they speak. (Momma, 2021). In other words, comprehension requires children to recognize the meaning of words and the syntactic dependencies in which the words enter. But in the light of production, children must actively plan the structure and its complexity impacts on this. It is thus more economical to start from a simpler structure than generating a more complex

structure, particularly when having to produce structures that are not very frequently used in the input children receive, as is the case for multiple interrogatives. Furthermore, the finding that differences between the heritage and the monolingual children surface only in production, while similar patterns emerge in the two groups in the real-time comprehension of multiple interrogatives, also suggests that on-line methods can better reflect competence in HL (see Villegas (2014) for similar results with adult heritage speakers).

An interesting observation here is the fact that children's production patterns mirror the errors that they produce in the comprehension task. For example, reversed role errors are linked to children's preference to mainly produce questions with a subject-object order for the conditions containing *which*-elements. Their production of simple *wh*-questions mirrors the list answers they give in the comprehension task. To recall, these were answers in which children answered only one *wh*-word (either by listing all the Agent characters in the image or by listing the Patient characters), indicating that they treat multiple *wh*-questions as simple questions. These errors in comprehension, together with children's production patterns, show that even monolingual Romanian-speaking children take longer to produce structures with multiple *wh*-fronting and that they also make use of the option of having one *wh*-fronted and one *in-situ*. This corroborates Grebenyova's (2011) findings for Russian as she shows that monolingual Russian-speaking children 4 to 6 years of age also have more difficulties with the production of multiple *wh*-questions compared to monolingual English-speaking children of the same age. She postulates that Russian-speaking children go through an intermediate phase when acquiring the syntax of multiple interrogatives in Russian and that, with enough Russian input, they will produce questions with multiple *wh*-fronting, similarly to adults. Although it is not very clear what would count as "enough Russian input" given that children rarely hear such multiple *wh*-questions to begin with, this view has interesting implications for our study as it suggests that, while monolingual Romanian children will eventually converge on the correct production of multiple *wh*-fronting questions, for the Romanian heritage children this will depend on the amount of Romanian input they receive. Future research with Romanian-dominant bilinguals and with Romanian heritage adults can shed light on this.

Conclusion and future research

Our study makes a substantial contribution to the understanding of child HL development, as it investigates the production and comprehension of the same phenomenon in HL speakers and provides insight into the language development stages of both monolingual and heritage bilingual children. The current study is among the first to investigate cross-linguistic influence in bilingual children

during real-time sentence processing and the first to use the self-paced listening paradigm with heritage children. However, the study also paves the way to future questions that remain unaddressed, such as when the syntax of multiple interrogatives is fully acquired as well what happens with adult heritage speakers or child heritage speakers whose dominant language(s) also have multiple wh-movement or lack this type of questions entirely. Studies that compare multiple bilingual groups will further increase our understanding of the impact of cross-linguistic influence on heritage language development.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Ethics Committee of the University of Reading. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

AB carried out the data collection and was responsible for data analysis. AB and TM shared responsibility for the conception of the work and the interpretation of results. AB wrote the manuscript with input from TM. All authors contributed to the article and approved the submitted version.

References

- Aarts, F. (1994). Relative who and whom: prescriptive rules and linguistic reality. *Am. Speech* 69, 71–79. doi: 10.2307/455950
- Adani, F., and Fritzsche, T. (2015). "On the relation between implicit and explicit measures of child language development: evidence from relative clause processing in 4-year-olds and adults," in *Proceedings of the 39th Annual Boston University Conference on Language Development*. eds. E. Grillo and K. Jepson (Somerville, MA: Cascadia Press), 14–26.
- Alboiu, G. (2002). *The Features of Movement in Romanian*. Bucharest: Bucharest University Press.
- Argyri, E., and Sorace, A. (2007). Crosslinguistic influence and language dominance in older bi-lingual children. *Biling. Lang. Cogn.* 10, 79–99. doi: 10.1017/S1366728906002835
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bayram, F., Pisa, G., Rothman, J., and Slabakova, R. (2021). "Current trends and emerging methodologies in charting heritage language grammars," in *The Cambridge Handbook of Heritage Languages and Linguistics Cambridge Handbooks in Language and Linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 545–578.
- Benmamoun, E., Montrul, S., and Polinsky, M. (2013). Heritage languages and their speakers: opportunities and challenges for linguistics. *Theor. Ling.* 39, 129–181. doi: 10.1515/tl-2013-0009
- Bentea, A. (2010). "Multiple Wh-questions in child language: a cross-linguistic perspective," in *New Directions in Language Acquisition: Romance Languages in the Generative Perspective*. eds. P. Guijarro-Fuentes and L. Domínguez (Newcastle upon Tyne: Cambridge Scholars Publishing), 1–485.
- Bentea, A., and Durrleman, S. (2013). "A'-dependencies in French: a study in L1 acquisition," in *Romance Languages and Linguistic Theory 2011. Selected Papers From 'Going Romance' Utrecht 2011*. eds. S. Baauw, F. A. C. Drijkoningen, L. Meroni and M. Pinto (Amsterdam: John Benjamins), 1–16.
- Bentea, A., and Marinis, T. (2021). Not all wh-dependencies are created equal: processing of multiple wh-questions in Romanian children and adults. *Appl. Psycholinguist.* 42, 825–864. doi: 10.1017/S0142716421000059
- Blom, E. (2010). Effects of input on the early grammatical development of bilingual children. *Int. J. Biling.* 14, 422–446. doi: 10.1177/1367006910370917
- Bosch, J. E., and Unsworth, S. (2020). Cross-linguistic influence in word order: effects of age, dominance and surface overlap. *Ling. Approach. Biling.* 11, 783–816. doi: 10.1075/lab.18103.bos
- Chomsky, N. (1973). "Conditions on transformations," in *A Festschrift for Morris Halle*. eds. S. Anderson and P. Kiparsky (New York: Holt, Reinhart & Winston), 232–286.
- Chondrogianni, V., and Marinis, T. (2012). Production and processing asymmetries in the acquisition of tense morphology by sequential bilingual children. *Biling.: Lang. Cogn.* 15, 5–21. doi: 10.1017/S1366728911000368
- Chondrogianni, V., Marinis, T., Edwards, S., and Blom, E. (2015a). Production and on-line comprehension of definite articles and clitic pronouns by Greek

Funding

This research was supported by an Early Postdoc.Mobility grant for AB from the Swiss National Science Foundation (grant number 174870).

Acknowledgments

Parts of this paper were presented at the 3rd International Symposium on Bilingual and L2 Processing in Adults and Children (ISBPAC 2021) and the Heritage Languages Around the World (HLAW 2022) conference. We thank the audiences there for their valuable input. We also extend our gratitude to all the children who took part in this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

sequential bilingual children and monolingual children with specific language impairment. *Appl. Psycholinguist.* 36, 1155–1191. doi: 10.1017/S0142716414000101

Chondrogianni, V., Vasic, N., Marinis, T., and Blom, E. (2015b). Production and on-line comprehension of definiteness in English and Dutch by monolingual and sequential bilingual children. *Second. Lang. Res.* 31, 309–341. doi: 10.1177/0267658314564461

Contemori, C., Carlson, M., and Marinis, T. (2018). On-line processing of English which - questions by children and adults: a visual world paradigm study. *J. Child Lang.* 45, 415–441. doi: 10.1017/S0305000917000277

Cuza, A. (2016). The status of interrogative subject–verb inversion in Spanish-English bilingual children. *Lingua* 180, 124–138. doi: 10.1016/j.lingua.2016.04.007

Daskalaki, E., Chondrogianni, V., Blom, E., Argyri, F., and Paradis, J. (2019). Input effects across domains: the case of Greek subjects in child heritage language. *Second. Lang. Res.* 35, 421–445. doi: 10.1177/0267658318787231

Friedmann, N., Belletti, A., and Rizzi, L. (2009). Relativized relatives: types of intervention in the acquisition of A-bar dependencies. *Lingua* 119, 67–88. doi: 10.1016/j.lingua.2008.09.002

Grebenyova, L. (2005). “Multiple interrogatives in child language,” in *Proceedings of the 30th Annual Boston University Conference on Language Development*. eds. D. Bamman, T. Magnitskaia and C. Zaller (Somerville, MA: Cascadia Press), 225–236.

Grebenyova, L. (2011). Acquisition of multiple questions in English, Russian, and Malayalam. *Lang. Acquis.* 18, 139–175. doi: 10.1080/10489223.2011.580672

Grimm, A., Müller, A., Hamann, C., and Ruigenduijk, E. (2011). *Production-Comprehension Asymmetries in Child Language*. Berlin and Boston: De Gruyter Mouton.

Haiden, M., Prévost, P., Tuller, L., Ferré, S., and Scheidnes, M. (2009). “Production and comprehension of wh-questions in acquisition of French: Comparing L2 children and L1 children with SLI,” in *Paper presented at the Workshop on Production-comprehension asymmetries in child language, Annual Meeting of the German Linguistics Society*, 3–6 March 2009; University of Osnabrück, Osnabrück, Germany.

Haznedar, B., and Schwartz, B. (1997). “Are there optional infinitives in child L2 acquisition?” in *Proceedings of the 21st Annual Boston University Conference on Language Development*. eds. E. Hughes, M. Hughes and A. Greenhill (Somerville, MA: Cascadia Press), 257–268.

Hendriks, P. (2014). *Asymmetries Between Language Production and Comprehension*. Dordrecht: Springer.

Hendriks, P., and Koster, C. (2010). Production/comprehension asymmetries in language acquisition. *Lingua* 120, 1887–1897. doi: 10.1016/j.lingua.2010.02.002

Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., and Snider, N. (2013). The source ambiguity problem: distinguishing the effects of grammar and processing on acceptability judgments. *Lang. Cogn. Process.* 28, 48–87. doi: 10.1080/01690965.2011.572401

Hulk, A., and Müller, N. (2000). Bilingual first language acquisition at the interface between syntax and pragmatics. *Biling. Lang. Cogn.* 3, 227–244. doi: 10.1017/S1366728900000353

Jegerski, J., and Sekerina, I. (2021). “The psycholinguistics of heritage languages,” in *The Cambridge Handbook of Heritage Languages and Linguistics Cambridge Handbooks in Language and Linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 449–470.

Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Laenzlinger, C., and Soare, G. (2005). A cartographic approach to Wh-movement in Romanian. *Working Papers in Linguistics* 1, 23–60. Bucharest: Editura Universității București.

Lemmerth, N., and Hopp, H. (2019). Gender processing in simultaneous and successive bilingual children: cross-linguistic lexical and syntactic influences. *Lang. Acquis.* 26, 21–45. doi: 10.1080/10489223.2017.1391815

Lenth, R. V. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.7.3. Available at: <https://CRAN.R-project.org/package=emmeans>

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Măniță, C. (2017). *Asymmetries in the Acquisition of Wh-questions: The Role of Intervention and Interference*. [Doctoral dissertation]. University of Bucharest, Bucharest.

Marinis, T. (2007). “On-line processing of passives in L1 and L2 children,” in *Proceedings of the 2nd Conference on Generative Approaches to Language Acquisition*

North America (GALANA). eds. A. Belikova, L. Meroni and M. Umeda (Cascadia Proceedings Project: Somerville, MA), 265–276.

Marinis, T., and Saddy, D. (2013). Parsing the passive: comparing children with specific language impairment to sequential bilingual children. *Lang. Acquis.* 20, 155–179. doi: 10.1080/10489223.2013.766743

Martinez-Nieto, L., and Restrepo, A. M. (2022). Production and comprehension of grammatical gender by Spanish heritage speakers: evidence from accusative clitic pronouns. *Int. J. Biling.* doi: 10.1177/13670069211057318

Meir, N., and Janssen, B. (2021). Child heritage language development: an interplay between cross-linguistic influence and language-external factors. *Front. Psychol.* 12:651730. doi: 10.3389/fpsyg.2021.651730

Meir, N., Walters, J., and Armon-Lotem, S. (2017). Bi-directional cross-linguistic influence in bilingual Russian-Hebrew children. *Ling. Approach. Biling.* 7, 514–553. doi: 10.1075/lab.15007.mei

Momma, S. (2021). Filling the gap in gap-filling: long-distance dependency formation in sentence production. *Cogn. Psychol.* 129:101411. doi: 10.1016/j.cogpsych.2021.101411

Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.

Müller, N., and Hulk, A. (2001). Crosslinguistic influence in bilingual language acquisition: Italian and French as recipient languages. *Biling. Lang. Cogn.* 4, 1–21. doi: 10.1017/S1366728901000116

Nicoladis, E. (2006). Cross-linguistic transfer in adjective–noun strings by preschool bilingual children. *Biling. Lang. Cogn.* 9, 15–32. doi: 10.1017/S136672890500235X

Nicoladis, E. (2012). Cross-linguistic influence in French–English bilingual children’s possessive constructions. *Biling. Lang. Cogn.* 15, 320–328. doi: 10.1017/S1366728911000101

Nicoladis, E., and Gavrilu, A. (2015). Cross-linguistic influence in Welsh–English bilingual children’s adjectival constructions. *J. Child Lang.* 42, 903–916. doi: 10.1017/S0305000914000440

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., et al. (2019). PsychoPy2: experiments in behavior made easy. *Behav. Res. Methods* 51, 195–203. doi: 10.3758/s13428-018-01193-y

Pesetsky, D. (2000). *Phrasal Movement and its Kin*. Cambridge, MA: MIT Press.

Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.

Polinsky, M., and Scontras, G. (2020a). A roadmap for heritage language research. *Biling. Lang. Cogn.* 23, 50–55. doi: 10.1017/S1366728919000555

Polinsky, M., and Scontras, G. (2020b). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245

R Core Team (2022). R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org/> (Accessed March 10, 2022).

Richards, N. (1997). What moves where when in which language? Dissertation thesis. Cambridge (MA): MIT.

Rizzi, L. (1997). “The fine structure of the left periphery,” in *Elements of Grammar*. ed. L. Haegeman (Dordrecht: Kluwer), 281–337.

Rizzi, L. (2004). “Locality and the left periphery,” in *Structures and beyond: The cartography of syntactic structures*. ed. A. Belletti, vol. 3 (New York: Oxford University Press), 223–251.

Rodina, Y., Kupisch, T., Meir, N., Mitrofanova, N., Urek, O., and Westergaard, M. (2020). Internal and external factors in heritage language acquisition: evidence from heritage Russian in Israel, Germany, Norway, Latvia and the United Kingdom. *Front. Educ.* 5:20. doi: 10.3389/educ.2020.00020

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *J. Mem. Lang.* 110:104038. doi: 10.1016/j.jml.2019.104038

Serratrice, L. (2013). Cross-linguistic influence in bilingual development determinants and mechanisms. *Ling. Approach. Biling.* 3, 3–25. doi: 10.1075/lab.3.1.01ser

Serratrice, L., Sorace, A., Filiaci, F., and Baldo, M. (2012). Pronominal objects in English–Italian and Spanish–Italian bilingual children. *Appl. Psycholinguist.* 33, 725–751. doi: 10.1017/S0142716411000543

Serratrice, L., Sorace, A., and Paoli, S. (2004). Crosslinguistic influence at the syntax–pragmatics interface: Subjects and objects in English–Italian bilingual and monolingual acquisition. *Biling. Lang. Cogn.* 7, 183–205. doi: 10.1017/S1366728904001610

Soare, G. (2009). *The Syntax-Information Structure Interface: A Comparative View from Romanian*. Dissertation thesis. Geneva: University of Geneva.

Strik, N., and Pérez-Leroux, A. T. (2011). Jij doe wat girafe? Wh-movement and inversion in Dutch-French bilingual children. *Ling. Approach. Biling.* 1, 175–205. doi: 10.1075/lab.1.2.03str

Tuller, L. (2015). “11. Clinical use of parental questionnaires in multilingual contexts,” in *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. eds. S. Armon-Lotem, J. de Jong and N. Meir (Bristol, Blue Ridge Summit: Multilingual Matters), 301–330.

van Dijk, C., Dijkstra, T., and Unsworth, S. (2022). Cross-linguistic influence during online sentence processing in bilingual children. *Biling. Lang. Cogn.* 25, 691–704. doi: 10.1017/S1366728922000050

van Dijk, C., van Wonderen, E., Koutamanis, E., Kootstra, G. J., Dijkstra, T., and Unsworth, S. (2021). Cross-linguistic influence in simultaneous and early sequential bilingual children: a meta-analysis. *J. Child Lang.* 49, 897–929. doi: 10.1017/S0305000921000337

Villegas, Á. (2014). *The role of L2 English immersion in the processing of L1 Spanish sentence complement/relative clause ambiguities*. [Doctoral dissertation]. Pennsylvania State University.

Yip, V., and Matthews, S. (2006). Assessing language dominance in bilingual acquisition: a case for mean length utterance differentials. *Lang. Assess. Q.* 3, 97–116. doi: 10.1207/s15434311laq0302_2



OPEN ACCESS

EDITED BY

Maki Kubota,
UiT The Arctic University of Norway,
Norway

REVIEWED BY

Vsevolod Kapatsinski,
University of Oregon,
United States
Mike Putnam,
The Pennsylvania State University (PSU),
United States
Cristina Maria Flores,
University of Minho,
Portugal

*CORRESPONDENCE

Silvia Perez-Cortes
sp1019@camden.rutgers.edu

[†]These authors share first authorship

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 25 July 2022

ACCEPTED 14 October 2022

PUBLISHED 23 November 2022

CITATION

Perez-Cortes S and Giancaspro D (2022)
(In)frequently asked questions: On types of
frequency and their role(s) in heritage
language variability.
Front. Psychol. 13:1002978.
doi: 10.3389/fpsyg.2022.1002978

COPYRIGHT

© 2022 Perez-Cortes and Giancaspro. This
is an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

(In)frequently asked questions: On types of frequency and their role(s) in heritage language variability

Silvia Perez-Cortes^{1*†} and David Giancaspro^{2†}

¹Department of World Languages and Cultures, Rutgers University–Camden, Camden, NJ, United States, ²Department of Latin American, Latino and Iberian Studies, University of Richmond, Richmond, VA, United States

In recent years, researchers have become increasingly interested in exploring frequency as a source of variability in heritage speakers' (HSs) knowledge of their heritage language (HL). While many of these studies acknowledge that frequency can affect the shape of HL grammars, there is still no clear consensus about (a) what "frequency" means in the context of HL acquisition and (b) how to operationalize its multiple subtypes. In this paper, we provide a critical overview of frequency effects in HL research and their relevance for understanding patterns of inter/intra-speaker variability. To do so, we outline how prior research has defined, measured, and tested frequency, and present—as well as evaluate—novel methodological approaches and innovations recently implemented in the study of frequency effects, including a new analysis of how self-reported lexical frequency reliably predicts HSs' production of subjunctive mood in Spanish. Our aim is to highlight the immense potential of such work for addressing long-standing questions about HL grammars and to propose new lines of inquiry that will open up additional pathways for understanding HL variability.

KEYWORDS

frequency, variability, heritage speakers, activation, lexicon

Introduction

Despite its exponential growth in recent years, the field of heritage bilingualism is still relatively young—especially in comparison to first (L1) and second (L2) acquisition research. While every year heritage language (HL) investigators continue to expand the reach of their work in terms of linguistic content and methodological approaches, research in this field has primarily focused on two lines of inquiry. The first one is centered around *between-group comparisons* and focuses on identifying areas where heritage speakers (HSs) differ from—or pattern like—other speaker groups, perhaps due to the enduring influence of L2 acquisition research—much of which examines differences between L2 learners and (monolingual) native speakers (e.g., [White and Genesee, 1996](#)). The second area of research involves *between-property comparisons*, which address the relative difficulty—or, to use a

more commonly employed term, vulnerability (e.g., Polinsky and Scontras, 2020)—of different properties of the HL grammar (e.g., tense/aspect vs. mood morphology in the verbal domain: Montrul, 2009).

In our estimation, the pursuit of these two lines of inquiry has led to the vast majority of what we now know about HSs and their grammars. Nonetheless, we must recognize that, as informative as they have been, neither line of research addresses certain fundamental—and often overlooked—puzzles of HL research, especially those that involve the study of variability at the intraspeaker level. In this article, we advocate for increased attention to two promising yet less commonly investigated areas of study, each of which we will summarize here and then elaborate upon throughout the remainder of the paper. Critically, both approaches open the door to fine-grained analyses of frequency, a variable that has thus far received much more attention in L1 (e.g., Ambridge et al., 2015) and L2 (e.g., Ellis, 2002) research than in work with HSs, in spite of its evident potential in this area (e.g., O'Grady et al., 2011; Putnam and Sánchez, 2013; Montrul et al., 2014; Hur, 2020; López-Beltrán and Carlson, 2020; Giancaspro et al., 2022; Perez-Cortes, 2022b).

The first of the two categories of frequency that we will examine in this article, frequency of HL activation, involves the study of *between-speaker comparisons*, that is, the analysis of the unique grammatical systems that develop in the minds of individual HSs. Although inter-speaker heterogeneity is a defining—and well-recognized—characteristic of HSs (e.g., Montrul, 2016), we still have a lot to learn about the underlying factors that might cause two HSs with seemingly similar demographic/linguistic profiles to end up with what appear to be very differently shaped HL grammars. Why, for example, might one HS produce high rates of a certain inflection while another goes to great lengths to avoid it (e.g., Giancaspro, 2019)? Conceivably, part of this gap in our understanding results from the field's longstanding reliance on group-level inferential analyses, which often take center stage in HL acquisition studies. Regardless of the reason for the scarcity of between-speaker analyses, variability at this level is a well-attested pattern that needs to be explored if we are to improve our understanding of the gradience of outcomes observed among HSs. After all, as much as we rely on group-level analyses in our quest to comprehend heritage bilingualism, linguistic systems form (and transform) in individual minds—not in groups. As such, our models and theories must also speak to the nature of these individual acquisitional paths.

The second type of frequency examined, which includes lexical frequency in its many instantiations, allows for *within-speaker comparisons*, that is, the study of variability that arises in individual HSs with a single HL property. When HSs differ from comparison groups, as they often do, contrasts tend to emerge in a variable rather than a categorical manner. For instance, for a given grammatical property, X, HSs will often produce both the instantiation that we usually see in control groups of monolingual speakers, as well as other (often innovative) variants (Flores,

2015). Despite its near ubiquity in HL research, this pattern of intra-speaker variability has also received relatively little attention from HL researchers, perhaps partially due to the complexity of accounting for such variability *via* formal linguistic theory, which often—though not exclusively (e.g., Sorace and Keller, 2005; Putnam et al., 2018)—views grammatical operations as categorical rather than gradient.

In the present paper, we argue for the importance of prioritizing the inclusion of frequency-based analyses in future empirical work, which enable the exploration of *between-speaker* and *within-speaker* comparisons. Despite their superficial differences, both types of comparisons have the potential to illuminate the multidimensional relationship between HSs' linguistic knowledge, on one hand, and their language experience—operationalized via frequency—on the other. As we will outline in “Between-speaker comparisons: frequency of heritage language activation,” the analysis of HSs' frequency of use/activation of their HL offers the unique chance for researchers to draw informative connections between speakers' individual linguistic experience and their command of—or variability with—specific HL properties. We propose that by examining the extent to which individual HSs differ in their patterns of HL use, we can shed new light on HL heterogeneity, a puzzle that cannot be solved with *between-group* or *between-property* comparisons. Lexical frequency, which will be the focus of “Lexical frequency and its role in heritage grammars,” provides an opportunity for researchers to further interrogate HL variability, this time, at the level of individual speakers. Like between-speaker variability, intra-speaker variability, too, is a micro-level pattern that simply cannot be addressed by looking at the macro-level comparisons—specifically, between-speaker and between-property comparisons—that continue to predominate in the field. After reviewing different approaches to conceptualizing lexical frequency, we argue that subjective or self-reported lexical frequency—that is to say, HSs' own evaluation of how often they hear/use certain words—can help us to explain individual HSs' alternation between “target” and innovative variants of a given HL property. We conclude the paper in “Discussion and conclusion: Some final thoughts” by (a) proposing that frequency-based analyses should play a key role in our study of between-speaker and within-speaker HL research and (b) sketching out future directions for investigations that follow this blueprint.

Between-speaker comparisons: frequency of heritage language activation

Regardless of theoretical background, HL researchers largely agree that HSs' overall amount of experience with the HL, broadly conceived, will strongly affect the HL grammars that they ultimately develop (e.g., Kupisch and Rothman, 2018; Polinsky and Scontras, 2020; Montrul, 2021b *inter alia*). Thus,

one would expect—all else equal—that a HS with extensive HL experience would perform in a less innovative (or more “target-like”) manner than a comparable peer whose use of the HL is relatively less frequent. Despite the consensus on this general—and perhaps obvious—point, researchers have dedicated relatively little attention to the question of how (and to what extent) between-speaker differences in HL usage/exposure/experience might lead different HSs to develop distinct patterns of grammatical knowledge with a given HL property. In fact, as noted by Daskalaki et al. (2019: 423), “no study to date has examined the differential effect of input quantity, as a continuous variable, within the same group of heritage speakers.” Before reviewing five studies that have shed light—directly or indirectly—on effects of HL usage/exposure/experience, we first summarize two papers that formalize how differences in HSs’ experience with the HL, sometimes referred to as frequency of activation, might lead to between-speaker differences in HL knowledge.

Putnam and Sánchez (2013), working from a generative theoretical framework, argue that HSs’ frequency of activation of their HL will impact the shape of their HL grammar. From this vantage point, HSs who use their HL less frequently might experience “a decline in the availability of FFs [functional features]” (p. 484) of their HL, which often manifests as innovative patterns of HL production and/or comprehension. Putnam and Sánchez’s conceptualization of heritage bilingualism creates a basic framework for accommodating differences in the performance of HSs. While those who frequently activate the HL—Stage 1 HSs in their terminology—are unlikely to exhibit major morphosyntactic innovations in their HL, HSs who use this language far less often (e.g., “Stage 3” and “Stage 4” HSs) will demonstrate much more variability and innovation, in large part due to the increasing inaccessibility of FFs in their (relatively less activated) HL system. In a study of HSs’ production and comprehension of subjunctive mood, Perez-Cortes et al. (2019) provide evidence that is consistent with Putnam and Sánchez’s proposed stages. While high-activation HSs—operationalized as HSs with higher HL proficiency—performed in a more “target-like” manner with subjunctive mood, lower-activation HSs performed more variably, exhibiting increasingly prominent production/comprehension asymmetries as their proficiency in the HL decreased. More important than the specific details of their proposals, we believe, is these authors’ novel attempt to articulate how HSs with different levels of HL activation might end up with differentially innovative HL grammars.

Measuring language activation and its effects on HL grammars: a complex enterprise

Since the publication of Putnam and Sánchez’s (2013) foundational work, additional evidence has emerged that aligns with its basic principles. A key example comes from studies that

identify variability across HSs and descriptively explore the extent to which differences in exposure to the HL could be the source of said between-speaker contrasts. Cuza (2016), in a study of subject-verb (SV) inversion in the Spanish of younger and older child HSs, observed that younger HSs performed in a more “target-like” manner than their older counterparts, a finding that Cuza attributes—at least in part—to patterns of HL usage. Relative to the younger HSs, the older HSs in the study reportedly used less Spanish with their parents and siblings, which could, in principle, contribute to their differential knowledge of SV inversion. While suggestive, this trend could also have been caused by other differences between the two groups, such as older HSs’ emerging dominance in English.

Montrul and Sánchez-Walker (2013), in an extensive investigation of simultaneous and sequential HSs’ production of differential object marking (DOM) in Spanish, found that HSs produced less DOM in expected contexts than both monolingual and bilingual “baseline” groups. Far more revealing than these between-group differences, however, were the extensive between-speaker differences in the HS group—particularly in the case of the simultaneous HSs, whose production of DOM in expected contexts ranged from 0% to 100%. In an attempt to better understand these between-speaker differences, Montrul and Sánchez-Walker (2013) divided their HS participants into two groups—omitters, who produced DOM less than 80% of the time in expected contexts, and non-omitters—who produced DOM categorically. Post-hoc analyses revealed that, relative to the omitters, the non-omitters reported using Spanish more often in a variety of different situations, including with their parents, siblings, and friends. Though only a few of the differences between these two groups were statistically significant, these analyses point to the possibility that differences in HL usage can, in fact, trigger measurable between-speaker grammatical differences.

More recently, two additional studies have strengthened the claim that HSs’ frequency of experience with their HL shapes the variability of their HL grammatical systems. What sets these studies apart from Cuza (2016) and Montrul and Sánchez-Walker (2013) is that in each case, HL experience is seamlessly integrated into the inferential statistical modeling, thereby allowing for more reliable insight into the effects of this potentially critical explanatory variable. Dracos and Requena (2022) investigated child HSs’ production of a few different types of subjunctive mood morphology in Spanish, including, most relevantly for the present study, and volitional subjunctive forms (e.g., *quiero que bailes_{SUBJ}* ‘I want you to dance’). Critically—and in contrast with previous studies of HSs and subjunctive mood—Dracos and Requena’s (2022) analyses incorporated information about HL usage/exposure, which they combined into a single variable that was included as a fixed factor in their mixed-effects statistical models. Results indicated that HSs with higher use of the HL (as reported by their caretakers) were significantly more likely to produce subjunctive mood morphology, pointing to HL experience as a factor in explaining certain between-speaker

differences¹. Complicating this finding, however, is the fact that HL proficiency—which is strongly correlated with HL use—was an even stronger predictor of HSS' performance, thereby highlighting the difficulty of isolating HL experience as a cause of between-speaker differences.

Perhaps the most thorough attempts to connect the linguistic experiences of HSSs to the grammatical systems they develop are López-Beltrán (2021), who tested the subjunctive mood knowledge of (adult) HSSs of Spanish living in long-standing bilingual communities in New Mexico, and López Otero et al. (2021), who tested Spanish clitic production by adult HSSs living in Brazil. Simplifying greatly, López-Beltrán found, using sophisticated statistical modeling, that HSSs who reported higher use of Spanish were more sensitive to mood violations, as measured in a study of their pupil dilations, which reflect processing difficulty. Similarly, López Otero et al. (2021), who also explored effects of HL usage by treating it as a continuous variable in mixed effects models, found that HSSs with higher HL usage were less likely to exhibit innovative clitic pronoun production.

Moving forward: the future of studying frequency of HL use and exposure

In the studies reviewed above, we have seen preliminary evidence that between-speaker differences—that is to say, differences in the grammatical knowledge of different HSSs—seem to be caused, at least in part, by differences in HSSs' frequency of use of and exposure to the HL. Weakening this conclusion, however, are two major methodological and epistemological concerns. First, it is not yet clear that language background questionnaires offer an accurate or reliable assessment of HSSs' actual patterns of HL use both (a) at the moment of data collection as well as (b) in earlier stages of their lives. This is especially concerning if HL use/exposure is more impactful during early childhood, as suggested by several researchers (e.g., Montrul, 2016; Silva-Corvalán, 2018; López-Beltrán, 2021). So how, exactly, have researchers attempted to quantify HSSs' frequency of HL use/exposure/experience? To illustrate the complexity of this task—and underline the need for new, methodologically-oriented work in this area—we briefly review the approaches employed in two of the studies presented in the previous section.

Dracos and Requena (2022) calculated (child) HSSs' experience with the HL by asking the HSSs' parents to provide approximate percentages of the time that their children hear/use the HL during the week, as well as on weekends. After receiving the responses, the researchers recoded the data into five different categories (0%–19%, 20%–39%, 40%–59%, 60%–79% and 80%–100%), which became levels of a fixed factor (HL usage/exposure) in their subsequent statistical modeling. While we applaud Dracos and Requena's (2022) utilization of this variable in their analyses,

we must acknowledge, too, the potential unreliability of the percentage estimates that they received. When a parent estimates the percentage of the time that their child uses the HL, what factors might they be considering (or not)? For example, how do they acknowledge, among other potential concerns, language mixing, asymmetric communication (e.g., parent speaks Spanish, and child responds in English), and the generally dynamic nature of HL use in a majority-language dominant society? The proportion of the HL that a child hears, for example, might vary greatly from week to week. In light of these challenges, it should not surprise us, perhaps, that HL proficiency—which may be a more direct measure of HL use/experience than questionnaire data—was a better predictor of subjunctive production than parental HL estimates².

If it is challenging for HSSs' parents to estimate their children's (current) HL use, it is likely even harder for adult HSSs to accurately pinpoint the percentage of the time that they themselves used their HL (at the time of data collection or—cumulatively—throughout their lives). The adult HSSs in López-Beltrán's (López-Beltrán, 2021) study, for example, were asked to determine the percentages of English and Spanish that they heard at home *before* beginning school, a period of time that likely predated their study participation by 13+ years. In households where parents exclusively used Spanish—and required their children to do the same—such estimates may, in fact, be quite reliable. (This may be why HL use was, after all, a statistically significant predictor in López-Beltrán's study). In households with more varied language practices, however, it is difficult to imagine that college-aged HSSs could accurately and reliably recall percentages of their overall language usage during childhood. Consequently, differences in the middle of the (estimated) HL usage spectrum—e.g., between HSSs who reported using their HL 60% of the time vs. those who reported using it 40% of the time—seem far less likely to effectively predict differences in speakers' command of the HL in adulthood. Critically, this might be the case even if the “true” difference between 40% and 60% HL use does have important effects on adult HSSs' eventual grammatical knowledge.

A second (and related) concern with quantifying the effects of HL use, as mentioned in the preceding paragraphs, is that HL use/exposure is often strongly correlated with—and therefore hard to disentangle from—other potentially influential factors such as HL proficiency, age of acquisition of the majority language, and even formal education in the HL. As noted above,

2 Some studies, such as Unsworth (2013), have found that parental estimates of children's language use/exposure are, in fact, effective predictors of children's grammatical knowledge. The fact that different studies find differentially predictive effects of these estimates, however, may very well be attributable to (a) inconsistencies across the conceptualization of language use/exposure across different language background questionnaires (Kascelan et al., 2022) and/or (b) practical difficulties in disentangling language use/exposure from related and potentially confounding variables (e.g., proficiency).

1 For a similar finding with child HSSs of Greek, see Daskalaki et al. (2019).

Dracos and Requena (2022)—as well as López-Beltrán—found that both HL use/experience and HL proficiency were statistically significant predictors of HSs' grammatical performance, making it impossible to isolate the specific influence of HL use itself. (In the first of these studies, recall that proficiency was actually a stronger predictor than reported HL experience). Given this conceptual difficulty³, it may be the case—in spite of the suggestive evidence presented above—that our knowledge of how HL exposure/use affects HL grammars remains quite limited. Furthermore, even if we could design a perfectly reliable background questionnaire that allowed us to isolate the effects of HL usage from other potentially confounding variables, we would still face another major conceptual challenge. When differences in HL use lead different HSs to exhibit differential knowledge of a HL property, *where* exactly do these differences emerge?

To illustrate this conundrum, consider the following hypothetical. John and Carlos are both HSs of Spanish, though John reports using his HL two times as often as Carlos. (For the sake of argument, let us assume that John and Carlos are equivalent in terms of other pertinent background variables, thereby allowing us to isolate the effect of HL usage.) When John and Carlos complete an experimental task designed to assess their knowledge of mood morphology, John produces subjunctive in 80% of expected contexts while Carlos only does so in 40% of the same contexts. This hypothetical between-speaker difference would appear to indicate that Carlos' HL usage affects his production of subjunctive mood. Nonetheless, this finding does not tell us where, at a fine-grained grammatical level, the two speakers differ from one another. It is possible, for example, that Carlos tends to use subjunctive with irregular verbs, or with forms that are more frequent, or even in contexts that are more likely to appear in academic/formal registers. In any case, the purpose of this example is to show that identifying between-speaker differences, though important, only provides indirect insight into individual patterns of HL development. To understand variability within a single speaker, that is to say, what factors lead Carlos to alternately produce both subjunctive, *and* indicative when subjunctive is expected, we will need to make

within-speaker comparisons (e.g., with lexical frequency), which offer the micro-level perspective necessary for understanding individual HL grammatical patterns. In “Lexical frequency and its role in heritage grammars,” we elaborate on this point, using lexical frequency effects as an illustrative test case.

Lexical frequency and its role in heritage grammars

In “Between-speaker comparisons: frequency of heritage language activation,” we discussed the advantages as well as limitations of how the field of HL acquisition has examined and modeled the development and outcomes of heritage bilinguals based on their patterns of language activation, that is, the frequency (or lack thereof) with which HSs use (and are exposed to) their HL. There is, however, another more fine-grained way in which frequency has been implemented to analyze patterns of language maintenance and change/innovation among heritage bilinguals. This second “type” of frequency (henceforth *lexical frequency*), which has recently emerged as an area of interest in HL research (Zyzik, 2016, 2019; Giancaspro, 2020; Gonzalez, 2020; Hur et al., 2020; Camacho, 2022; Giancaspro et al., 2022; Perez-Cortes, 2022b; *inter alia*), addresses the question of how the rate of occurrence of certain forms or structures in the HL input/output may affect their representation, processing, and use (Bybee, 1985, 2007).

While generative approaches to language acquisition have paid relatively little attention to the effects of lexical frequency (see Yang, 2004, 2015 for an exception), usage-based approaches, in contrast, have placed significant importance on this factor, arguing that “the structure and organization of a speaker's linguistic knowledge is the product of language use or performance” (Diessel and Hilpert, 2016). From this perspective, increased (or decreased) exposure to a particular lexical item—based on its likelihood of appearing in the input—would affect how it is accessed, retrieved, and stored (Bybee, 1985, 2007; Poplack et al., 2013). Thus, highly frequent lexical items become the building blocks of grammatical categories, acting as exemplars around which related tokens cluster and establishing—and reinforcing—connections across multiple elements of language in what is known as entrenchment.

The nature and directionality of frequency effects appear to vary depending on the area of language under analysis. While high frequency collocations such as *I do not know* in English are especially vulnerable to phonological change or reduction (Bybee, 2006), high frequency morphological inflections, such as irregular past forms in English (i.e., *bought*, *went*), prove to be much more resistant to overregularization or simplification than less frequent counterparts (i.e., *snuck*, *dove*: Bybee, 1985). Since morphosyntax is the most commonly studied locus of variability in HL grammars (Putnam et al., 2022), and, furthermore, the data that we present in this paper comes from this domain, we narrow our focus in

³ This challenge may also exist in the opposite direction, that is to say, in studies where another between-speaker variable (e.g., formal education) is identified as the underlying driver of between-speaker differences, but overall patterns of HL usage are not considered. Bayram et al. (2017), for example, report that HSs with higher literacy in Turkish are more likely to produce Turkish passive constructions than their less Turkish-literate counterparts, a finding that they attribute to literacy itself. (Notably, all HSs in the study produced passive forms, meaning that literacy cannot be a necessary precondition for their acquisition.) Since the authors do not present data about these groups' overall usage of and exposure to Turkish, it is possible—especially if these variables are highly correlated with literacy—that their inclusion in the statistical modeling might have washed out some of the observed literacy effect.

“Lexical frequency and its role in heritage grammars” to the effects of lexical frequency on HSs’ knowledge of morphosyntactic properties of the HL. To do so, we outline ways to operationalize lexical frequency, summarize cutting-edge studies and their proposals, and suggest future areas of research related to this topic.

As we will argue throughout this section, the limited exposure to (and use of) the HL often observed among HSs provides the perfect backdrop for the study of lexical frequency effects, in part, because this factor establishes a direct—and quantifiable—connection between speakers’ linguistic experience and how they represent and use language. It seems feasible, for example, that the relatively reduced input to which many HSs are exposed could drive them to rely more extensively on highly frequent HL items or structures, likely at the expense of lower frequency forms with which they have much less experience. As a result, properties or forms that are highly frequent in the (baseline) input might become more entrenched in the grammars of HSs, leading to lower levels of optionality in their use⁴. In contrast, less frequent HL forms would be more likely to favor the emergence of grammatical innovations (Backus, 2020) or morphosyntactic variability (Poplack et al., 2013; Perez-Cortes, 2022a). These hypotheses are compatible with recent theoretical proposals regarding the nature of the lexicon, especially those that advocate for an exoskeletal approach to morphology (Embick, 2015; Lohndal and Putnam, 2021). In particular, the adoption of a distributed view of lexical items (as the result of abstract morphosyntactic (*synsem*) features being mapped onto specific (morpho) phonological exponents) provides us with a systematic way to model and predict how frequency in the input could either reinforce such mappings, or allow for a disassociation between them, generating a wide range of outcomes that could have consequences at the level of production as well as representation (Perez-Cortes et al., 2019).

What makes considering the effects of lexical frequency most critical in future HL research, though, is that it allows us to account for differences that emerge at the individual level, that is, those that appear *within* speakers rather than *between* them. This change in perspective provides new and additional explanations to long-standing questions, such as why morphological variability tends to appear in certain forms but not in others, or how HSs’ lexical knowledge affects their overall linguistic development in the HL (Montrul and Mason, 2020; Montrul, 2021a).

4 Consistent with this claim is a recent study by Torregrossa et al. (2022), which found that for HSs of European Portuguese living in Switzerland, HL experience factors (e.g., HL instruction) facilitated performance with more difficult (possibly, less frequent) cloze-test items in the HL (e.g., *ansiosa*: ‘anxious’) but not with lower difficulty (and, presumably, higher frequency) HL words/forms (e.g., *sol*: ‘sun’).

Operationalizing lexical frequency in research: the role of token, type, and lemma

Although it is common for acquisitional studies to refer to lexical frequency in broad terms, Ambridge et al. (2015), who work from a usage-based perspective, argue for a more specific use of this construct. With that in mind, what do HL researchers mean when they talk about the effects of lexical frequency? More often than not, observations about lexical frequency are centered around a word’s *token frequency*—i.e. its overall occurrence in the input⁵—as documented in large language corpora. According to Bybee (2007), forms that exhibit high token frequencies tend to 1) be more autonomous; and 2) have more lexical strength. Together, these factors make it more likely that speakers will access and retrieve frequent forms—which may be stored directly in their lexicon—as whole units or constructions, rather than assembling them derivationally (e.g., *walk* + *-ed* = *walked*). In the context of HSs, this would predict that more frequent items—from a token frequency perspective—would be, as a result of their autonomy/strength, more easily recognized and decoded in comprehension and less likely to exhibit variability in production⁶. Token frequency, however, is not the only existing category of lexical frequency, nor is it the only one that could generate predictions for HL acquisition. The construct of *type frequency*, likely the second most studied frequency category, captures the productivity of a particular pattern in language and accounts for analogical leveling in language acquisition, that is, the (over) application of a specific rule to forms that present relatively less common patterns (Hopper and Bybee, 2001).

Since token and type frequency interact with one another in complex ways (Bybee, 1995; Bybee and Thompson, 2000), we believe that researchers must—to the extent possible—carefully manipulate (or at least, control) these factors when conducting empirical analyses of HSs’ morphological knowledge. A perfect example of this can be found in the formation of past participles, tested among Spanish HSs by Mason (2019). As in English, Spanish past participles are classified based on whether

5 Theoretically, one could (as seen in Giancaspro et al., 2022) account for token frequencies in the output (i.e., production) as well. Although work on monolingual acquisition has postulated that input/output frequency distributions tend to be extremely similar (DePaolis et al., 2011; Ambridge et al., 2015), this might not be the case for HSs, whose opportunities to productively use the HL are much more reduced than those of their monolingual peers (Montrul, 2016).

6 While these tendencies have also been documented by Schmid and Köpcke (2017) in the realm of L1 attrition, it is necessary to note—as pointed out by a reviewer—that they only hold in cases where variability is unintentional, that is, in situations where the speaker is not purposefully modifying a specific item (often prosodically) with a particular communicative intention in mind. See Kapatsinski et al. (2020) for a more detailed discussion.

TABLE 1 Token frequencies of regular and irregular Spanish past participles.

Regularity	Form	Token frequency (ranking/total participles)
Regular	Ser-sido ('been')	4,550,546 (1)
	Llegar-llegado (‘arrived’)	822,102 (14)
Irregular	Hacer—hecho ('had')	1,577,077 (2)
	Poner-puesto ('put')	463,155 (15)

their formation is considered regular or irregular. Regular participles are formed by adding the suffix *-(i)do/-a* to the root of the verb, as in the case of *llegar-llegado* ('to arrive/arrived') or *ser-sido* ('to be/been'). Irregular past participles, by contrast, can present a wider range of morphological instantiations, following patterns such as those in *hacer-hecho* ('to do/done'), or *poner/puesto* ('to put/put'). However, even within the subcategory of irregular past participles that end in *-to*, there are a number of different subpatterns, e.g., *romper/roto* ('to break/broken') or *escribir/escrito* ('to write/written'), to give two quick examples. Thus, if we were to describe Spanish past participles based on their lexical frequency, we could do so in at least two different ways:

a) From a *token* frequency perspective, we could report and contrast their frequencies of occurrence in the input as documented by participant self-reports or by language corpora, such as the [Davies' Spanish NOW corpus \(2012–2016\)](#) that we used to extract the information provided in [Table 1](#). This would allow us to establish differences between how often specific verbal inflections (e.g., *sido*) are used relative to others (e.g., *hecho*). Though both *sido* and *llegado* are regular past participles, for example, *sido* is over five times more frequent; similarly, while *hecho* is an irregular participle, it is used about twice as often as the regular participle, *llegado*. These differences in token frequency—both within and across different types of regularity—could very well affect how HSs (and other Spanish speakers) learn and use participial forms.

b) From a *type* frequency perspective, we could analyze the productivity of the different word-formation patterns involved in the forms under consideration. As proposed by [Mason \(2019\)](#), in this particular case we would be able to identify two large clusters: those observed within regular participles such as *sido* and *llegado*, which present one of two different instantiations (*-ado* or *-ido*); and those observed in irregular forms (i.e., *hecho* and *puesto*). While the regular past participles *-ado* and *-ido* exhibit similarly high type frequencies, irregular participles have a wider range of different allomorphic subpatterns—[Mason \(2019:43\)](#) identifies up to seven—each of which may be relatively more or less common. Presumably, the differences in type frequency across irregular past participles, to give one example, could influence how HSs (and other Spanish speakers) develop their knowledge of participial forms.

Prior work on the acquisition of these structures among Spanish speakers (bilingual and monolingual) has found that irregular verbs (such as *hecho* or *puesto*) tend to be overregularized—e.g., to *hacido* or *ponido*—, especially during the first stages of acquisition ([Clahsen et al., 2002](#); [Soto-Corominas, 2021](#)). In some cases, these “non-target-like” forms may even remain in the repertoire of adult bilinguals, especially if their exposure to Spanish is limited and/or they are not familiar with the specific verb where the suffix is featured ([Montrul and Mason, 2020](#)). [Mason \(2019\)](#) attributes this trend to differences in type frequency across past participles, whereby regular forms present more productive formation patterns (i.e., the use of *-ado/-ido*) than irregular forms. Pattern productivity alone, however, does not explain the gradience of outcomes observed in language acquisition, where overregularizations appear to be resolved in some verbs earlier than in others (i.e., *dicho* ('said') vs. *resuelto* ('resolved'), as documented by [Galaz et al. \(2008\)](#)). In this case, token frequency could help determine which particular irregular forms HSs might be more likely to regularize in an innovative way.

The effects of (type and/or token) frequency may also be examined by controlling their presence through careful study design and stimuli selection. This is precisely the strategy we adopted in our ongoing work on subjunctive mood among Spanish HSs in the US (see [Giancaspro et al., 2022](#) for more information). The objective of this project was to revisit the study of a popular area of research among HSs of Spanish (i.e., subjunctive mood) by taking into account the effect of variables that had not been systematically controlled for in the past, such as the morphological regularity of the subjunctive forms tested, their type, and token frequency, and the modality of the proposition where the subjunctive forms were expected to appear. In contrast with previous research, we decided to control for the type frequency of the forms under analysis, limiting our selection of irregular verbs to those featuring a velar insert in their third person singular subjunctive inflections—instead of including forms with other types of irregularities, such as vocalic changes.

By controlling for type frequency, we were able to sidestep a key, potentially confounding variable and examine the effects of token frequency more directly, which gave rise to important differences across irregular verbs. The results in [Table 2](#) indicate that while all high frequency irregular forms were similarly likely to elicit subjunctive mood from HSs, lower frequency irregular verbs elicited much more variability from HSs, as evidenced by both (a) HSs' lower predicted probabilities of subjunctive production and (b) the wider confidence interval ranges for those predicted probabilities⁷. One clear exception to this pattern occurs in the case of *retenga* ('to retain'), which, though infrequent, still elicits a very high rate of subjunctive production, a finding that opens up new areas of inquiry concerning word compositionality

⁷ As noted by a reviewer, it is also possible that the width of the CI ranges was—in part—a result of log-odds mapping onto probabilities non-linearly, which could have contributed to the widening reported.

TABLE 2 Predicted probability of subjunctive use as a function of token frequency.

	Form elicited [token frequency ^a]	Predicted probability	95% CI	
			Lower	Upper
More Frequent	Tenga [712,671]	0.96	0.90	0.99
	Salga [119,654]	0.97	0.93	0.99
	Ponga [111,607]	0.99	0.98	1.00
	Venga [91,453]	0.96	0.90	0.99
	Traiga [13,048]	0.98	0.96	0.99
Less frequent	Proponga [11,192]	0.93	0.83	0.97
	Convenga [9,780]	0.79	0.54	0.92
	Retenga [1,912]	0.97	0.91	0.98
	Extraiga [1,345]	0.93	0.84	0.97
	Sobresalga [959]	0.72	0.48	0.88

^aToken frequencies were extracted from Davies' NOW corpus (2012–2016).

and the opacity/transparency of seemingly compound verbs. (Out of the irregular verbs tested, *tenga* is, by far, the most frequent, possibly making it easier for HSs to access and retrieve closely related subjunctive mood inflections such as *retenga*.) Overall, these findings not only provide additional insight on the effects of token frequency on subjunctive use, they also present a more nuanced description of morphological irregularity, which has generally been presented as a uniform, somewhat monolithic category.

Although less common, lexical frequency can also be examined from a lemmatic perspective, that is, considering the effects of all the inflectional variants of a particular form (verbal or nominal) all of which are represented with a single lemma⁸. The lemma *CUT*, for instance, includes all possible forms of this verb, such as {*cutting*, *cut*, *cuts*...}, as well as the word's nominal variants {*cut/cuts*}. Choosing to analyze the effects of verbal or nominal lexical frequency from a lemmatic perspective carries theoretical implications regarding how words are represented and accessed in the lexicon. In particular, it is assumed that the (cumulative) frequency of the paradigm will affect the lexical strength of individual—morphologically related—forms, making them more/less recognizable and likely to be retrieved. Adopting lemmatic frequency might be suitable for research where individual word differences are not central (i.e., measuring the extent to which the effects of (lemmatic) lexical frequency modulate the complexity of a text). Recent work dedicated to the study of frequency effects on morphological families, however, reports that the frequency of individual forms is more likely to predict variability in production, even if form similarity (between members of the same paradigm,

for example) might play a role in how related items compete with each other (Bybee, 2002; Kapatsinski, 2010). These findings suggest that the adoption of lemmatic frequency might not be fitting if the focus of the study is on the development and acquisition of particular forms, where their individual token frequency—rather than the frequency of their complete inflectional paradigm—is relevant for the analysis. Let us imagine, for example, that we were interested in examining whether lexical frequency modulates the interpretation and use of different types of future (periphrastic vs. morphological) among US HSs of Spanish. In principle, frequency could be analyzed in two different ways: a) including information about the token frequencies of each type of future (i.e., *comprará* ('(he/she) will buy') [7936] vs. *va a comprar* ('(he/she) will buy') [3754]), or b) reporting the frequency of the complete paradigm in the form of lemmatic frequency (*COMPRAR* [509875]). If we include the token frequency of all the forms involved, we would be able to explore whether (and how) the individual frequency of each inflection could affect HSs' performance. The use of lemmatic frequency, in contrast, would limit our analysis to general frequency effects, allowing us to gauge the extent to which the frequency of a particular verb, regardless of its inflection, might drive HSs' preference for one type of future over another. This broader perspective on frequency would allow us to capture verb-general effects, e.g., that Spanish speakers tend to use one type of future more with verbs that are collectively more frequent—that is to say, when all of its paradigms are collapsed together.

The previous discussion highlights the potential contributions of lexical frequency (in its different instantiations) to HL research, underscoring how individual speakers' experience with a HL might shape their grammatical knowledge and use. In the next section, we provide a summary of recent investigations that have examined the effects of lexical frequency on HSs' morphosyntactic development of the HL. After summarizing these studies, we present novel evidence that HSs' subjective assessment of lexical frequency more effectively predicts their patterns of subjunctive mood production than corpus-based frequency metrics.

Experimental approaches to lexical frequency effects in HL grammars

Putnam and Sánchez's (2013) predictions regarding lexical frequency sparked a renewed interest in the study of how this variable might modulate HSs' performance (Hur, 2020; Hur et al., 2020; Karayayla, 2021; López-Beltrán, 2021; Giancaspro et al., 2022; Perez-Cortes, 2022b, *inter alia*). As previously mentioned, the majority of the research in this area has focused on the domain of morphosyntax, with a particular emphasis on the acquisition of nominal and verbal inflection. Rather than manipulating it, some studies (Gor, 2019; López-Beltrán, 2021) have used token frequencies as a way to control HS participants' familiarity with a particular selection of lexical items. Thus, instead of including

⁸ Our operationalization of lemma differs from the one traditionally used in the production literature (i.e., Jescheniak and Levelt, 1994), aligning with the definition in the work of Gries (2009) and Knowles and Don (2004) instead).

verbs that are less frequent in the input, which are likely to have been less activated—and, as a result, more likely to exhibit increased variability—, these investigations only included highly frequent forms in the input, thereby giving participants the best chance to exhibit their HL knowledge. In the case of López-Beltrán's (López-Beltrán, 2021) auditory pupillometry study, verb selection was made based on data compiled from the *Corpus Sociolingüístico de la Ciudad de México* (CSCM; Martín Butragueño and Lastra, 2011). Specifically, the researcher ensured that the frequency range of all subjunctive-triggering governors included in this receptive task (i.e., *Deseo que* 'I wish that' or *Quiero que* 'I want that') was between 1 and 72 per 400,000 words. Additionally, the number of sentences that featured each governor was made proportional to its frequency, meaning that frequent triggers in the corpus appeared proportionally more often in the experimental task. Gor (2019) adopted a similar strategy in her study on the morphosyntactic knowledge of L2 learners and HSs of Russian. In particular, the investigator limited the vocabulary used in her grammaticality judgment task to words that appeared frequently in Russian language textbooks and that were also among the one thousand most frequent words in the Russian National Corpus.

The experimental designs adopted by Gor (2019) and López-Beltrán (2021) highlight the need to include stimuli that adequately represent the experience participants have with language. This is particularly relevant in the case of HSs, who may be more familiar with registers, styles or subsets of the lexicon that are not usually represented in traditional corpora. Karayayla (2021) addressed this particular question in her study of adult Turkish HSs' use of inflectional suffix templates and the level of sophistication of the morphological forms they produce (when compared to Turkish monolinguals and recent immigrants). Based on previous work by Durrant (2013), Karayayla suggests that it is imperative to use frequency data that captures the characteristics of the input experienced by heritage bilinguals to reproduce as closely as possible their patterns of exposure. Accordingly, all type and lemmatic frequencies of the words and suffixes that appeared in her study were based on a corpus that included (informal) oral language that is spoken around UK-born HSs of Turkish. Information about the type frequency of the suffixes represented in the corpus was implemented to ensure that only those that were more productive would appear in the stimuli. Results from this study indicated that HSs exhibited lower nominal productivity than other groups, which translated into the application of nominal suffixes to a reduced—and primarily, high frequency—subset of Turkish nouns.

Lexical frequency, in particular token frequency, can also be manipulated to determine the extent to which it affects HSs' ability to abstract grammatical knowledge from the input and generalize it across a wide range of lexical items. Perez-Cortes (2022b) sought to explore previously reported patterns of intraspeaker variability by focusing on the effects of token frequency on HSs' preference and use of subjunctive in predicates that allow for variable mood selection. Participants in the study

were a group of 35 intermediate-proficiency HSs of Spanish, who are among the most notoriously variable groups in HL research (Perez-Cortes et al., 2019). In two tasks (truth-value judgment and elicited production), Perez-Cortes tested two matrix verbs—*decir* ('to say') and *repetir* ('to repeat')—that represented both ends of the frequency spectrum, as seen in the contrasts illustrated in Table 3.

Results from a mixed-effects binary logistic regression indicated that Spanish HSs were more likely to interpret embedded clauses featuring subjunctive mood as commands, as would be expected in "baseline" Spanish, when the matrix verb introducing them was higher frequency ($M = 0.65$) rather than lower frequency ($M = 0.51$). Even though the type of matrix verb did not significantly affect HSs' performance in a separate production task, a descriptive analysis of the data indicated that their probability of using subjunctive in jussive (indirect command) contexts was higher when the matrix verb was frequent ($M = 0.64$) than when it was not ($M = 0.54$). Token frequency has also been shown to affect intermediate Spanish HSs' likelihood of using DOM in the expression of animate direct objects. In particular, Hur (2020) found that this group of bilinguals was more likely to favor the use of DOM with telic verbs that were more frequent ($M = 0.21$), such as *cuidar* ('to take care of' [7531]) than with less frequent ones ($M = 0.07$), such as *acariciar* ('to pet' [427]). Crucially, this pattern was not replicated among advanced-proficiency HSs, suggesting that as experience/proficiency with the HL grows, so does HSs' ability to employ grammatical morphemes across a wider range of lexical items.

In line with the suggestions documented in Karayayla (2021), several studies have moved towards a more ecologically-valid approach of obtaining lexical frequency data, putting speakers' individual experience with the HL at the forefront. (As Uygun & Clahsen (2021: 424) note, "frequencies for lexical entries may be highly variable for heritage speakers given their individual linguistic experience.") Hur et al. (2020), in their investigation of the effects of token frequency on gender assignment and agreement in heritage Spanish, implemented a self-rating lexical frequency task (SRLFT)—adapted from López Otero (2020)—with this particular purpose in mind. In the SRLFT, HSs reported their use of and exposure to the 32 lexical items included in the subsequent elicited production and forced-choice tasks. Participants were asked how often they heard and used the items under examination using a 9-point Likert scale (1 = never, 2 = hardly ever, 3 = a few times a year, 4 = once a month, 5 = a few

TABLE 3 Frequency values adapted from Perez-Cortes (2022: 158).

Matrix verb	Context	Token frequency (Davies NOW corpus)
Decir	Assertive (que + indicative)	76,962
	Jussive (que + subjunctive)	3,362
Repetir	Assertive (que + indicative)	237
	Jussive (que + subjunctive)	18

times a month, 6 = once a week, 7 = several times a week, 8 = once a day, 9 = several times a day), which resulted into a composite score for each lexical item that ranged from 2 to 18 (see Hur et al., 2020).

Results from a generalized linear mixed model including HSs' responses across tasks revealed that lexical frequency—as measured by the SRLFT described above—facilitated gender assignment and agreement. In general, items that were deemed more frequent by participants favored the expected gender assignment and agreement, while those that were less frequently used and heard exhibited more variability.

The studies summarized thus far obtained (token) frequency information in two distinct ways: through language corpora or participant self-reports. To explore whether (and how) the adoption of these measures could affect how we conceptualize the effects of frequency in HSs' performance, we reanalyzed data from Giancaspro et al. (2022)'s study of Spanish HSs' use of subjunctive mood in desiderative constructions (i.e., *Maria quiere que salgas pronto* 'Maria wants you to leave [3psgSUBJ] early'). Using the Davies NOW Corpus (2012–2016), we collected the token frequency of all the subjunctive verbs used in our production task ($N = 20$), which included items along a wide frequency spectrum: from highly frequent forms (i.e., *tenga* 'have') to very infrequent ones (i.e., *sobresalga* 'exceed'). Self-rated frequency was also examined using the results of a Lexical Experience Survey, which assessed participants' ($N = 42$) use of and exposure to all experimental verbs using a four-point frequency scale where 1 meant that participants 'never' used a verb and 4 meant that they used that verb 'very frequently'.

The data reported in Table 4 reveal a series of interesting observations. First, verbs at the low end of the (token) frequency spectrum based on their occurrence in the Davies' NOW corpus, such as *parta* ('split/cut') [6,470] or *traiga* ('bring') [13,048] appeared to be rather frequent for HSs in both use (*parta*: 3.59/4; *traiga*: 3.91/4) and exposure (*parta*: 3.66/4; *traiga*: 3.92/4), perhaps because these forms may be more common in the household setting. Participants' ratings also provided information about asymmetries in exposure and use that simply cannot be captured by traditional corpus data. Verbs like *proponga* ('to propose'), *convenga* ('to convene') and *ceda* ('to yield') are good examples of this: in each case, HSs' average exposure [range: 2.21–3.27] easily exceeds their self-reported use [range: 1.65–2.83].

Differences between these ways of capturing lexical frequency also emerged when we examined their statistical effects on HSs' performance. To do so, we ran three separate binary logistic regression models—each with a different fixed factor (participants' self-reported use (#1), exposure (#2) or items' token frequency based on corpus data (#3))—and with subjunctive use, dummy-coded as 1 for subjunctive and 0 for indicative, as their dependent variable. In all cases, the best fitting models that converged included random slopes for Participant, as well as random

TABLE 4 Corpus-based and self-rated token frequency (exposure and use) data.

Lexical item	Token frequency (Davies corpus)	Frequency of use (HS' average out of 4) ^a	Frequency of exposure (HS' average out of 4)
Tenga	712,671	3.85	3.85
Salga	119,654	3.98	4.00
Ponga	111,607	3.98	3.94
Venga	91,453	4.00	3.96
Traiga	13,048	3.91	3.92
Proponga	11,192	2.83	3.27
Convenga	9,780	1.67	2.21
Retenga	1,912	1.98	2.50
Extraiga	1,345	2.46	2.87
Sobresalga	959	2.30	2.67
Meta	21,597	3.76	3.80
Corra	10,993	3.63	3.71
Viva	27,637	3.93	3.88
Ceda	6,228	1.65	2.25
Parta	6,470	3.59	3.66
Prometa	1,366	3.30	3.46
Exceda	6,528	2.33	2.63
Comparta	56,158	3.78	3.85
Recorra	3,215	2.28	2.77
Sobreviva	3,800	2.91	3.25

^aSince our study tested HSs' knowledge and use of both indicative and subjunctive mood, participants' self-ratings were based on the stimuli's lemmas rather than their inflected indicative/subjunctive forms.

intercepts for Item⁹. Results from these regressions revealed that while participants' self-reported use ($\beta = 0.837$, $SE = 0.1538$, $t = 5.444$, $p < 0.001$) and exposure ($\beta = 1.181$, $SE = 0.2004$, $t = 5.894$, $p < 0.001$) were statistically significant predictors of their subjunctive use, token frequency based on the Davies' corpus was not ($\beta = 0.0054$, $SE = 0.3187$, $t = 1.739$, $p = 0.082$). These findings suggest that relative to frequency metrics derived from large-scale corpora, self-reported frequency measures that reflect participants' lived linguistic experience more accurately predict their likelihood of producing variability/grammatical innovations. Figure 1 depicts how participants' self-reported use of the verbs in the study (Model #1) affected their production of subjunctive in expected subjunctive items¹⁰:

9 Best-fitting models were determined to be those with the lowest Akaike information criterion (AIC) values, in accordance with Heck et al. (2012). The first model (AIC = 3976.022), had an overall correct classification rate of 88.4%, the second one (AIC = 4243.107), had a classification rate of 88.3%, and the third one (AIC = 3884.784), with a classification rate of 86.9%.

10 Despite finding several asymmetries in participants' reported use and exposure to the verbs tested in our study, differences in how these two factors modulated their actual performance were minimal, hence our decision to only provide a graphic representation of one of them.

The results plotted in Figure 1 show that for most participants, the more frequently that they report using a lexical item, the more likely they are to produce it in the subjunctive ($OR = 2.31$; 95% CI [1.70, 3.12], $p < 0.001$). As observed in the graph, verbs that participants reported using rarely (2) or never (1)—marked by smaller-sized blue circles—usually yielded the lowest predicted probabilities of subjunctive production. In contrast, verbs that participants reported using somewhat frequently (3) or very frequently (4)—marked by larger-sized blue circles—were more likely to elicit subjunctive mood inflections. Interestingly, participants whose performance was categorical at both ends of the probability scale—almost 40% of the sample—were not as affected by frequency as those who exhibited more variability, as in the case of Participant 9, whose verb-by-verb data we highlight in Table 5 below.

As indicated in Table 5, this participant, whose overall predicted probability of using subjunctive mood averaged 40%, did not produce subjunctive with any verb that they reported using either infrequently or “never”. In fact, 50% (6/12) of this participant’s innovative, indicative responses occurred with verbs that were relatively unfamiliar to them, and according to the self-rating task.

The information summarized thus far suggests that the study of lexical frequency—whether it is at the level of type, token, or lemma frequency—grants researchers the opportunity to tap into patterns of *intra-speaker variability*. However, despite the relevant role exerted by lexical frequency on HSS’ morphosyntactic development, we agree with Ambridge et al. (2015) that this factor alone cannot explain variability on its own. As these researchers

note, lexical frequency—which in most cases is operationalized as the occurrence of individual tokens in the input—is likely to interact with other variables (i.e., regularity, phonological salience or semantic content) when modulating HL acquisition and maintenance, as reported in Giancaspro et al. (2022) and evident in Participant 9’s individual data. (Notably, Participant 9 uses much more subjunctive with irregular, as opposed to regular verbs.) The explanatory limitations of frequency, though, should not be seen as a disadvantage, especially given that potential interactions between frequency and other pertinent variables offer researchers multiple new avenues for better explaining HL variability.

Discussion and conclusion: some final thoughts

The purpose of this article was twofold: first, we sought to clarify what is meant by ‘frequency effects’ in the field of HL acquisition research. To do so, we provided clear operationalizations of frequency both from a language activation lens, as well as from a lexical perspective. After laying out this critical groundwork, we then illustrated how further exploration of these frequency subtypes will help to illuminate two long-standing, yet relatively less studied patterns: (i) between-speaker variability, that is to say, differences in the linguistic knowledge of different HSS and (ii) within-speaker variability, meaning variability in individual HSS’ knowledge of particular HL forms (e.g., subjunctive mood). A second goal of the article was to serve as a point of departure for HL researchers who are interested in

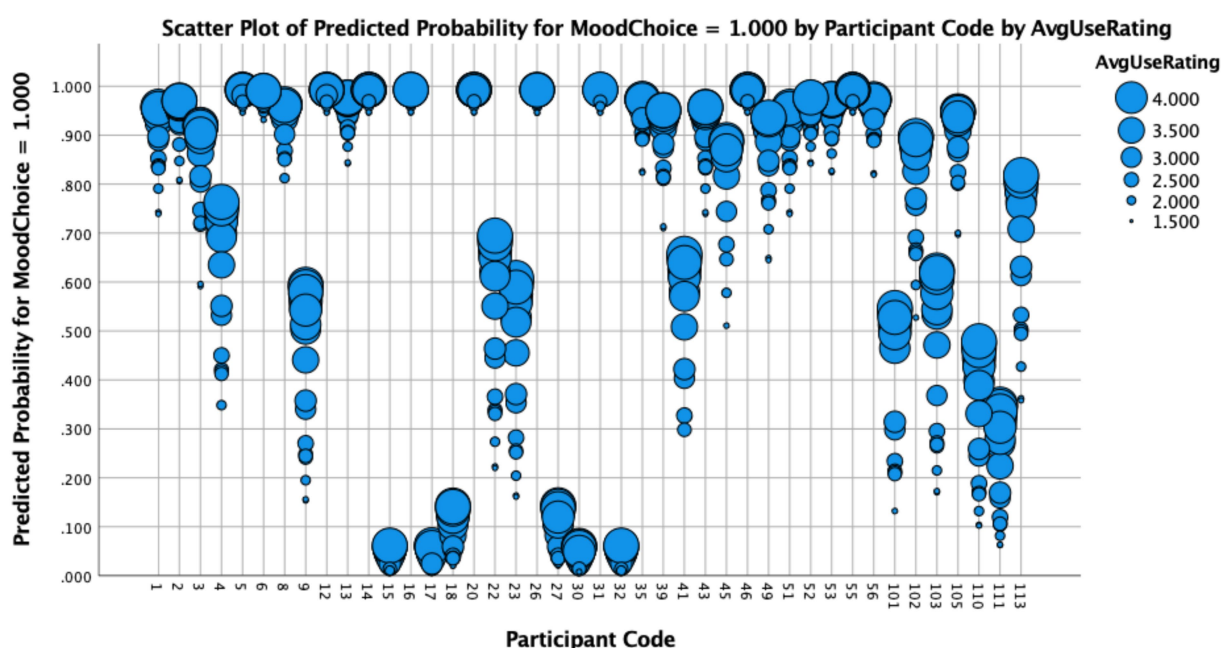


FIGURE 1
Individual participants’ predicted probability of subjunctive use as a function of self-reported average use.

TABLE 5 Participant 9 (advanced HS) individual results as a function of reported use.

Verb	Self-reported use (out of 4)	Use of subjunctive (0 or 1)
Tener	4	1
Retener	3	1
Venir	4	1
Convenir	3	1
Traer	4	1
Extraer	1	0 (indicative)
Poner	4	1
Proponer	2	0 (indicative)
Salir	4	1
Sobresalir	1	0 (indicative)
Meter	3	0 (indicative)
Prometer	4	0 (indicative)
Ceder	1	0 (indicative)
Exceder	2	0 (indicative)
Correr	4	0 (indicative)
Recurrer	2	0 (indicative)
Partir	3	0 (indicative)
Compartir	4	0 (indicative)
Vivir	4	0 (indicative)
Sobrevivir	3	1

The bolded rows indicate verbs that participants reported using infrequently.

examining frequency—from either one (or both) of the perspectives mentioned—in their future studies. To this end, we presented a critical analysis of some of the field’s most relevant and recent work on frequency effects in the HL, paying particular attention to what should be considered best practices from theoretical as well as empirical vantage points. Among the most novel contributions of this overview, we believe, is the finding that self-reported lexical frequency—that is to say, HSs’ own subjective assessment of how frequently they hear/use certain words—appears to be a better predictor of their subjunctive mood variability than traditional, corpus-derived frequency metrics.

Before going any further, we believe that a couple of key clarifications are in order. First, while the present paper has prioritized the discussion of between-speaker and within-speaker comparative analyses, it is not our intention to dismiss the importance of more commonly studied contrasts—namely, between-group and between-property comparisons—in the study of heritage bilingualism. In fact, as we noted in the “Introduction,” the vast majority of the foundational work in our field has emerged from those two lines of inquiry, a reality which should not be overlooked. Our claim, instead, is that different comparative vantage points—including those that we have showcased in this paper—have different epistemological blind spots, meaning, essentially, that in order to appreciate the immense complexity of HL grammars, we must look at them from a more diverse variety of viewpoints. Just like between-group comparisons—e.g., comparing HSs to a baseline/control group—cannot shed light on why individual HSs might alternately produce two variants of a

single form in a single HL context, within-speaker comparisons—like the analyses of lexical frequency effects presented in “Lexical frequency and its role in heritage grammars”—cannot explain why some HL properties (e.g., mood morphology) appear to be more “vulnerable” for HSs than others (e.g., tense/aspect morphology)¹¹. Given the inherently complementary nature of between-group, within-speaker, and other perspectives on heritage bilingual knowledge, focusing (nearly) exclusively on one or two specific perspectives will necessarily lead to oversimplified understandings of HSs and the sophisticated linguistic systems that they develop and maintain. An even more concerning consequence of such epistemological uniformity, we believe, is that it could, if sufficiently conventionalized, make it increasingly difficult for researchers to even imagine other types of research questions whose answers might be needed in order to illuminate new paths forward for the field as a whole. Summarizing, then, it is our hope that the between-speaker and within-speaker comparisons that we promote in this paper both (a) complement, rather than replace, other types of comparisons, and (b) stimulate novel lines of inquiry, possibly (though not necessarily) related to the categories of frequency we discuss here.

While we recognize the enormous potential of the two varieties of frequency outlined in this paper, it is important to clarify, too, that neither is powerful enough to obviate other types of linguistic and non-linguistic explanations of HL grammars. In fact, as [Ambridge et al. \(2015\)](#) note, “a frequency effect can never be an explanation or answer in its own right” (p. 248), a point to which we will return later in this section. That said, if frequency is not—and cannot be—an explanation, why should researchers invest the time to address it carefully in their HL grammatical work? Do not we already have enough to worry about without diving into the frequency deep end?

One reason to embrace frequency is that frequency-effects—broadly conceived—appear to be an empirical reality of HL grammars. At the between-speaker level, differences in HSs’ frequency of experience with the HL seem to result in differences in the HL grammars that they ultimately develop. Recall, to recap an example from “Between-speaker comparisons: frequency of heritage language activation,” that the simultaneous HSs in [Montrul and Sánchez-Walker \(2013\)](#), who produced DOM in Spanish at rates ranging from 0% to 100%, were less likely to omit DOM if they used Spanish more frequently. Much more research is needed in this area—especially, work that builds patterns of HL use into statistical modeling—but the early returns, so to speak, certainly suggest that HSs who use the HL more often are more likely to develop generalized—rather than item-by-item—knowledge about HL grammatical properties, such as DOM or subjunctive mood. Relatedly, at the within-speaker level, it appears to be the case that HSs often develop “item-based” lexically-specific sensitivity to HL grammatical properties, that is,

¹¹ For two mood-related examples of the between-property comparative vantage point needed for that type of question, see [Montrul, 2009](#) and [van Osch and Sleeman, 2018](#).

knowledge of certain morphemes/structures that only applies to specific subsets of the HL lexicon (e.g., gender with frequent nouns; mood with irregular verbs...etc....) rather than to the HL lexicon in its entirety. To the extent that we can agree on the existence of these patterns—and the evidence, from our view, seems undeniable—posing (in)frequently asked questions about frequency in HL grammars is a necessary step in the field's quest to understand HL grammatical systems as they are, and not just as they fit into our models.

There's another reason to pursue frequency-based analyses in heritage bilingualism research. Though frequency is not, to reiterate, an explanation itself, investigating it and identifying some of its previously undiscovered effects can open the door to a number of novel analyses and research questions, many of which have the potential to reverberate far beyond HL research itself. As Ambridge et al. (2015) point out, when a so-called frequency effect is identified, it does not provide answers as much as it “poses a question: What type of learning mechanism is needed to yield the *particular type* of frequency effect observed?” Therefore, when Perez-Cortes (2022) documents token frequency effects on HSs' interpretation and use of subjunctive mood or Mason (2019) finds that type frequency modulates HSs' knowledge and use of present perfect and preterit forms, what might these specific patterns reveal about how HSs go about building (and maintaining) abstract grammatical knowledge? It is still very early, of course, but we suspect that facing—and then interrogating—these common HL patterns will challenge some of the binary conceptualizations that have thus far dominated not just HL acquisition research but also much of linguistic theory.

For reasons of space, we will conclude this paper by presenting two brief—and hopefully, inspirational—examples of how reflecting on—and taking into consideration—frequency effects, broadly defined, could deepen our understanding of HL grammatical complexity and actually improve existing explanations of widespread HL patterns and phenomena. A substantial proportion of research on HSs has focused on what they do not know and how they diverge from so-called baseline speakers (see Polinsky, 2018 or Montrul, 2016 for an overview). These between-group differences are undeniable, if not inevitable (Polinsky, 2016), yet, considering how little attention has been dedicated to controlling for (or manipulating) lexical frequency in experimental research on HSs, one wonders if the differences between-groups—which have formed the foundation of HL theories and models—may have been inadvertently inflated by the inclusion of infrequent (and/or high register) lexical items that are peripheral to HSs' own linguistic life experiences¹². Recent work, as highlighted

in “Lexical frequency and its role in heritage grammars,” has started to address this oversight by considering lexical frequency when creating experimental items that are drawn from HS-specific corpora—and other sources—that more directly reflect participants' linguistic experiences with the HL. A perfect example can be found in López-Beltrán's innovative (López-Beltrán, 2021) study, where stimuli only consisted of highly frequent forms that were representative of HSs' input in the HL. This methodological change had direct consequences in the results obtained, as HSs who participated in her tasks exhibited clearer sensitivity to subjunctive mood morphology than HSs in previously reported studies. This finding has the potential to serve as a methodological rebuke to so-called deficit perspectives on HL acquisition. If researchers test HSs on frequent items that form a key part of their HL experience, perhaps many of the HS vs. baseline differences will greatly diminish or even disappear altogether.

We have seen, thus far, that being more intentional about lexical frequency might help us to gain a more reliable representation of what HSs really know about their HL. On a similar note, we believe that lexical frequency might also help us to understand the nature of promising—yet still relatively underexplored—explanations of between-speaker differences, such as HL literacy/formal education. To illustrate this final point, let us reflect on Bayram et al.'s (2017) work on passives in heritage Turkish. Summarizing briefly, Bayram et al. found that adolescent HSs were more likely to produce passive structures in their HL if they were more literate in Turkish. At one level, this finding constitutes an explanation of why some HSs appear to exhibit different knowledge than others. (This is a great example, in fact, of the type of between-speaker analysis that we hope to see more of in the field.) At another level, however, the finding that literacy drives between-speaker differences in passive production only raises another series of deeper, and perhaps more revealing questions, whose answers may be at least partially addressed by looking at lexical frequency. Are the more literate HSs in Bayram et al. (2017) more likely to use Turkish passives in general or only with higher register/lower frequency subsets of the Turkish lexicon, which they might be more likely to encounter in educational/written sources and settings? In asking this question, and we believe that other, similar questions can be asked of many other impactful studies in the field, we can better pinpoint the specific grammatical muscles that are strengthened by additional, formal HL experience, a finding that would have both theoretical and classroom implications.

In any case, we do not wish to belabor the point, but frequency-based analyses, in our view, raise interesting—even

¹² This seems especially likely to happen, we believe, in cases where researchers' own education in the target language far exceeds that of their HS participants. If we aren't intentional in the selection of lexical items for our experiments—and we should note here that very few studies in HL verbal morphology comment on the verbs that are chosen for experimental

tasks—it seems likely that lexical items that are infrequent (either in terms of their frequency in a corpus or, more importantly, from the perspective of individual HS participants themselves) will be oversampled in our tasks.

stimulating—questions and broaden our perspective of heritage grammars and their speakers. In a field as relatively young as HL acquisition research, pursuing new empirical questions and charting new methodological paths can only be a positive development, especially if those new directions, in acknowledging new layers of complexity, push us to more deeply reflect on the near ubiquitous (yet still understudied) patterns of between-speaker and within-speaker variability.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We would like to thank the three reviewers who provided thoughtful comments and suggestions in the review process.

References

- Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *J. Child Lang.* 42, 239–273. doi: 10.1017/S030500091400049X
- Backus, A. (2020). “Usage-based approaches” in *The Routledge Handbook of Language Contact*. eds. E. Adamou and Y. Matras (New York, NY: Routledge), 110–126. doi: 10.4324/9781351109154-8
- Bayram, F., Rothman, J., Iverson, M., Kupisch, T., Miller, D., Puig-Mayenco, E., et al. (2017). Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *Int. J. Biling. Educ. Biling.* 22, 919–939. doi: 10.1080/13670050.2017.1324403
- Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins Publishing.
- Bybee, J. (1995). Regular morphology and the lexicon. *Lang. Cogn. Process.* 10, 425–455. doi: 10.1080/01690969508407111
- Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Lang. Var. Chang.* 14, 261–290. doi: 10.1017/S0954394502143018
- Bybee, J. (2006). From usage to grammar: the mind's response to repetition. *Language* 82, 711–733. doi: 10.1353/lan.2006.0186
- Bybee, J. (2007). *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.
- Bybee, J., and Thompson, S. (2000). Three frequency effects in syntax. *Berkeley Linguistics Society* 23, 378–388. doi: 10.3765/bls.v23i1.1293
- Camacho, J. (2022). Paradigmatic uniformity: evidence from heritage speakers of Spanish. *Languages* 7:14. doi: 10.3390/languages7010014
- Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cogn. Neuropsychol.* 14, 177–208. doi: 10.1080/026432997381664
- Clahsen, H., Avello, F., and Roca, I. (2002). The development of regular and irregular verb inflection in Spanish child language. *J. Child Lang.* 29, 591–622. doi: 10.1017/S0305000902005172
- Cuza, A. (2016). The status of interrogative subject-verb inversion in Spanish-English bilingual children. *Lingua* 180, 124–138. doi: 10.1016/j.lingua.2016.04.007
- Daskalaki, E., Chondrogianni, V., Blom, E., Argyri, F., and Paradis, J. (2019). Input effects across domains: the case of Greek subjects in child heritage language. *Second. Lang. Res.* 35, 421–445. doi: 10.1177/0267658318787231
- Davies' Spanish, NOW corpus (2012–2016). Corpus of news on the web (NOW). Available at: <https://www.corpusdelespanol.org/nw/> (Accessed July 20, 2022).
- DePaolis, R. A., Vihman, M. M., and Keren-Portnoy, T. (2011). Do production patterns influence the processing of speech in prelinguistic infants? *Infant Behav. Dev.* 34, 590–601. doi: 10.1016/j.infbeh.2011.06.005
- Diessel, H., and Hilpert, M. (2016). Frequency effects in grammar. in *Oxford Research Encyclopedia of Linguistics*.
- Dracos, M., and Requena, P. (2022). Child heritage speakers' acquisition of the Spanish subjunctive in volitional and adverbial clauses. *Lang. Acquis.*, 1–28. doi: 10.1080/10489223.2022.2071156
- Durrant, P. (2013). Formulaicity in an agglutinating language: the case of Turkish. *Corpus Linguist. Linguist. Theory* 9, 1–38. doi: 10.1515/cllt-2013-0009
- Ellis, N. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Stud. Second. Lang. Acquis.* 24, 143–188. doi: 10.1017/S0272263102002024
- Embick, D. (2015). *The Morpheme: A Theoretical Introduction*. Berlin, Germany: Mouton de Gruyter.
- Flores, C. (2015). Understanding heritage language acquisition: some contributions from the research on heritage speakers of European Portuguese. *Lingua* 164, 251–265. doi: 10.1016/j.lingua.2014.09.008
- Galaz, X., Norambuena Muñoz, C., and Rivera Lazo, M. (2008). *Errores de sobreregularización frecuentes en niños de entre tres y cinco años de edad*. *Cyber Humanitatis*, 45. (Santiago, Chile: Universidad de Santiago de Chile).
- Giancaspro, D. (2019). Over, under and around: Spanish heritage speakers' production (and avoidance) of subjunctive mood. *Herit. Lang. J.* 16, 44–70. doi: 10.46538/hlj.16.1.3
- Giancaspro, D. (2020). “Not in the mood: Frequency effects in heritage speakers' knowledge of subjunctive mood,” in *Lost in Transmission: The Role of Attrition and Input in Heritage Language Development* eds. B. Brehmer and J. Treffers-Daller (Amsterdam: John Benjamins), 72–97.
- Giancaspro, D., Perez-Cortes, S., and Higdon, J. (2022). (Ir) regular mood swings: lexical variability in heritage speakers' oral production of subjunctive mood. *Lang. Learn.* 72, 456–496. doi: 10.1111/lang.12489
- Gonzalez, B. H. (2020). The syntactic distribution of object experiencer psych verbs in heritage Spanish. *Languages* 5:63. doi: 10.3390/languages5040063
- Gor, K. (2019). Morphosyntactic knowledge in late second language learners and heritage speakers of Russian. *Herit. Lang. J.* 16, 124–150. doi: 10.46538/hlj.16.2.2
- Gries, S. T. (2009). What is corpus linguistics? *Lang. Linguist. Compass* 3, 1225–1241. doi: 10.1111/j.1749-818x.2009.00149.x
- Heck, R. H., Thomas, S. L., and Tabata, L. N. (2012). *Multilevel and Longitudinal Modeling with IBM SPSS*. New York, NY: Routledge.
- Hopper, P. J., and Bybee, J. (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam, Netherlands: John Benjamins Publishing.
- Hur, E. (2020). “Verbal lexical frequency and DOM in heritage speakers of Spanish” in *The Acquisition of Differential Object Marking*. eds. A. Mardale and S. Montrul (Amsterdam and Philadelphia: John Benjamins), 207–235.

We would also like to thank Josh Higdon, who contributed greatly to the initial Giancaspro et al. (2022) study, as well as our participants, without whom this study would not be possible.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Hur, E., Lopez Otero, J. C., and Sanchez, L. (2020). Gender agreement and assignment in Spanish heritage speakers: does frequency matter? *Languages* 5:48. doi: 10.3390/languages5040048
- Jeschaniak, J. D., and Levelt, W. J. (1994). Word frequency effects in speech production: retrieval of syntactic information and of phonological form. *J. Exp. Psychol. Learn. Mem. Cogn.* 20, 824–843. doi: 10.1037/0278-7393.20.4.824
- Kapatsinski, V. (2010). What is it I am writing? Lexical frequency effects in spelling Russian prefixes: uncertainty and competition in an apparently regular system. *Corpus Linguist. Linguist. Theory* 6, 157–215. doi: 10.1515/clt.2010.007
- Kapatsinski, V., Easterday, S., and Bybee, J. (2020). Vowel reduction: a usage-based perspective. *Rivista di Linguistica* 32, 19–44. doi: 10.26346/1120-2726-146
- Karayayla, T. (2021). A usage-based approach to productive use of inflectional patterns and level of lemma sophistication in adult heritage speakers' performance: convergence on the immigrant variety. *Linguist. Approaches Biling.* 11, 753–782. doi: 10.1075/lab.18019.kar
- Kascelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., and De Cat, C. (2022). A review of questionnaires quantifying bilingual experience in children: do they document the same constructs? *Biling. Lang. Cogn.* 25, 29–41. doi: 10.1017/S1366728921000390
- Knowles, G., and Don, Z. M. (2004). The notion of a “lemma”: headwords, roots and lexical sets. *Int. J. Corpus Linguist.* 9, 69–81. doi: 10.1075/ijcl.9.1.04kno
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Lohndal, T., and Putnam, M. T. (2021). The tale of two lexicon: decomposing complexity across a distributed lexicon. *Herit. Lang. J.* 18, 1–29.
- López Otero, J. C. (2020). The acquisition of the syntactic and morphological properties of Spanish imperatives in heritage and second language speakers. Doctoral dissertation. New Brunswick, NJ: Rutgers the State University of New Jersey.
- López Otero, J. C., Cuza, A., and Jiao, J. (2021). Object clitic use and intuition in the Spanish of heritage speakers from Brazil. *Second. Lang. Res.* 026765832110176. doi: 10.1177/02676583211017603
- López-Beltrán, P. (2021). Heritage speakers' online processing of the Spanish subjunctive: a comprehensive usage-based study. Doctoral dissertation. College Park, PA: Penn State University.
- López-Beltrán, P., and Carlson, M. (2020). How usage-based approaches to language can contribute to a unified theory of heritage grammars. *Linguist. Vanguard* 6. doi: 10.1515/lingvan-2019-0072
- Martín Butragueño, P., and Lastra, Y. (2011). *Corpus sociolingüístico de la ciudad de México*. México: El Colegio de México.
- Mason, S. A. (2019). *The influence of task factors and language background on morphological processing in Spanish*. Doctoral dissertation. Urbana-Champaign, IL: University of Illinois at Urbana-Champaign.
- Montrul, S. (2009). Knowledge of tense-aspect and mood in Spanish heritage speakers. *Int. J. Biling.* 13, 239–269. doi: 10.1177/1367006909339816
- Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge, United Kingdom: Cambridge University Press.
- Montrul, S. (2021a). “Morphology in Spanish heritage language grammars” in *The Routledge Handbook of Spanish Morphology*. eds. A. Fábregas, V. Acedo-Matellán, G. Armstrong, M. C. Cuervo and I. P. Payet (New York, NY: Routledge), 538–549.
- Montrul, S. (2021b). Representational and computational changes in heritage language grammars. *Herit. Lang. J.* 18, 1–30. doi: 10.1163/15507076-12340011
- Montrul, S., Davidson, J., De La Fuente, I., and Foote, R. (2014). Early language experience facilitates the processing of gender agreement in Spanish heritage speakers. *Biling. Lang. Cogn.* 17, 118–138. doi: 10.1017/s1366728913000114
- Montrul, S., and Mason, S. A. (2020). Smaller vocabularies lead to morphological overregularization in heritage language grammars. *Biling. Lang. Cogn.* 23, 35–36. doi: 10.1017/s1366728919000427
- Montrul, S., and Sánchez-Walker, N. (2013). Differential object marking in child and adult Spanish heritage speakers. *Lang. Acquis.* 20, 109–132. doi: 10.1080/10489223.2013.766741
- O'Grady, W., Kwak, H. Y., Lee, O., and Lee, M. (2011). An emergentist perspective on heritage language acquisition. *Stud. Second. Lang. Acquis.* 33, 223–245. doi: 10.1017/S0272263110000744
- Perez-Cortes, S. (2022a). On complexity and divergence in heritage language grammars: the case of double mood selection in reported speech contexts. *Stud. Second. Lang. Acquis.* 44, 818–842. doi: 10.1017/s0272263121000589
- Perez-Cortes, S. (2022b). Lexical frequency and morphological regularity as sources of heritage speaker variability in the acquisition of mood. *Second. Lang. Res.* 38, 149–171. doi: 10.1177/0267658320918620
- Perez-Cortes, S., Putnam, M. T., and Sánchez, L. (2019). Differential access: asymmetries in accessing features and building representations in heritage language grammars. *Languages* 4:81. doi: 10.3390/languages4040081
- Polinsky, M. (2016). Structure vs. use in heritage language. *Linguist. Vanguard* 2. doi: 10.155/lingvan-2015-0036
- Polinsky, M. (2018). *Heritage languages and their speakers* (Cambridge, United Kingdom: Cambridge University Press).
- Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/s1366728919000245
- Poplack, S., Leal, A., and Dion, N. (2013). The evolving grammar of the French subjunctive. *Int. J. Latin Roman. Linguist.* 25, 139–195. doi: 10.1515/probus-2013-0005
- Putnam, M., Carlson, M., and Reitter, D. (2018). Integrated, not isolated: defining typological proximity in an integrated multilingual architecture. *Front. Psychol.* 8:2212. doi: 10.3389/fpsyg.2017.02212
- Putnam, M., and Sánchez, L. (2013). What's so incomplete about incomplete acquisition? A prolegomenon to modeling heritage language grammars. *Linguist. Approaches Biling.* 3, 478–508. doi: 10.1075/lab.3.4.04put
- Putnam, M., Schwarz, L., and Hoffman, A. D. (2022). “Morphology of heritage languages” in *The Cambridge Handbook of Heritage Languages and Linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 613–643.
- Schmid, M. S., and Köpke, B. (2017). The relevance of first language attrition to theories of bilingual development. *Linguist. Approaches Biling.* 7, 637–667. doi: 10.1075/lab.17058.sch
- Silva-Corvalán, C. (2018). Simultaneous bilingualism: early developments, incomplete later outcomes? *Int. J. Biling.* 22, 497–512. doi: 10.1177/1367006916652061
- Sorace, A., and Keller, F. (2005). Gradience in linguistic data. *Lingua* 115, 1497–1524. doi: 10.1016/j.lingua.2004.07.002
- Soto-Corominas, A. (2021). “Morphology and L1 acquisition” in *The Routledge Handbook of Spanish Morphology*. eds. A. Fábregas, V. A. Matellán, G. Armstrong, M. C. Cuervo and I. P. Payet (New York, NY: Routledge), 513–525.
- Torregrossa, J., Flores, C., and Rinke, E. (2022). What modulates the acquisition of difficult structures in a heritage language? A study of Portuguese in contact with French, German, and Italian. *Biling.* 1–14. doi: 10.1017/S1366728922000438
- Unsworth, S. (2013). Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: The case of Dutch gender. *Biling. Lang. Cogn.* 16, 86–110. doi: 10.1017/S1366728912000284
- Uygun, S., and Clahsen, H. (2021). Morphological processing in heritage speakers: a masked priming study on the Turkish aorist. *Biling. Lang. Cogn.* 24, 415–426. doi: 10.1017/S1366728920000577
- van Osch, B., and Sleeman, P. (2018). Spanish heritage speakers in the Netherlands: linguistic patterns in the judgment and production of mood. *Int. J. Biling.* 22, 513–529. doi: 10.1177/1367006916654365
- White, L., and Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second. Lang. Res.* 12, 233–265. doi: 10.1177/026765839601200301
- Yang, C. (2004). Universal grammar, statistics or both? *Trends Cogn. Sci.* 8, 451–456. doi: 10.1016/j.tics.2004.08.006
- Yang, C. (2015). For and against frequencies. *J. Child Lang.* 42, 287–293. doi: 10.1017/s0305000914000683
- Zyzik, E. (2016). “Toward a prototype model of the heritage language learner,” in *Innovative Strategies for Heritage Language Teaching: A Practical Guide for the Classroom*. eds. M. Fairclough and S. Beaudrie (Washington, DC: Georgetown University Press), 19–38.
- Zyzik, E. (2019). Incomplete acquisition from a usage-based perspective: a response to Domínguez, Hicks, and Slabakova. *Stud. Second. Lang. Acquis.* 41, 279–282. doi: 10.1017/s0272263119000330



OPEN ACCESS

EDITED BY

Maki Kubota,
UiT the Arctic University of Norway,
Norway

REVIEWED BY

Anamaria Bentea,
University of Konstanz,
Germany
Eliane Lorenz,
University of Giessen,
Germany

*CORRESPONDENCE

Ethan Kutlu
ethan-kutlu@uiowa.edu

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 01 September 2022

ACCEPTED 03 November 2022

PUBLISHED 24 November 2022

CITATION

Kutlu E, Chiu S and McMurray B (2022)
Moving away from deficiency models:
Gradiency in bilingual speech
categorization.
Front. Psychol. 13:1033825.
doi: 10.3389/fpsyg.2022.1033825

COPYRIGHT

© 2022 Kutlu, Chiu and McMurray. This is
an open-access article distributed under
the terms of the [Creative Commons
Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Moving away from deficiency models: Gradiency in bilingual speech categorization

Ethan Kutlu^{1,2*}, Samantha Chiu¹ and Bob McMurray^{1,2}

¹Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA, United States,

²Department of Linguistics, University of Iowa, Iowa City, IA, United States

For much of its history, categorical perception was treated as a foundational theory of speech perception, which suggested that quasi-discrete categorization was a goal of speech perception. This had a profound impact on bilingualism research which adopted similar tasks to use as measures of nativeness or native-like processing, implicitly assuming that any deviation from discreteness was a deficit. This is particularly problematic for listeners like heritage speakers whose language proficiency, both in their heritage language and their majority language, is questioned. However, we now know that in the monolingual listener, speech perception is gradient and listeners use this gradiency to adjust subphonetic details, recover from ambiguity, and aid learning and adaptation. This calls for new theoretical and methodological approaches to bilingualism. We present the Visual Analogue Scaling task which avoids the discrete and binary assumptions of categorical perception and can capture gradiency more precisely than other measures. Our goal is to provide bilingualism researchers new conceptual and empirical tools that can help examine speech categorization in different bilingual communities without the necessity of forcing their speech categorization into discrete units and without assuming a deficit model.

KEYWORDS

speech perception, gradiency, categorical perception, bilingualism, sound acquisition, heritage bilingualism

Introduction

Listeners encounter highly variable speech signals every day. Much of the research on speech perception has focused on understanding this problem of lack of invariance – how does a given listener categorize a highly variable acoustic signal into discrete units like features, phonemes or words to extract the linguistic information relevant for that utterance? For a long time, these issues were investigated under the umbrella of categorical perception (henceforth CP; Liberman et al., 1957). Theoretically, CP argues that perception—the pre-categorical auditory encoding—is warped by the presence of categories. One consequence of this is that during speech perception, listeners discard continuous acoustic information that is irrelevant to category identity and only perceive the category.

For example, voice onset time (VOT) is a continuous cue that distinguishes voiced and voiceless/aspirated consonants across many languages (Lisker and Abramson, 1964; Lisker, 1986; Abramson and Whalen, 2017). It is defined as the period of time between the release of a stop consonant and the onset of voicing. In English, voiced sounds have VOTs near 0 msec, and voiceless near 60 msec, though this varies cross linguistically. Even though VOT scales continuously, CP argued that English-speaking listeners are less capable of hearing the difference between VOTs of 40 and 50 msec (both of which indicate a voiceless sound) than the difference between 15 and 25 msec (which spans the boundary), despite the fact that each contrast has an equivalent physical distance.

CP led to two contributions that shaped subsequent work on multilingualism. The first was methodological: the extensive use of speech continua and forced choice tasks along with a set of theoretical assumptions about how to interpret them. The second was theoretical: CP led to an implicit view that a sort of quasi-discrete representation of speech was desirable and any deviation from that may represent a deficit. Importantly, this representational system emerges during the first year of life. This impact can be seen in two examples.

First, research on bilinguals has long known that adult L2 learners face challenges in acquiring the categories of their second language (Strange and Shafer, 2008). The question is why? Classic developmental work argued that speech categories are formed during the first year of life and that the emergence of these categories and their structure was associated with a sensitive period (Werker and Tees, 1984) (though see McMurray, 2022). If we assume CP as a model of speech perception, this can then explain adult learners: many new L2 distinctions comprise within-category distinctions in the native language (e.g., the English /l/r distinction which lies within the Japanese category). If listeners cannot hear these distinctions due to the effect of early experience, this can explain why L2 learning is so hard.

More broadly, the assumption of CP (and the methods) also served as a sort of linking hypothesis to understanding bilingual abilities. In particular, the forced-choice task has been extensively interpreted such that a steeper slope (i.e., categorical) reflects better perceptual encoding, and a shallower slope (i.e., gradient) reflects a deficiency in perceptual encoding. Consequently, even a slight deviance from monolingual-like performance led to discussions of whether bilinguals can form monolingual-like categories. That is the steep slope is considered ideal and any departure reflects a limitation.

However, this sort of simple framing may be inappropriate when we consider the wide variety of forms that bilingualism can take. In a heritage bilingual context, the first years of life might have more emphasis on the heritage language compared to the majority language. Nonetheless, their continued exposure to the majority language may overcome this background. Alternatively, the dynamics between the heritage language and the majority language might change depending on the bilingual context (e.g., code-switching vs. a more linguistically homogenous context).

In these cases, it may be more appropriate to evaluate cues like VOT gradiently across different contexts, rather than attempting to impose a single sharp (and inflexible) boundary.

While work on bilingualism has operated on the assumption of CP, research on monolingual adults has begun to move away from it [for a review (McMurray, n.d.)]. As we describe, this work suggests that adult listeners show robust sensitivity to within-category differences, and that speech categories may be highly gradient. In fact, unlike the claims made by categorical perception, this more recent work suggests that having a shallow slope (i.e., being gradient) is not an indicator of deficiency. On the contrary, it might be the marker of better information encoding. It also proposes new methods (and new ways of understanding existing measures) that may allow more sensitive ways to probe individual differences and are more aligned with this theoretical development.

The goal of this manuscript is to challenge the assumption of CP in bilingualism research, particularly in heritage bilingualism. We first describe the debates over CP in monolingual speech perception. We then focus on how assumptions of CP impacted bilingualism research and how it led to a deficiency model of bilingual speech perception. We will then introduce the Visual Analogue Scaling task (VAS task), which has been a trademark of our research group, to examine speech perception in developing children (both monolingual and bilingual) and adults (monolingual, bilingual, and cochlear implant users). We will present preliminary data from an ongoing experiment that exemplifies how the VAS task can capture profiles of gradient speech perception in bilinguals, and we introduce new statistical modeling that builds the notion of individual variability into the analysis. The ultimate goal of this manuscript is to move bilingualism research away from the theoretical assumptions produced by categorical perception and show how methodological reconsiderations are necessary to fully capture different bilingual profiles without the deficiency model.

Categorical perception in language science research

Historically, a large majority of speech perception research has focused on the problem of lack of invariance (Kluender, 1994; Liberman and Whalen, 2000; Perkell and Klatt, 2014). This problem arises from the fact that the same phoneme varies in its own acoustic manifestation depending on the speaker's speech rate, phonetic context, and many other variables. Moreover, the same bundle of acoustic cue values can be consistent across multiple phonemes. The problem then is how can a listener efficiently map a continuous acoustic signal onto a set of discrete units (e.g., phonemes) in the face of a non-invariant mapping between individual cue values and categories?

CP (Liberman et al., 1957) was central to early theoretical approaches to this problem. CP was initially an empirical phenomenon which was observed when listeners showed poor discrimination for two speech sounds that arose from the same

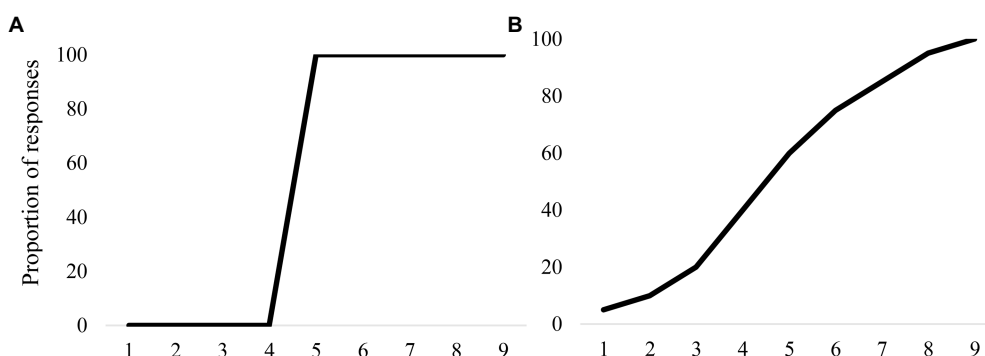


FIGURE 1
Proportion of responses in a typical 2AFC task where the black curve signifies the mean responses in both graphs. **(A)** A categorical profile which has a sharper slope. **(B)** A less categorical profile, with a shallower slope.

category, but good discrimination for tokens that span the boundary, even when the acoustic difference was the same. For instance, VOT is a critical cue for stop consonant voicing. Voiced sounds like /b/ have short VOTs of around 0 msec, while voiceless sounds like /p/ have a longer VOT of around 60 msec and a boundary at around 20 msec. CP is thus observed when discrimination of two sounds with 40 and 60 msec VOTs (both /p/'s) is poor, but discrimination of 10 vs. 30 (a /b/ vs. /p/) is quite good. Because discrimination does not require overt labeling, this suggested that listeners can perceive acoustic differences that are relevant for discriminating categories but disregard differences that are not. CP suggests that listeners ignore any variability on this continuum, and one perceives a /b/ when it is below the 20 msec boundary no matter whether the VOT was 0, 10, or 15.

Theoretically, CP suggested that listeners solve the problem of lack of invariance by collapsing a continuous variable signal into discrete categories. That is, by ignoring within-category variation listeners could rapidly abstract a “quasi-symbolic” representation of the input (Goldstone and Hendrickson, 2010) such that a stimulus can be identified based solely on its relationship to the boundary: all VOTs <20 are /b/ and all VOTs greater than that are /p/.

While discrimination tasks comprise the core empirical definition of CP, it is the forced choice identification task that has left the most vivid impact on fields like bilingualism. Empirically, forced-choice identification tasks require participants to listen to stimuli from a continuum and report which of several categories provided is the best match. If there are two, then the task is a two-alternative forced-choice task (2AFC), but larger response sets are possible (nAFC). What made this task so compelling is that in these tasks, monolingual or so-called typical listeners often show a near-perfect step function (Figure 1A), which seemed to capture the discrete nature of the system. Consequently, any deviation from this ideal may be informative.

Moreover, unlike discrimination tasks, nAFC tasks are feasible in diverse populations [i.e., younger children see (Slawinski and Fitzgerald, 1998; Hazan and Barrett, 2000); people with language

impairments (Serniclaes, 2006); as well as bilinguals (Sebastián-Gallés and Bosch, 2002; Aoyama et al., 2004; Goriot et al., 2020)]. Many of these sorts of studies conducted on LX learners¹ and clinical populations using forced-choice identification tasks use the slope of the categorization function as an index of speech categorization ability. Here, a categorical or step-like response function (i.e., sharper slope) is interpreted as having a “strong” ability (Serniclaes, 2006). Any deviation from being categorical is interpreted as a deficiency in the system (Figure 1B).

This interpretation aligns with the assumptions of CP as to how a good listener should behave – sharp categorical boundaries indicate that the listener ignores variation and successfully reaches the category decision. Having a shallow slope indicates some “deficit in categorical precision,” either through noise in the system or being unable to map the categories successfully (Serniclaes, 2006). When a bilingual individual does not show monolingual-like categorization in their second language or their heritage language, they are perceived to be differing from native-like proficiency levels (even as the concept of native-like proficiency is also arbitrary as not all bilinguals have the same goals or needs for proficiency). Indeed, the fact that both clinical language disorders and multilinguals show these kind of shallower response functions further emphasizes the deficit interpretation.

However, as we describe in the next section, mounting evidence shows that the interpretation of shallow slope reflecting a deficiency is problematic and does not capture the essence of speech categorization, even in normal hearing, “typical” monolingual listeners. If this is the case, differences in an nAFC task that are standardly interpreted from the lens of deficiency may in reality be driven from methodological and theoretical limitations and could reflect a unique approach to speech perception that may be more flexible or efficient for a bilingual. Therefore, before considering whether bilinguals’ have “native-like

¹ We use the term “LX learners” to embrace the diversity of L2, L3, heritage language speakers and other forms of multilingualism (Dewaele, 2018).

categories,” it is crucial to understand what categories are in monolinguals.

Speech categories are gradient

Recent studies on speech perception have challenged the idea that speech input is carved into discrete categories and have shown clear evidence of gradiency in speech perception. While a history of studies has directly challenged the foundations of categorical perception using discrimination and other tasks to show continuous encoding (Massaro and Cohen, 1983; Schouten et al., 2003; Gerrits and Schouten, 2004; Toscano et al., 2010) our emphasis is not on criticisms of CP *per se*, but rather on a growing body of work that challenges the broader theoretical claim that discrete categorization is the ideal.

Evidence from studies that used priming (Andruski et al., 1994), continuous rating scales (Massaro and Cohen, 1983; Miller and Volaitis, 1989), the visual world paradigm with eye-tracking (VWP; McMurray et al., 2002, 2009; Kapnoula et al., 2021), and event-related potentials (ERPs; Toscano and McMurray, 2010; Sarrett et al., 2020; Kapnoula and McMurray, 2021) all converge on the idea that categorization is highly gradient. For instance, McMurray et al. (2002) tested monolinguals on a VOT continuum (e.g., spanning *beach* to *peach*) in a VWP task in which eye movements to each option were used to assess activation of the options (*beach* and *peach*) leading up to the ultimate nAFC response. They assessed *via* fixations to the picture of the competing word (e.g., *peach* when the target was *beach*), and found that this was linearly related to the continuous changes in VOT. That is listeners looked more to *peach* for a 10 msec VOT than a 0 msec VOT, even when they considered only trials where participants clicked on the target word (i.e., *beach*). This suggests that listeners are tracking continuous differences in VOT, within a category – not attempting to suppress these differences.

In fact, these gradient (rather than discrete) representations may be useful when listeners are coping with ambiguity and integrating different pieces of information in speech perception (McMurray et al., 2002, 2008, 2009; Clayards et al., 2008). For example, a more gradient representation may help listeners recover from misperceptions. McMurray et al. (2009) tested listeners on lexical garden paths words such as *barricade*, where the onset sound came from a/b/to/p/continuum. Here, if the VOT was high (e.g., 40 msec), the word may be briefly interpreted as both *parakeet*, and resolution would not occur until late in the word (at-cade or-keet). However, if listeners were preserving gradient representations, they may be able to recover more quickly when the VOT was near the boundary. They found that if the VOT was around 40 msec, listeners were initially biased to interpret the input as the beginning of *parakeet* and then revised their decision when-cade arrives. In contrast, when the VOT was around 25 msec, listeners were still biased to *parakeet* but recovered faster because *barricade* was more active. If listeners were categorical,

the activation of /p/ should have fully suppressed the activation of /b/. In this case, a more gradient commitment (rather than a firm commitment to a discrete category) may help listeners be more flexible to integrate later cues efficiently to recover from misperceptions. Similar results have been seen with a variety of sources of ambiguity, suggesting that a partial commitment is the norm in speech perception (Szostak and Pitt, 2013; Brown-Schmidt and Toscano, 2017; Gwilliams et al., 2018).

Beyond flexibility, a gradient commitment is also important in learning and adaption, particularly when speech is inherently varied. Listeners need to learn and adapt to the talker's speech to account for different factors such as their dialect, coarticulation patterns, rate of speech, or indexical differences. In fact, dozens of studies have documented the remarkable plasticity of speech perception (McQueen, 1996; Fenn et al., 2003; Bent et al., 2009). However, if listeners fully disregard fine-grained differences within a category, they would not be able to do this kind of learning (McMurray and Jongman, 2011). Importantly, these factors may interact. Clayards et al. (2008), for example, used a similar eye-tracking paradigm as McMurray et al. (2002), but with a perceptual learning twist. For some listeners, VOTs were highly consistent – most trials had VOTs near the prototypes for/b/ and/p/ with very little variation; for other listeners, VOTs were more variable. She found that after a brief exposure, listeners with high variance distributions adopted a more gradient representation. That is, people were learning to be more gradient when noise was expected (see also: (Theodore and Monto, 2019)).

This leads to the broader conclusion that underlying speech categorizations are highly gradient, and the degree of activation or consideration of one category or another is sensitive to fine-grained differences in continuous cues like VOT. However, it is unclear how to rectify this with traditional 2AFC tasks that show a sharp categorization. This is illustrated by a recent VWP study on the development of speech categorization. McMurray et al. (2018) used the same VWP paradigm from their 2002 study (McMurray et al., 2002) with children ages 7–18. Children heard tokens from VOT (e.g., *beach/peach*) or fricative spectra (*sip/ship*) continua and selected the corresponding picture while their eye movements were recorded as an index of lexical activation. The examination of the ultimate responses (i.e., the mouse click on the pictures) showed that older children had slightly steeper slopes. This appears to support the standard view – children's categorization is getting steeper (more discrete) with development, and the younger children's results mirror those of people with language impairments, or multilinguals (i.e., the association to the deficit model). However, the eye movements revealed a different story.

Similar to the McMurray et al. (2002) results, there was an overall gradient effect. As the VOT or fricative spectra moved toward the participant's category boundary, there were increased looks to the competitor, indicating that children were sensitive to these fine-grained acoustic details. However, the youngest children showed the *least* sensitivity to these fine-grained acoustic details in the eye-tracking experiment, and this sensitivity grew with

development. Under a quasi-discrete view, children with steeper slopes have *stronger* categories and should therefore be less sensitive to within-category detail. However, eye movements showed the exact opposite. In fact, it looks like children were achieving this sharper 2AFC categorization by becoming more sensitive to fine-grained details.

These findings have huge implications for how the slope of the identification is interpreted. However, beyond that, there are three deeper implications for work on multilingualism. First, even in monolingual children, speech perception skills develop slowly. This is unlike the most common views of speech categorization, which argues that speech categorization stabilizes in infancy (Werker and Curtin, 2005) (but see McMurray, 2022); critically in the context of multilingualism, it offers a gentle challenge to the notion that plasticity tapers off at later ages due to some kind of critical period – in fact, speech perception is developing quite slowly, implicating plasticity that may be available throughout the lifespan. Second, a gradient representation, rather than a discrete or categorical one, seems to be something desirable that people are attempting to develop. Finally, standard 2AFC tasks may show the complete opposite pattern of the underlying picture revealed by more sensitive measures like eye-tracking – a steep slope can accompany a highly gradient underlying representation.

This suggests serious problems with the traditional forced-choice identification tasks. In fact, it has long been known the discrimination tasks that formed the basis of support CP, involve other cognitive and decision processes that might create confounding factors for any given study (Gerrits and Schouten, 2004). However, nAFC identification is perhaps worse and the same pattern of data can be the product of completely different mechanisms of categorization (Kapnoula et al., 2017; Kapnoula and McMurray, 2021).

In classic two-alternative forced-choice tasks, listeners need to make a discrete judgment on a given trial. Because of the discrete nature of the response – and the effect of averaging – this can lead to enormous interpretative ambiguity. Consider a listener with a shallower-than-average identification curve (e.g., Figure 1B). Under the standard CP model, it would be assumed that this listener is responding variably from trial to trial—that is on some trials, a VOT of 15 msec (a /b/) is misperceived as a VOT of 25 msec (a /p/), leading to a different response. When averaged, we see a shallower curve. However, a shallow slope or a gradient profile might emerge from a completely gradient categorization. Here, listeners activate the competing category /p/ more near the boundary, and they attempt to approximate the frequency of their responses to the underlying gradient patterns. Therefore, they might respond 60% of the time indicating that the sound that they heard was a /d/ and 40% of the time that they heard /t/. These two profiles – both of which show shallower slopes – emerge from completely different underlying category structures. They cannot be differentiated from one another in a two-alternative forced-choice task. While the first scenario has a listener who does have discrete mapping between cues and categories, the second scenario has a listener whose underlying processes are gradient mapping.

The same is true for a steep (step-like) function. If we assume CP, this means that listeners have underlyingly discrete categories. However, if a gradient listener simply assumed a winner take all response mapping, one would see the same thing. That is, even if a speech token was perceived as 60% /b/like (e.g., near the boundary) if they always said/b/, one could observe a steep categorization curve even if the underlying categorization were gradient.

Thus, once we acknowledge that the underlying category structure could be gradient (as it clearly is in monolingual listeners), the 2AFC task is completely ambiguous. Clearly, not every shallower slope is due to gradiency – in many cases (e.g., hearing loss) it may be a marker of a problem. However, at the same time, in other cases, a shallower slope could be a sign of an adaptive and flexible gradient representation. Despite this ambiguity, both discrimination and forced-choice identification tasks are still widely used in language science research, and the assumptions that a steeper slope indicates more robust categorization are still commonly made. This is problematic not only for the larger language science community but also for research on bilingualism.

Categorical perception and bilingualism

As we have described CP exerts a dominant force on the study of multilingual speech perception. In fact, the study of bilingualism has its own share of methodological and theoretical misconceptions (see for a review (Surrain and Luk, 2019)). Early bilingualism research was built on deficit models and in part due to methodologies adopted from other disciplines. Consequently, for a long time, bilingualism was treated as a discrete category in comparison to monolingualism, failing to consider variability in language experience, proficiency, and sociolinguistic contexts of each language (on the other hand see Bice and Kroll, 2019; Surrain and Luk, 2019; Bayram et al., 2021; López et al., 2021; Tiv et al., 2021; Castro et al., 2022; Kutlu et al., 2022). These early approaches specifically focused on LX learners' ability to produce *native-like* utterances in their LX (Flege et al., 1995a; Piske et al., 2001; Alario et al., 2010). However, more recently, many scholars have begun to challenge this paradigm, assessing bilinguals on their own terms, given their own functional needs and environment. This work suggests bilingualism is better to be treated continuously and multi-dimensionally rather than as a discrete category (Surrain and Luk, 2019). In keeping with the classic views, the assumption of discrete categories and the methods of CP have also contributed to these deficit models. This shows up in at least two ways.

First, early bilingualism research assumed CP played a mechanistic role in explaining how well bilinguals learn categories or fully/partially transfer their L1 categories to their LX. That is, novel LX categories span regions of the perceptual space that lie within an L1 category (e.g., an English listener for whom the dental and alveolar /t's/ of Hindi lie within a single category).

Given CP, people cannot hear these distinctions, causing a barrier in learning them.

The strength of this account led CP to become a dominant component of theories of bilingual sound acquisition (MacKain et al., 1981). This was in part due to the emphasis on perceptual narrowing in speech perception which argues that starting in the first year of life, infants lose the sensitivity to discriminate sounds in other languages but get better at discriminating contrasting sounds in the language that surrounds them (Best et al., 1988; Kuhl et al., 2006; Werker et al., 2012). Critically, this loss was seen as the end of a critical or sensitive period, blocking further plasticity (e.g., LX learning).

Second, bilingualism research also makes heavy use of the discrimination and identification tasks that were pioneered in monolingual speech categorization (Werker and Tees, 1987; Sebastián-Gallés and Bosch, 2002; Aoyama et al., 2004; Goriot et al., 2020). Given the assumption of CP, a shallower slope of the identification function has been associated with deficits or as an inability to map LX categories accurately due to factors such as age of acquisition or proficiency. Crucially, many studies linked early age of acquisition and higher proficiency to successful outcomes (i.e., steeper slopes) of discrimination tasks (Bosch, 2011). However, as we described, these identification tasks may not be truly estimating the nature of speech categorization. As we described in the previous section, a listener can have a steep curve while underlyingly having gradient categorization or they can have a gradient curve while having a steep curve underlyingly. This is the fundamental ambiguity of the slope function in a 2AFC task. It is unknown what the underlying mechanism is as the 2AFC task is not capturing these differences accurately.

What is interpreted as a shallow slope, and hence an inability to robustly categorize the stimulus, may actually be a mark of listeners' flexibility and adaptation to categorizing overlapping categories. In fact, it may be almost impossible to impose fully discrete categories on the multiple phonological systems of a bilingual listener. Bilingual listeners need to adapt and learn from those cues more so than monolinguals. It is, therefore, more optimal to maintain a gradient mapping between cues and categories to permit more flexibility.

Categorical perception in the context of heritage bilingualism

Most of the bilingualism enterprise in the late 90s through early 2010s primarily focused on balanced bilingualism (e.g., Peltola et al., 2012). This is the type of bilingualism where the use or the proficiency in both languages are somewhat equal. However, this picture of a bilingual as two monolinguals (Grosjean, 1989) does not accurately describe bilinguals who do not have the societal support or educational platforms to help them maintain their languages. For example, someone from a Spanish-speaking home in an English-majority country may only have access to a more specialized Spanish vocabulary (that which is needed at

home) and may never learn to read Spanish, as their L2 (English) has much stronger support from school.

Historically, these minoritized bilinguals were consistently labeled as deficit language users (e.g., Bloomfield, 1927). That is, their abilities in their heritage language were seen as deficient (relative to a monolingual speaker of that language). To illustrate this point, consider two large bilingual populations in North America: (1) French/English bilinguals in Canada, and (2) Spanish/English bilinguals in Florida. If one strictly looks at age of acquisition for these two groups, it is possible to find early bilinguals in both contexts. It is also possible to find late learners of one of the languages in both contexts. What differentiates these two groups are primarily sociolinguistic factors.

In Canada, French is officially recognized as one of the two official languages spoken. Children in Canada (particularly Quebec) get some immersion in both languages, they are taught both in school, and they can maintain both Canadian French and Canadian English to some extent (even as bilingual groups in Canada who speak other languages and face other societal injustice toward their languages). Thus, many bilinguals in Canada are likely to fit the balanced bilingual definition, and many others are likely to be at least proficient in both languages.

In contrast, in the United States, cultural factors led to the stigmatization of Spanish (Kutlu and Kircher, 2021; Kircher and Kutlu, 2022), and as a result, immersion programs are rare and there are regional and racial disparities in access to general education in Spanish (Rosa, 2016). Children in the United States mostly do not receive any support in Spanish beyond the foreign language classroom, and for those who do, it is not sustainable at the national level. These children experience what is known as the heritage bilingual experience, where their home language is limited to the home settings due to societal prejudice and stigmatization. This prejudice is not only disrupting heritage speaker children's heritage language development but also their bilingual development.

In these listeners, Spanish is perceived as a problem that needs to be fixed when bilingual children start schooling (Rosa, 2016). Children who have categories that are somewhat ambiguous in their comprehension and production are asked to fix this problem by immersing themselves in a solely English educational context (Rosa, 2016; García et al., 2021). The majority of the work on Spanish heritage-speaker children in the United States has argued that communities of minoritized languages should find ways to increase heritage-speaker children's exposure to "native English speakers" to prevent them from having a gap in their English (Place and Hoff, 2011). Such recommendations use individual variability in development as a case for the assumption that heritage-speaker children cannot develop or are delayed in developing English proficiency as they are exposed to English at a later age or with a reduced amount while ignoring the social stigmatization towards bilingualism (Kutlu, 2020; Kutlu and Wiltshire, 2020).

Heritage speakers (of any language) are not a homogenous group (see Polinsky, 2018; Montrul and Polinsky, 2021).

There are substantial differences in terms of exposure to the heritage language and the majority language, feelings of attachment to these languages, as well as perceived fluency in these languages. A survey of the literature reveals a wide variety of definitions and classifications of heritage bilinguals and heritage languages [e.g., (Benmamoun et al., 2013); also see (Ortega, 2020) for a detailed discussion]. In this context, there has been substantial work suggesting that heritage speakers are “deficient” in the majority language (Oller et al., 2007; Hoff, 2013), and that they cannot be considered “native speakers” of that language [but also see new conceptualizations on how heritage speakers can be placed in the native speaker continuum (Wiese et al., 2022)]. At the same time, research on heritage bilingualism has also focused on how heritage speakers retain and process their heritage language – which may also be deficient by this standard. This anti-nativization of heritage speakers from both their heritage and the majority language provided an array of places where heritage speakers experience a state of languagelessness (Rosa, 2016). Neither their heritage language nor their majority language fits into the “standard” norms. Much of the research on heritage bilingualism was done with the purpose of “fixing their languages” by providing more of the majority language.

Work on speech perception in heritage language speakers has the potential to fall into the trap of the deficit model. As we have described, the standard approach to speech perception in bilinguals was motivated by perceptual narrowing and CP (Caramazza et al., 1973; Werker et al., 1981; Werker and Tees, 1984; Flege, 1987; Flege et al., 1995b; Mayo et al., 1997; Sebastián-Gallés and Bosch, 2002, 2009; Bosch and Sebastián-Gallés, 2003; Aoyama et al., 2004; Kuhl et al., 2006; García-Sierra et al., 2011; Stölten et al., 2014; Liu and Kager, 2015; Pan et al., 2022). In this context, any deviation in endpoints was interpreted as noisy encoding, deficiency in categories in their languages, or an unstable state of language use. However, a compelling alternative that cannot yet be ruled is that heritage speakers or bilinguals may be more gradient than monolinguals or individuals primarily exposed to one language or one language variety. This may serve to help them flexibly shift between languages.

Such interpretation does not require “fixing” a non-existent problem but focuses on the strengths of language-diverse individuals and how it informs our theories and methodologies. In fact, a theory based on deficiency arguments that do not consider language diversity has more potential to lead to educational outcomes that actually hinder language-diverse individuals from achieving the specific skills they need to navigate the educational system. That is an intervention designed to make such individuals perceive speech more categorically may actually be harmful. However, a fundamental limit is that the 2AFC task simply cannot distinguish a noisier or poorer category representation from a more gradient one. Thus, to better inform theories of speech perception, we must move towards a continuous measure of speech perception.

Moving away from categorical understanding of speech: Measuring gradiency

Given the interpretive ambiguity in the 2AFC task and the strong likelihood that categories are underlyingly gradient, there is a clear need for methods that can more directly assess this. Both eye-tracking and EEG studies have previously captured this and can clearly show the underlying gradient profiles (McMurray et al., 2002; Toscano and McMurray, 2010). However, using these methods is not trivial: they have large technical requirements, can be expensive, and have a great deal of trial-to-trial noise, requiring longer experiments.

In contrast, several recent studies have suggested that gradient categorization can also be measured by a simple behavioral experimental tool that can be implemented in online studies, lab studies, or field studies. This task, which we refer to as the Visual Analogue Scaling (VAS) task (Kong and Edwards, 2011, 2016; Kapnoula et al., 2017, 2021; Kapnoula and McMurray, 2021), is similar to the 2AFC task, however, as we argue below, its psychometric properties nearly eliminate the interpretive ambiguity of 2AFC.

In the VAS task, as in the 2AFC task, a sound from a speech continuum is presented. However, instead of making a discrete binary choice, listeners are given a continuous scale on which to indicate where the sound falls between endpoints. For instance, if a listener is responding to members of a *beach/peach* continuum, the screen has an image of a *beach* on the left and a *peach* on the right with a straight scale in between (see Figure 2). They can then click anywhere on the line to indicate where they perceived this token.

In contrast to ERP and VWP tasks, the VAS task is straightforward and efficient to implement. It can be employed on any experiment building software (e.g., Experiment builder, PsychoPy, Matlab, Gorilla). It generally takes 15–20 min to complete with 2 to 6 repetitions per continua (moderate test–retest reliability ($r = 0.48$) of gradiency estimates was achieved with three repetitions of each stimulus see (Kong and Edwards, 2016)).

Critically, it overcomes concerns with the 2AFC task. Consider a situation in which listeners have a categorical or discrete boundary, but experience noise. In this case, the average VAS function should look highly gradient (like the 2AFC). However, if we look at individual trials, we should see that most trials have a VAS response that is near one endpoint or the other (Figure 3B grey points). That is, on each trial, they discretely heard /b/ or /p/, but shifts from trial to trial. In contrast, if the averaged response was because of an underlyingly gradient representation, we should see that individual responses are clustered tightly near the averaged (Figure 3B black points). Thus, by looking at individual trials relative to the average response, we can achieve more insight into the underlying nature of categorization.

Much of the recent work using the VAS task has examined individual differences in typical monolingual adults. For example, individual differences in gradiency in the VAS task predict the

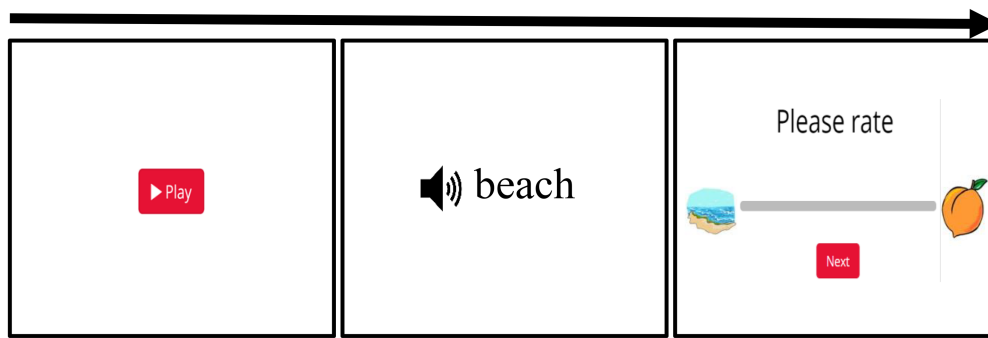


FIGURE 2

The layout of the experiment where the first panel shows the first page that the participants see when they are completing the online experiment. The second panel is when they hear the auditory stimuli. The last panel shows the VAS rating scene. Participants were asked to click on the scale to indicate where they think the auditory stimulus falls.

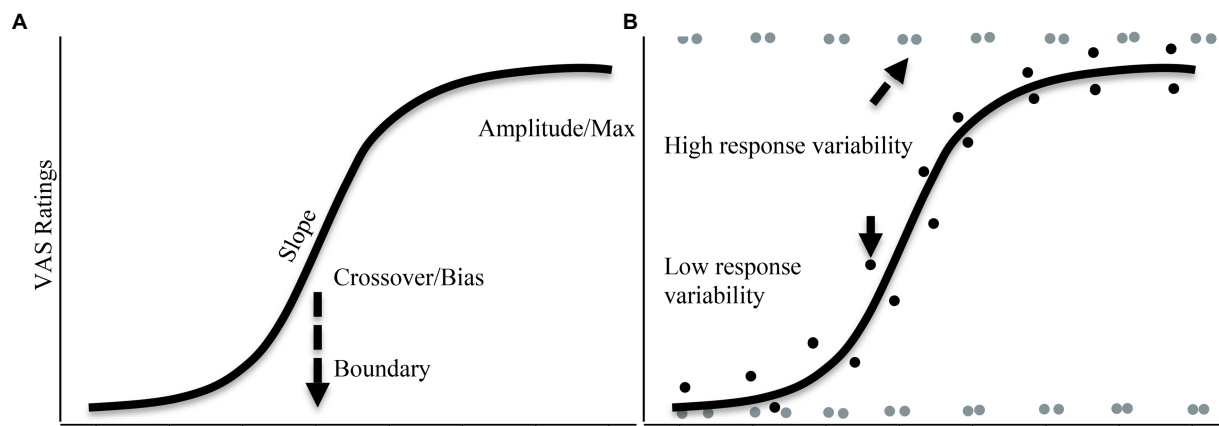


FIGURE 3

(A) Three parameters that can be extracted from VAS rating data. Averaged data across a continuum can provide measures such as amplitude, slope, and crossover/bias. (B) High response variability where a listener repeatedly uses the endpoints (in gray dots). This listener has underlyingly categorical representation. On the other hand, black dots show a listener who has an underlyingly gradient representation whose response variability is lower as their responses are tightly distributed around the average.

degree of gradiency in standard ERP and VWP paradigms (Kapnoula and McMurray, 2021), providing validation of the underlying constructs. It also predicts other skills. For example, adult listeners who more gradiently categorize stop voicing are more likely to use secondary cues (i.e., F0) when categorizing voicing (Kong and Edwards, 2016; Kapnoula et al., 2017), and they are better able to recover from misperceptions (Kapnoula et al., 2021).

Importantly, and most relevant to bilingualism research, in monolinguals there was little correlation between cognitive control tasks and the VAS task (Kapnoula and McMurray, 2021) or between gradiency in a non-linguistic visual continuum (Kapnoula et al., 2021) (e.g., an apple/pear visual continuum). This lack of an influence of more domain-general cognition may make it easier to isolate differences in speech perception in

varying groups. Ongoing work in our lab is now successfully using this technique with monolingual children, monolingual children with bilingual exposure, bilingual adults, and cochlear implant users. Here we present preliminary data from a study in progress on bilingual adults to illustrate both how to use the VAS paradigm with this population, and how it can lead to greater clarity than prior approaches.

We note that this is an ongoing study and no statistical analyses have been conducted (as we have not reached our pre-planned sampling goal). Thus, our goal in presenting this data is not to make any specific claims about differences across bilinguals. Rather we examine this subsample of the data to illustrate how a more sophisticated approach to speech categorization could offer the kind of person-centered approach to speech that naturally fits with a more sensitive approach to

bilingualism. Thus, our analyses are really meant more as a kind of case report for illustrative purposes, and we do not report many of the methodological details so as to avoid distracting from our goals here.

A case study

To illustrate how we have used the VAS task (both methodologically and statistically) we present examples of data from an ongoing project. The goal of this project was to understand a diverse array of Spanish/English bilinguals in terms of relative proficiency, age of acquisition, social environment, and how these factors give rise to differences in speech categorization. We used online testing to recruit individuals with experience with Spanish and English in the United States and conducted an extensive language background questionnaire, social network questionnaire, and other measures along with a VAS task assessing eight different continua.

We note that the goal of this paper is solely to illustrate how we can assess gradient speech categorization in bilingualism research and separate it from poor categorization in a way that cannot be captured by standard nAFC tasks. Thus, we did not conduct any statistical analysis which awaits our true sample.

Subjects

We have currently tested 73 listeners of various backgrounds. For ease of exposition below, we roughly group these subjects using age of acquisition to create four groups similar to those used in previous bilingualism research. The *Spanish-English heritage* ($n=32$) speakers are defined as those with experience with both languages during the first 10 years of their lives and who also self-identify as dominant English speakers. The *L2 Spanish* ($n=9$) group consists of individuals who acquired Spanish after the age of 10 only through schooling experiences, and for whom English is their dominant language. Next, participants in the *L2 English group* ($n=6$) are those whose first language is Spanish and who acquired English as a second language after the age of 10. These individuals have the least dominance in English. Finally, the *English monolingual* group ($n=26$) consists of speakers who acquired only English. These categories are designated solely for illustrative purposes and were not the groupings that originally motivated our ongoing study.

Auditory stimuli

Auditory stimuli used in VAS experiments consist of monosyllabic minimal pairs in any language. Here, we provide an example of the eight continua used in our experiment. Our continua included two voicing contrasts (*beach-peach*, *dime-time*), five vowel contrasts (*beet-boot*, *bet-bat*, *pen-pan*, *hat-hot*, *net-nut*),

and one fricative contrast (*sip-ship*). To construct the stimuli, we started by recording each endpoint word, spoken by an adult male with an American Mid-Western accent. The recordings were done in mono at a sampling rate of 44,100 Hz. Exemplars for endpoints were recorded in a carrier sentence to ensure uniform prosody and rate. We then selected one exemplar for each endpoint for each continuum.

VOT continua were created with a progressive cross-splicing procedure similar to (McMurray et al., 2008). Aspirated tokens were created by copying segments of the aspiration from *peach* and *time* and replacing the corresponding section of the onset of *beach* and *dime*, respectively. Fricative continua were created by a morphing procedure from McMurray and Jongman (2016). The frication portions from *sip* and *ship* were extracted, centered, and cut to be equal in length. Next, the spectral mean was calculated from the long-term average spectra. Both spectra were aligned to the average spectral mean. Then, weighted averages of the spectral shapes were extracted to create 0% /s/ to 100% /s/ in nine steps. Next, the frequency means of the spectra were shifted to create nine steps and a white noise filter was applied to each spectrum. Then, we imposed an average amplitude envelope on the filtered noise. Finally, the vowel continua were created by using TANDEM STRAIGHT (Kawahara et al., 1999). To create vowel continua, periodic information was first extracted for each endpoint. Then, temporal anchors were placed at the beginning, middle, and end of the target sounds. Spectral anchors were placed at the first and second formants. Finally, continua were morphed from one endpoint to the other across nine steps.

Visual stimuli were developed using a picture norming process adapted from McMurray et al. (2010). Candidates for stimuli were downloaded from a commercial clipart database², then selected by a committee of undergraduate and graduate students for the most prototypical image. Images were then edited based on committee feedback (changing colors, removing or adding parts to the image), and edited to a uniform size and brightness.

Procedures

The VAS task can be completed in the lab *via* touch-screen tablets and computers, or online *via* an internet browser. The example that is provided here was for online experiments which were implemented in Gorilla [³ (more about touch-screen testing with children can be found here: OSF⁴)].

In this task, participants hear a token from the continuum and report how closely it matched either endpoint by clicking along a line between the two pictures (see Figure 2). They are allowed to practice responding before proceeding to three practice trials,

² <https://clipart.com>

³ www.gorilla.sc

⁴ <https://osf.io/q39yt>

identical to experimental trials. After three practice trials, the participant begins the task.

On each trial, participants press a red PLAY button to initiate the word. After the word plays, the line appears and remains until the participant responds. Crucially, the line does not contain a slider or any marker until the participant makes a response (at which point a marker is shown). This avoids anchoring biases. The participant can change their response and indicate they are done by pressing the space bar.

Generally, when using multiple continua, we find it is much more efficient to present multiple trials from the same continuum in a block, and to maintain the sides of the pictures (e.g., the *beach* is consistently on the left and the *peach* on the right for some block of trials). This minimizes the amount of time that the participant needs to reorient to the task on each trial. However, in order to control for order effects and side bias, participants completed two blocks of each continuum which counterbalance the location of the endpoints along the response line. For example, the participant may see a picture of a beach on the left and a peach on the right in the first block, then see a *peach* on the left and a *beach* on the right in the eighth block (with blocks from other continua interspersed). Consequently, each continuum appears both early and late in the trial, with each endpoint on each side. In this study, each block consisted of 3 repetitions of 9 steps for each continuum or 27 trials/block. With two blocks for each of the 8 continua, this led to 432 total trials. The entire experiment took approximately 25–30 min.

Parametric analyses of VAS

Typically, in a 2AFC task, listeners must judge the endpoints as 0 (e.g., /b/) and 1 (e.g., /p/). A classic step-function of CP is thus when tokens on one side of a category boundary would be marked as 0, and all tokens on the other side of the boundary would be 1. The VAS data are different in the sense that these data are on a continuous rating scale which reflects how close each stimulus is to the endpoints.

Previous work on VAS (Kong and Edwards, 2016) utilized a simple histogram to illustrate the way that listeners vary in their use of these continuous responses. This approach simply counts how often listeners respond to each point along the VAS. This method showed that some listeners used only the endpoints when responding, and others used the whole scale (i.e., more gradient).

However, Kapnoula et al. (2017) pointed out that this ignores the actual continuum step – a listener could have a flat histogram (a uniform distribution) because their responses perfectly match the continuum step (e.g., step 1 gets a low rating, step 2 gets a slightly higher one and so forth), or because they are just guessing. They thus introduced a parametric approach using non-linear curvefitting. They first computed the average response for each participant at each step. This was then fit to a nonlinear function. This function provides parameters like the slope of the function,

the boundary, and the amplitude (the difference between the asymptotes; Figure 3A). These are described below and allow the researcher to directly characterize the shape of the function at an individual level.

To capture the nature of the trial-by-trial responses (e.g., Figure 3B), they then computed the difference between each individual trial rating and the mean. Therefore, if participants' responses to individual trials overlapped with the mean, the sum of squared differences should be minimal, creating low response variability (true gradiency, Figure 3B, dark points). However, if the participants are choosing the endpoints, the individual responses should show a large deviation from the overall mean, leading to a higher sum of squared differences (high response variability, Figure 3B, light points).

The typical non-linear function is a four-parameter logistic function. These four parameters are (see Figure 3A): the slope, the amplitude (asymptotes), the crossover, and the bias. These parameters can efficiently capture gradient responses. When coupled with the response variation, these parameters can differentiate between a gradient pattern from a categorical pattern, and crucially a gradient pattern from a noisy pattern.

Slope

The slope of the VAS dataset is analogous to that of a 2AFC task. It measures how the average function smoothly varies between the tokens or exhibits a more step-like function. However, as we have described in a 2AFC task, it is not possible to know whether a shallow slope emerges as a result of noisy or a more gradient encoding. The VAS task resolves this problem by incorporating response variability in the interpretation of the slope function as explained below.

Amplitude

The amplitude indicates an overall difference between the asymptotes of the response function (i.e., position at which extreme end of the continuum). This measure often was ignored in 2AFC experiments, since the expectation was that in a forced-choice task the response should not be ambiguous (i.e., choosing one or the other category). However, in many groups, the endpoints of the continuum may never be unambiguous. For example, competition from an LX category could destabilize the response, or the particular acoustic cue that was manipulated in the continuum may not be the same cue the listener is expecting. The use of a logistic function with variable amplitude can eliminate this problem – particularly in a VAS task where listeners can respond to tokens in the endpoints continuously and are not required to use the ends of the scale. In fact, the differences in endpoint ranges may play a crucial role in understanding individual differences. Differences in amplitude parameter may be independent of differences in slope. For instance, a listener

could have a low amplitude but steep slope or a low amplitude and shallow slope.

Crossover

The crossover is the point where the function shifts from being in one category to the other (i.e., the category boundary). This is strongly analogous to the boundary seen in 2AFC tasks and can be used for similar inferences.

Bias

The bias is the overall likelihood of a listener's use of one end of the scale or the other end (e.g., the degree of vertical shift). This is introduced by the fact that the asymptotes need not reach 0 and 1.

To demonstrate, the average slope of each language group is shown in Figure 4. Here, we see that participants learning English as a second language (i.e., blue curve) has the lowest slope, followed by English monolingual speakers (dark purple) and Spanish-English Heritage speakers (gold), then followed by participants learning Spanish as a second language (green). Notably, there appear to be differences between Spanish-English heritage speakers and English monolingual speakers. The average amplitude of each language group in Figure 4 shows that participants learning Spanish as a second language and Spanish-English heritage bilinguals have the highest amplitude compared to the English monolinguals and English as a second language learners.

Moreover, participants learning Spanish as a second language seem to have a lower minimum and higher maximum response as compared to other language groups. English monolingual and heritage speaker groups seem to have identical minimum value. While these parameters visually present group differences, the VAS data provide further insight which is the calculation of response variability.

Response variation

In addition to the parameters of the averaged estimated functions, we must also consider how closely the individual responses map onto these estimates. For example, a shallow average slope could arise from two distinct patterns of responding that cannot be captured by the 2AFC task but can be captured by the VAS task. First, a participant could be responding continuously to the different steps which results in a shallow slope with response points closely clustered around their slope. Alternatively, a shallow slope may emerge from a participant that responds primarily close to the endpoints of the line but does so inconsistently – the same step is categorized differently across responses. This latter pattern

also results in a shallow slope when responses are averaged, but the majority of the data points would fall far from the average.

A close consideration of these patterns suggests that residual variation captures individual differences in three different profiles: categorical (steep slope), gradient (shallow slope + low response variability), and noisy but looks gradient (shallow slope + high response variability).

We use the parameters of the non-linear function to compute a response variability index. For this, we simply compute the predicted value for each step for that subject and then compute the mean squared difference of each individual point from the predicted value. If a participant is responding continuously, they will have low response variability and a shallow slope. On the other hand, a participant who inconsistently responds will have a large residual variation value and may have the same shallow slope. Fundamentally, both of these participants have shallow slopes. However, while one has a shallow slope because they integrate and use fine-grained details of the speech continuum, the other has it because of noisy encoding. Therefore, bilinguals showing shallow slope being interpreted as noisy encoding might in fact be the opposite of fine-grained gradient encoding (see Figure 5A). That is, by using slope (or amplitude) along with response variability we can differentiate a noisy response pattern from a more gradient or flexible categorization (Figure 5A versus Figure 5B). Figure 6 shows it clearly. Theoretically, we argue that response variability occurs when listeners disregard the acoustic cues in a noisy manner (i.e., being categorical but noisy). Here, we show that such noisy responses are also possible to see in English monolinguals.

We are currently developing new ways of analyzing VAS data that allow us to simultaneously estimate the logistic function and the response variability in the same multi-level model. In particular, as part of our co-registration for the *Growing Words* Project, we have proposed using a non-linear Bayesian mixed model (see more here⁵). However, this is still an ongoing effort to implement better statistical tools to investigate such data (see Figure 6).

Limitations

As is the case with every experimental design, the VAS design also comes with its own limitations. For instance, it is unknown whether VAS has heavy demands on working memory. A listener who first categorizes the token and then converts their categorization to ratings might need to rely on their memory to remember the initial categorization. While this remains unknown, we argue that the VAS continuous responses are still a better option than an nAFC task as they minimally lack the interpretive ambiguity of that task. Future research should investigate whether

⁵ <https://osf.io/q39yt>

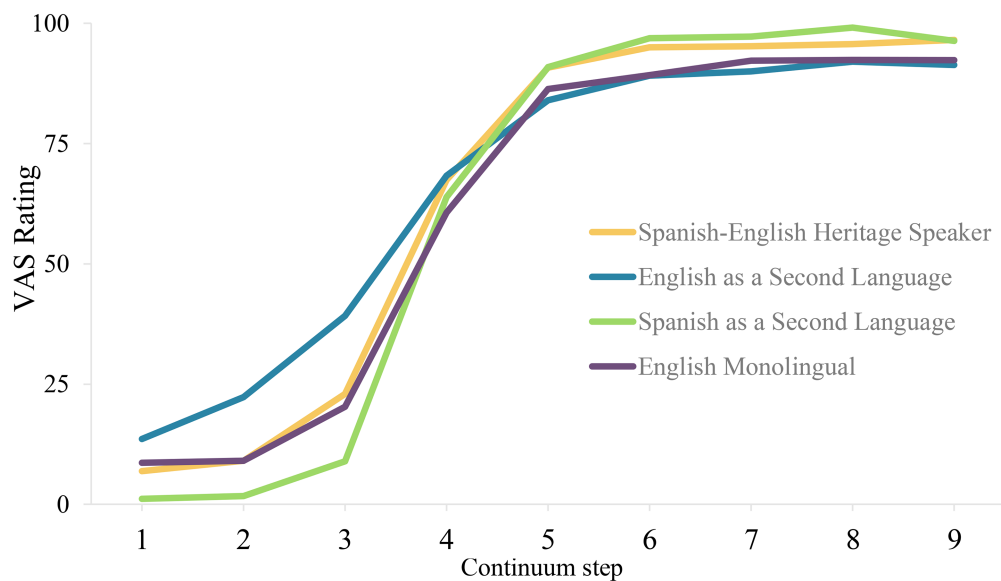


FIGURE 4

VAS ratings for four groups created based on Age of Acquisition of English and Spanish. Across four groups both Spanish as a second language and Spanish-English heritage speakers have the highest amplitudes. Spanish-English heritage speakers follow a more gradient slope compared to Spanish as a second language group.

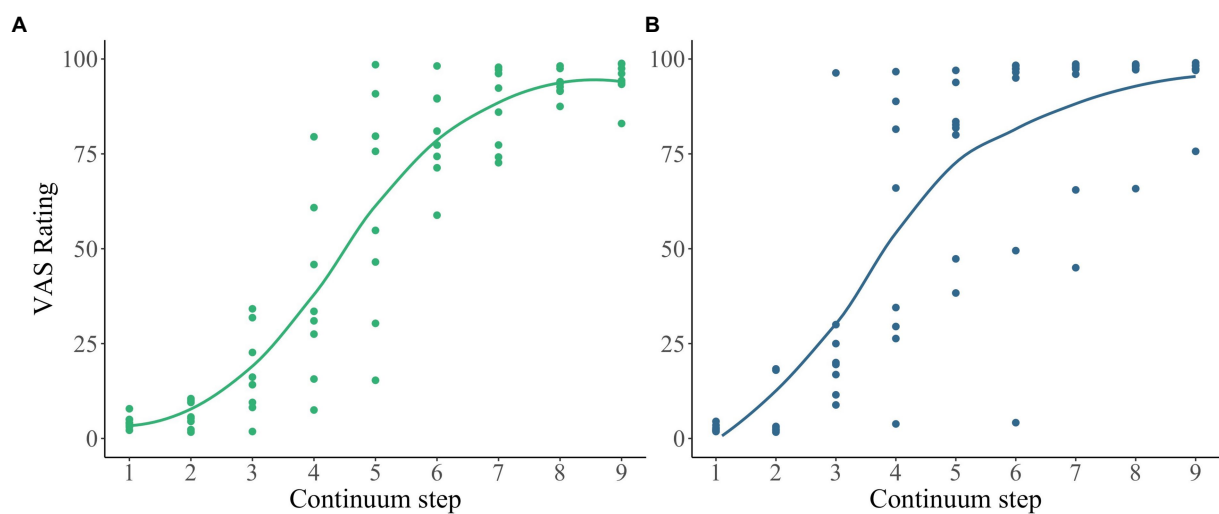


FIGURE 5

This figure shows the VAS ratings averaged across 9-steps for all continua for two participants. (A) Spanish-English heritage speaker who has a gradient profile with lower response variability. (B) English monolingual who has a categorical but somewhat noisy profile. Their response variability is larger compared to the heritage speaker.

bilingual and monolingual differences are driven due to potentially different memory use.

While the VAS could be combined with an analysis of reaction time, one challenge with this is that the use of RTs in the VAS task might not be ideal as listeners can change their decision before advancing to the next trial. If RTs are needed, we suggest that the experiment not allow the listeners the

option to change their decision and provide instructions that encourage fast responding.

A third issue is that the VAS task is not the most ecological experimental setup for speech perception. It depends on computer-generated speech tokens which might not be possible to hear in the real world and also asks listeners to do something fairly unnatural with speech. However, at the same time, it

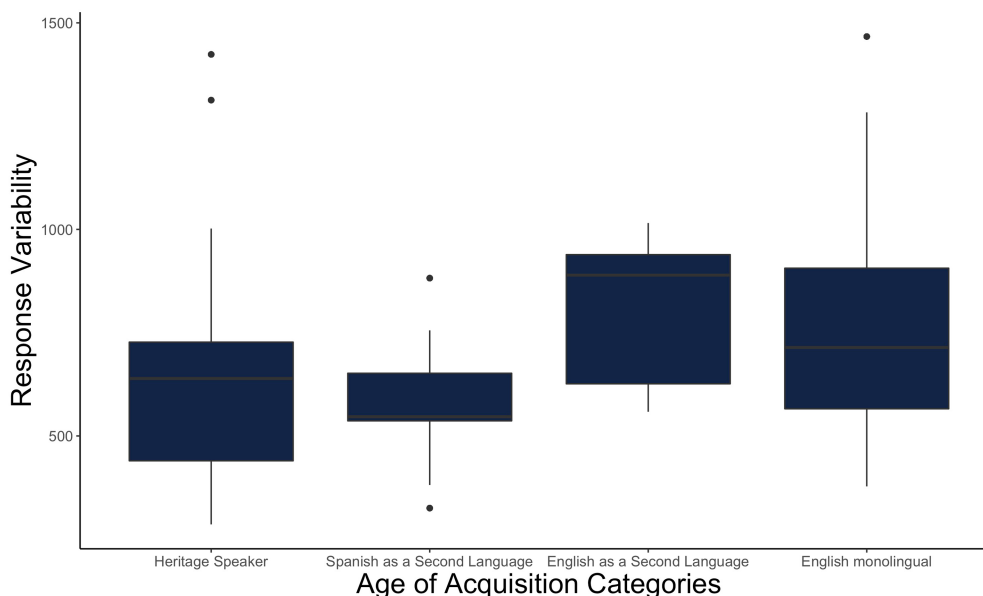


FIGURE 6

The residual variance calculated through the above formula on the y-axis. The higher the variance, the noisier the responses are. Here, we plotted four categories that were extracted from the Age of Acquisition variable from our dataset.

may allow researchers to isolate processes involved in speech categorization that cannot be seen in other ways. Moreover, we point out that gradiency estimates from the VAS correlate with the much more natural VWP, which involves matching sounds to pictured referents (Kapnoula et al., 2021), suggesting this is not a major problem.

Finally, the development of analytic tools for the VAS is still an ongoing project. While different parameters can be extracted and frequentist statistics can be applied, Bayesian modeling might provide unique ways to delve into the rich VAS data. While these models are complex, open science practices that allow the sharing of scripts and data should allow researchers to practice their preferred way of approaching the VAS analysis.

Conclusion

Classic work on bilingual speech perception has assumed categorical perception – both as a set of methods to be deployed and a theoretical “goal” of speech as providing quasi-discrete categories. In the context of bilingualism, the widespread use of these tasks suggested shallower identification slopes or poorer speech perception. This dovetailed with a deficit model of bilingualism in which any deviation from the monolingual performance was not accounted for until recently (Berthele, 2021; Wiese et al., 2022). Ultimately, this group of studies did not consider the fact that, just like with other language experiences, the bilingualism experience is not a static one. Bilinguals learn new languages, stop using those languages, or continue using those languages dominantly due to various personal and/or

societal reasons (Kutlu and Kircher, 2021; Tiv et al., 2022). Considering language acquisition as a short period of the learning process that closes during the early years of childhood puts bilinguals such as heritage speakers into a never-ending gray zone. Heritage speakers were too bilingual for monolingual comparisons but too monolingual for bilingual comparisons. However, as research in monolinguals has abandoned both the method and the theory, this creates new opportunities for understanding bilinguals. In particular, speech categories are highly gradient and may be important for flexibility. This, along with a richer understanding of the diversity of the bilingual experience, demands new methods for understanding speech.

Thus, this paper described a new experimental method—the VAS task—that offers a more in-depth understanding of how bilinguals might categorize speech sounds. This may help avoid a deficiency argument by allowing the researcher to better characterize the process along multiple dimensions, by helping to identify structural gradiency in categorization (which may be adaptive), and to discriminate it from patterns that reflect difficulty. Language science research is moving more towards such gradient analysis in other fields as well (Levshina et al., 2021), and our own contribution here builds on this important trend. Importantly, methods that embrace this kind of gradiency may ultimately help build a more interdisciplinary approach to language science as not all subfields of language sciences have historically ignored variation (i.e., years of sociolinguistic research that examine variation). Moreover, our argument is consistent with broader trends in psycholinguistic research to continuously integrate an understanding of variability (both within and across individuals) in our methods and theories (Titone and Tiv, 2022).

Forced-choice tasks may be useful in some contexts including work on bilingualism. However, if the primary concern is the slope or steepness of the function, this task is highly ambiguous, and a shallow slope could be due to a deficiency or to increased gradiency. Consequently, this task may lead to the interpretation of monolingual-bilingual differences as deficiencies, which may in fact not reflect the reality of speech perception neither in monolinguals nor in bilinguals. In fact, for listeners who are surrounded by variability in their everyday lives, gradiency might be more ecological and cognitively efficient than a discrete representation. Furthermore, bilingual research continues to move away from trying to fix so-called deficiencies that do not exist (Bayram et al., 2021). The VAS task, along with many recent theories and methodologies, is one of the tools that can continue to provide researchers with tools to account for individual differences.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

EK and BM contributed to conception and design of the study. SC organized the experiment. EK and SC performed the analysis. EK wrote the first draft of the manuscript. BM and SC both contributed to the writing and editing of this manuscript. All

authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by NSF grant number 2104015 to EK and BM, and NIH grant DC 008089 to BM.

Acknowledgments

We would like to thank the guest editors for allowing us to present our work. Special thanks to all the members of the MACLAB at the University of Iowa.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Abramson, A. S., and Whalen, D. H. (2017). Voice onset time (VOT) at 50: theoretical and practical issues in measuring voicing distinctions. *J. Phon.* 63, 75–86. doi: 10.1016/j.wocn.2017.05.002
- Alario, F. X., Goslin, J., Michel, V., and Laganaro, M. (2010). The functional origin of the foreign accent: evidence from the syllable-frequency effect in bilingual speakers. *Psychol. Sci.* 21, 15–20. doi: 10.1177/0956797609354725
- Andruski, J. E., Blumstein, S. E., and Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition* 52, 163–187. doi: 10.1016/0010-0277(94)90042-6
- Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese/r/ and English/l/ and /r/. *J. Phon.* 32, 233–250. doi: 10.1016/S0095-4470(03)00036-6
- Bayram, F., Kubota, M., Luque, A., Pascual y Cabo, D., and Rothman, J. (2021). You can't fix what is not broken: contextualizing the imbalance of perceptions about heritage language bilingualism. *Front. Educ.* 6:628311. doi: 10.3389/feduc.2021.628311
- Benmamoun, E., Montrul, S., and Polinsky, M. (2013). Heritage languages and their speakers: opportunities and challenges for linguistics. *Theor. Linguist.* 39, 129–181. doi: 10.1515/tl-2013-0009
- Bent, T., Buchwald, A., and Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *J. Acoust. Soc. Am.* 126, 2660–2669. doi: 10.1121/1.3212930
- Berthele, R. (2021). The extraordinary ordinary: re-engineering multilingualism as a natural category. *Lang. Learn.* 71, 80–120. doi: 10.1111/lang.12407
- Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 345–360.
- Bice, K., and Kroll, J. F. (2019). English only? Monolinguals in linguistically diverse contexts have an edge in language learning. *Brain Lang.* 196:104644. doi: 10.1016/j.bandl.2019.104644
- Bloomfield, L. (1927). Literate and illiterate speech. *Am. Speech* 2, 432–439. doi: 10.2307/451863
- Bosch, L. (2011). Precursors to language in preterm infants: speech perception abilities in the first year of life. *Prog. Brain Res.* 189, 239–257. doi: 10.1016/B978-0-444-53884-0.00028-2
- Bosch, L., and Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. *Lang. Speech* 46, 217–243. doi: 10.1177/00238309030460020801
- Brown-Schmidt, S., and Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Lang. Cogn. Neurosci.* 32, 1211–1228. doi: 10.1080/23273798.2017.1325508
- Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., and Carbone, E. (1973). The acquisition of a new phonological contrast: the case of stop consonants in French-English bilinguals. *J. Acoust. Soc. Am.* 54, 421–428. doi: 10.1121/1.1913594
- Castro, S., Wodniecka, Z., and Timmer, K. (2022). Am I truly monolingual? Exploring foreign language experiences in monolinguals. *PLoS One* 17:e0265563. doi: 10.1371/journal.pone.0265563

- Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition* 108, 804–809. doi: 10.1016/j.cognition.2008.04.004
- Dewaele, J.-M. (2018). Why the dichotomy 'L1 versus L2 user' is better than 'native versus non-native speaker'. *Appl. Linguis.* 39, 236–240. doi: 10.1093/applin/amw055
- Fenn, K. M., Nusbaum, H. C., and Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature* 425, 614–616. doi: 10.1038/nature01951
- Flege, J. E. (1987). Production and perception of English stops by native Spanish speakers. *J. Phon.* 15, 67–83. doi: 10.1016/S0095-4470(19)30538-8
- Flege, J. E., Munro, M. J., and MacKay, I. R. (1995a). Factors affecting strength of perceived foreign accent in a second language. *J. Acoust. Soc. Am.* 97, 3125–3134. doi: 10.1121/1.413041
- Flege, J. E., Munro, M. J., and MacKay, I. R. (1995b). Effects of age of second-language learning on the production of English consonants. *Speech Comm.* 16, 1–26. doi: 10.1016/0167-6393(94)00044-B
- García, O., Flores, N., Seltzer, K., Wei, L., Otheguy, R., and Rosa, J. (2021). Rejecting abyssal thinking in the language and education of racialized bilinguals: a manifesto. *Crit. Inq. Lang. Stud.* 18, 203–228. doi: 10.1080/15427587.2021.1935957
- García-Sierra, A., Rivera-Gaxiola, M., Percaccio, C. R., Conboy, B. T., Romo, H., Klarman, L., et al. (2011). Bilingual language learning: an ERP study relating early brain responses to speech, language input, and later word production. *J. Phon.* 39, 546–557. doi: 10.1016/j.wocn.2011.07.002
- Gerrits, E., and Schouten, M. E. (2004). Categorical perception depends on the discrimination task. *Percept. Psychophys.* 66, 363–376. doi: 10.3758/BF03194885
- Goldstone, R. L., and Hendrickson, A. T. (2010). Categorical perception. *WIREs Cogn. Sci.* 1, 69–78. doi: 10.1002/wcs.26
- Goriot, C., McQueen, J. M., Unsworth, S., Hout, R., and Broersma, M. (2020). Perception of English phonetic contrasts by Dutch children: how bilingual are early-English learners? *PLoS One* 15:e0229902. doi: 10.1371/journal.pone.0229902
- Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain Lang.* 36, 3–15. doi: 10.1016/0093-934X(89)90048-5
- Gwilliams, L., Linzen, T., Poeppel, D., and Marantz, A. (2018). In spoken word recognition, the future predicts the past. *J. Neurosci.* 38, 7585–7599. doi: 10.1523/JNEUROSCI.0065-18.2018
- Hazan, V., and Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *J. Phon.* 28, 377–396. doi: 10.1006/jpho.2000.0121
- Hoff, E. (2013). Interpreting the early language trajectories of children from low-SES and language minority homes: implications for closing achievement gaps. *Dev. Psychol.* 49, 4–14. doi: 10.1037/a0027238
- Kapnoula, E. C., Edwards, J., and McMurray, B. (2021). Gradient activation of speech categories facilitates listeners' recovery from lexical garden paths, but not perception of speech-in-noise. *J. Exp. Psychol. Hum. Percept. Perform.* 47, 578–595. doi: 10.1037/xhp0000900
- Kapnoula, E. C., and McMurray, B. (2021). Idiosyncratic use of bottom-up and top-down information leads to differences in speech perception flexibility: converging evidence from ERPs and eye-tracking. *Brain Lang.* 223:105031. doi: 10.1016/j.bandl.2021.105031
- Kapnoula, E. C., Winn, M. B., Kong, E. J., Edwards, J., and McMurray, B. (2017). Evaluating the sources and functions of gradiency in phoneme categorization: an individual differences approach. *J. Exp. Psychol. Hum. Percept. Perform.* 43, 1594–1611. doi: 10.1037/xhp0000410
- Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Comm.* 27, 187–207. doi: 10.1016/S0167-6393(98)00085-5
- Kircher, R., and Kutlu, E. (2022). *Multilingual Realities, Monolingual Ideologies: Social Media Representations of Spanish as a Heritage Language in the United States*. Available at: <https://psyarxiv.com/y54ru/>
- Kluender, K. R. (1994). "Speech perception as a tractable problem in cognitive science," in *Handbook of Psycholinguistics*. ed. M. A. Gernsbacher (Academic Press), 173–217.
- Kong, E. J., and Edwards, J. (2011). Individual Differences in Speech Perception: Evidence from Visual Analogue Scaling and Eye-Tracking. In *Proceedings of the XVIIth International Congress of Phonetic Sciences*. Hong Kong.
- Kong, E. J., and Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *J. Phon.* 59, 40–57. doi: 10.1016/j.wocn.2016.08.006
- Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. Sci.* 9, F13–F21. doi: 10.1111/j.1467-7687.2006.00468.x
- Kutlu, E. (2020). Now you see me now you mishear me: Raciolinguistic accounts of speech perception in different English varieties. *J. Multiling. Multicult. Dev.*, 1–15. doi: 10.1080/01434632.2020.1835929
- Kutlu, E., and Kircher, R. (2021). A corpus-assisted discourse study of attitudes toward Spanish as a heritage language in Florida. *Languages* 6:38. doi: 10.3390/languages6010038
- Kutlu, E., Tiv, M., Wulff, S., and Titone, D. (2022). Does race impact speech perception? An account of accented speech in two different multilingual locales. *Cogn. Res. Princ. Implic.* 7, 1–16. doi: 10.1186/s41235-022-00354-0
- Kutlu, E., and Wiltshire, C. (2020). Where do negative stereotypes come from? The case of Indian English in the USA. *Proc. Linguist. Soc. Am.* 5:74. doi: 10.3765/plsa.v5i1.4669
- Levshina, N., Namboodiripad, S., Allasonnière-Tang, M., Kramer, M. A., Talamo, L., Verkerk, A., et al. (2021). Why we need a gradient approach to word order. *PsyArXiv [Preprint]*. doi: 10.31234/osf.io/yg9bf
- Lieberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* 54, 358–368. doi: 10.1037/h0044417
- Lieberman, A. M., and Whalen, D. H. (2000). On the relation of speech to language. *Trends Cogn. Sci.* 4, 187–196. doi: 10.1016/S1364-6613(00)01471-6
- Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling/b/versus/p/in trochees. *Lang. Speech* 29, 3–11. doi: 10.1177/002383098602900102
- Lisker, L., and Abramson, A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word* 20, 384–422. doi: 10.1080/00437956.1964.11659830
- Liu, L., and Kager, R. (2015). Bilingual exposure influences infant VOT perception. *Infant Behav. Dev.* 38, 27–36. doi: 10.1016/j.infbeh.2014.12.004
- López, B. G., Luque, A., and Piña-Watson, B. (2021). Context, intersectionality, and resilience: moving toward a more holistic study of bilingualism in cognitive science. *Cult. Divers. Ethn. Minor. Psychol.* doi: 10.1037/cdp0000472
- MacKaig, K. S., Best, C. T., and Strange, W. (1981). Categorical perception of English/r/and/l/by Japanese bilinguals. *Appl. Psycholinguist.* 2, 369–390. doi: 10.1017/S0142716400009796
- Massaro, D. W., and Cohen, M. M. (1983). Categorical or continuous speech perception: a new test. *Speech Comm.* 2, 15–35. doi: 10.1016/0167-6393(83)90061-4
- Mayo, L. H., Florentine, M., and Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *J. Speech Lang. Hear. Res.* 40, 686–693. doi: 10.1044/jslhr.4003.686
- McMurray, B. (2022). The acquisition of speech categories: beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. *Lang. Cogn. Neurosci.* doi: 10.1080/23273798.2022.2105367
- McMurray, B. (n.d.). The myth of categorical perception. *J. Acoust. Soc. Am.* Available at: <https://psyarxiv.com/dq7ej/>
- McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., and Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *J. Exp. Psychol. Hum. Percept. Perform.* 34, 1609–1631. doi: 10.1037/a0011747
- McMurray, B., Clayards, M. A., Tanenhaus, M. K., and Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychon. Bull. Rev.* 15, 1064–1071. doi: 10.3758/PBR.15.6.1064
- McMurray, B., Danelz, A., Rigler, H., and Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Dev. Psychol.* 54, 1472–1491. doi: 10.1037/dev0000542
- McMurray, B., and Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychol. Rev.* 118, 219–246. doi: 10.1037/a0022325
- McMurray, B., and Jongman, A. (2016). What comes after /f/? Prediction in speech derives from data-explanatory processes. *Psychol. Sci.* 27, 43–52. doi: 10.1177/0956797615609578
- McMurray, B., Samelson, V. M., Lee, S. H., and Bruce Tomblin, J. (2010). Individual differences in online spoken word recognition: implications for SLI. *Cogn. Psychol.* 60, 1–39. doi: 10.1016/j.cogpsych.2009.06.003
- McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition* 86, B33–B42. doi: 10.1016/S0010-0277(02)00157-9
- McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: evidence against phoneme-level inhibition. *J. Mem. Lang.* 60, 65–91. doi: 10.1016/j.jml.2008.07.002
- McQueen, J. (1996). Phonetic categorization. *Lang. Cogn. Proc.* 11, 655–664. doi: 10.1080/016909696387060
- Miller, J. L., and Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Percept. Psychophys.* 46, 505–512. doi: 10.3758/BF03208147

- Montrul, S., and Polinsky, M. (2021). *The Cambridge Handbook of Heritage Languages and Linguistics*. Cambridge, United Kingdom: Cambridge University Press.
- Oller, D. K., Pearson, B. Z., and Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Appl. Psycholinguist.* 28, 191–230. doi: 10.1017/S0142716407070117
- Ortega, L. (2020). The study of heritage language development from a bilingualism and social justice perspective. *Lang. Learn.* 70, 15–53. doi: 10.1111/lang.12347
- Pan, L., Ke, H., and Styles, S. J. (2022). Early linguistic experience shapes bilingual adults' hearing for phonemes in both languages. *Sci. Rep.* 12:4703. doi: 10.1038/s41598-022-08557-7
- Peltola, M. S., Tamminen, H., Toivonen, H., Kujala, T., and Näätänen, R. (2012). Different kinds of bilinguals—different kinds of brains: the neural organisation of two languages in one brain. *Brain Lang.* 121, 261–266. doi: 10.1016/j.bandl.2012.03.007
- Perkell, J. S., and Klatt, D. H. (2014). *Invariance and Variability in Speech Processes*. Psychology Press.
- Piske, T., MacKay, I. R., and Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *J. Phon.* 29, 191–215. doi: 10.1006/jpho.2001.0134
- Place, S., and Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child Dev.* 82, 1834–1849. doi: 10.1111/j.1467-8624.2011.01660.x
- Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Vol. 159. Cambridge, United Kingdom: Cambridge University Press.
- Rosa, J. D. (2016). Standardization, racialization, Languagelessness: Raciolinguistic ideologies across communicative contexts. *J. Linguist. Anthropol.* 26, 162–183. doi: 10.1111/jola.12116
- Sarrett, M., McMurray, B., and Kapnoula, E. (2020). Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and semantic expectations. *Brain Lang.* 211:104875. doi: 10.1016/j.bandl.2020.104875
- Schouten, M. E. H., Gerrits, E., and Van Helsen, A. (2003). The end of categorical perception as we know it. *Speech Comm.* 41, 71–80. doi: 10.1016/S0167-6393(02)00094-8
- Sebastián-Gallés, N., and Bosch, L. (2002). Building phonotactic knowledge in bilinguals: role of early exposure. *J. Exp. Psychol. Hum. Percept. Perform.* 28, 974–989. doi: 10.1037/0096-1523.28.4.974
- Sebastián-Gallés, N., and Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: is the distributional account all there is to it? *Dev. Sci.* 12, 874–887. doi: 10.1111/j.1467-7687.2009.00829.x
- Serniclaes, W. (2006). Allophonic perception in developmental dyslexia: origin, reliability and implications of the categorical perception deficit. *Written Lang. Lit.* 9, 135–152. doi: 10.1075/wll.9.1.09ser
- Slawinski, E. B., and Fitzgerald, L. K. (1998). Perceptual development of the categorization of the/r-w/contrast in normal children. *J. Phon.* 26, 27–43. doi: 10.1006/jpho.1997.0057
- Stölen, K., Abrahamsson, N., and Hyltenstam, K. (2014). Effects of age of learning on voice onset time: categorical perception of Swedish stops by near-native L2 speakers. *Lang. Speech* 57, 425–450. doi: 10.1177/0023830913508760
- Strange, W., and Shafer, V. L. (2008). Speech perception in second language learners: the re-education of selective perception. *Phonol. Second Lang. Acquis.* 36, 153–192. doi: 10.1075/sibil.36.09str
- Surraín, S., and Luk, G. (2019). Describing bilinguals: a systematic review of labels and descriptions used in the literature between 2005–2015. *Bilingualism Lang. Cogn.* 22, 401–415. doi: 10.1017/S1366728917000682
- Szostak, C. M., and Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Atten. Percept. Psychophys.* 75, 1533–1546. doi: 10.3758/s13414-013-0492-3
- Theodore, R. M., and Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychon. Bull. Rev.* 26, 985–992. doi: 10.3758/s13423-018-1551-5
- Titone, D. A., and Tiv, M. (2022). Rethinking multilingual experience through a systems framework of bilingualism. *Bilingualism Lang. Cogn.* 1–16. doi: 10.1017/S1366728921001127
- Tiv, M., Kutlu, E., Gullifer, J. W., Feng, R. Y., Doucerain, M. M., and Titone, D. A. (2022). Bridging interpersonal and ecological dynamics of cognition through a systems framework of bilingualism. *J. Exp. Psychol. Gen.* 151, 2128–2143. doi: 10.1037/xge0001174
- Tiv, M., Kutlu, E., and Titone, D. (2021). “Bilingualism moves us beyond the ideal speaker narrative in cognitive psychology,” in *Bilingualism Across the Lifespan*. ed. W. S. Francis (Routledge), 29–46.
- Toscano, J. C., and McMurray, B. (2010). Cue integration with categories: weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 434–464. doi: 10.1111/j.1551-6709.2009.01077.x
- Toscano, J. C., McMurray, B., Dennhardt, J., and Luck, S. J. (2010). Continuous perception and graded categorization electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychol. Sci.* 21, 1532–1540. doi: 10.1177/0956797610384142
- Werker, J. F., and Curtin, S. (2005). PRIMIR: a developmental framework of infant speech processing. *Lang. Learn. Dev.* 1, 197–234. doi: 10.1080/15475441.2005.9684216
- Werker, J. F., Gilbert, J. H. V., Humphrey, K., and Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child Dev.* 52, 349–355. doi: 10.2307/1129249
- Werker, J. F., and Tees, R. C. (1984). Cross-language speech perception: evidence for perceptual reorganization during the first year of life. *Infant Behav. Dev.* 7, 49–63. doi: 10.1016/S0163-6383(84)80022-3
- Werker, J. F., and Tees, R. C. (1987). Speech perception in severely disabled and average reading children. *Can. J. Psychol.* 41, 48–61. doi: 10.1037/h0084150
- Werker, J. F., Yeung, H. H., and Yoshida, K. A. (2012). How do infants become experts at native-speech perception? *Curr. Dir. Psychol. Sci.* 21, 221–226. doi: 10.1177/0963721412449459
- Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., et al. (2022). Heritage speakers as part of the native language continuum. *Front. Psychol.* 12:5982. doi: 10.3389/fpsyg.2021.717973



OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway,
Norway

REVIEWED BY

Derrin Pinto,
University of St. Thomas, United States
Irina Dubinina,
Brandeis University,
United States

*CORRESPONDENCE

Sagit Bar On
shayelbar@gmail.com
Natalia Meir
natalia.meir@biu.ac.il

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 12 August 2022

ACCEPTED 24 November 2022

PUBLISHED 22 December 2022

CITATION

Bar On S and Meir N (2022) Requests and
apologies in two languages among
bilingual speakers: A comparison of
heritage English speakers and English- and
Hebrew-dominant bilinguals.
Front. Psychol. 13:1017715.
doi: 10.3389/fpsyg.2022.1017715

COPYRIGHT

© 2022 Bar On and Meir. This is an open-
access article distributed under the terms
of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that
the original publication in this journal is
cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Requests and apologies in two languages among bilingual speakers: A comparison of heritage English speakers and English- and Hebrew-dominant bilinguals

Sagit Bar On^{1*} and Natalia Meir^{1,2*}

¹Department of English Literature and Linguistics, Bar Ilan University, Ramat Gan, Israel, ²The Gonda Multidisciplinary Brain Research Center, Bar-Ilán University, Ramat Gan, Israel

Introduction: Linguistic research over the last two decades has uncovered a significant number of properties that identify heritage language (HL) speakers as a particular phenomenon within bilingualism. However, despite the rapid development of HL research, the sphere of HL speech act pragmatics is still in its infancy.

Methodology: The current study sought to cover part of this gap by investigating both languages of HL adult speakers for whom English is their HL and Hebrew is their dominant societal language (SL; $n = 20$, hereafter HS) in comparison with two other groups: Hebrew-dominant speakers who were born to Hebrew-speaking families and raised in Israel, and thus English is the language of their schooling ($n = 20$, hereafter HEB-D), and English-dominant speakers who were born to English-speaking families and immigrated to Israel from an English-speaking country after the age of 18, and thus Hebrew is their L2 ($n = 20$, hereafter ENG-D). The discourse-pragmatic tasks in English and in Hebrew consisted of the same 36 scenarios eliciting requests and apologies in each language. Each request was followed by an apology that is related.

Results: The results indicated that Hebrew-dominant speakers and English-dominant speakers, i.e., HEB-D and ENG-D, had different strategies for the realization of requests and apologies which they systematically transferred from their dominant language into their weaker one, confirming the cross-cultural and cross-linguistic differences between request and apology realizations in English and in Hebrew. However, the picture was more complex for the HL speakers in their HL-English and SL-Hebrew as in some cases their strategies paired up with the ENG-D in English, and with the HEB-D in Hebrew, while in other cases they developed a unique and hybrid linguistic style reflecting the strategies of both languages.

Discussion: The investigation of both languages of HL speakers enabled us to compare features of the two linguistic systems and make conclusions about their nature.

KEYWORDS

heritage language, L2, English, Hebrew, pragmatics

Introduction

Despite the rapid development of Heritage Language (HL) research, the sphere of speech act pragmatics has remained uncharted territory. The current study is devised to investigate pragmatic abilities of HL speakers of English who have acquired their HL-English in contact with Hebrew as the dominant SL. The pragmatic abilities of the HL-English speakers are investigated *via* comparing their request and apology realizations in English to English-dominant speakers (L1-English, L2-Hebrew), and in Hebrew to Hebrew-dominant speakers (L1-Hebrew, L2-English). The investigation of both languages, HL-English and SL-Hebrew, is expected to provide a more comprehensive picture of the pragmatic abilities of this unique bilingual group. It is important to note that the HL-speakers in this study might be more balanced bilinguals than prototypical HL speakers due to the fact that English is a mandatory subject in the Israeli national education system.

HL speakers are bilinguals who are exposed to their HL from birth *via* naturalistic input at home, yet the HL is not the dominant language of the larger surrounding society (Rothman, 2009; Montrul, 2018; Polinsky, 2018). Thus, HL speakers use and acquire their HL in a socio-linguistically, socio-culturally, and socio-politically complex situation (Xiao-Desai, 2019). HL speakers are considered to be asymmetrical bilinguals since they learned their HL as a first language in childhood at home, but, as adults, they become dominant in the majority societal language (SL) spoken by the community (Elabbas et al., 2013). There are two main factors which are reported to influence the linguistic systems of HL speakers: HL input characteristics and cross-linguistic influence (Montrul, 2016; Polinsky, 2018). Input relates to the timing, amount, and content of the exposure to the languages, while cross-linguistic influence is concerned with the effect that one language has on the other.

Numerous studies have discussed connections between language, culture, and identity (Wallace, 2004; Berard, 2005; Achugar, 2006). HL speakers belong simultaneously to two socio-linguistic and socio-cultural communities. The core identity of HL speakers involves the process of constant negotiation and self-positioning within a bilingual and bicultural environment (Val and Vinogradova, 2010). Rothman (2009) claims that HL speakers feel strong pressure to assimilate to the mainstream culture, and therefore, gradually begin to use their SL more and more at home. As a result, patterns of the HL gradually change and become modified in quality. Thus, by the time HL children reach adolescence and young adulthood, their HL might resemble, in some aspects, an L2 learned in adulthood as opposed to L1 acquired in childhood.

Most studies in interlanguage pragmatics investigate influences from L1 to L2, and do not consider a bi-directional

relationship between the two languages (Cenoz, 2003). However, Blum-Kulka (1990) and Blum-Kulka and Sheffer (1993) have found that the realization patterns of requests produced in English (L1) and in Hebrew (L2) by American immigrants to Israel who were fully competent in the two languages differed significantly from both the Israeli and the American patterns. Their requests presented features that can be situated in between American and Israeli requests. Blum-Kulka (1991) proposes the 'Intercultural Style Hypothesis' to define the development of an intercultural pattern that reflects bi-directional interactions between the languages. Cenoz (2003) found that native Spanish speakers who achieved a high level of proficiency in English developed an intercultural pattern that was reflected both in the similarity between their requests production in Spanish and in English and in the differences between these requests and those formulated by other native speakers of Spanish. These findings support the 'Intercultural Style Hypothesis' and show that the interaction between the languages of a proficient multilingual speaker is bi-directional. Research investigating immigrants in other countries is also compatible with this hypothesis. For example, Su (2010) investigated Chinese learners of English requesting behavior in both their L1 and L2 and found a bi-directional transfer at a pragmatic level. Kecskes and Papp (2000) also examined the influence of foreign language learning on the development of mother tongue skills from a cognitive-pragmatic perspective and found evidence that foreign language acquisition can influence different areas of L1 when exposure to the foreign language is intensive. However, as Kasper and Blum-Kulka (1993) point out, more research is needed in the field of intercultural pragmatics.

Cross-linguistic transfer and shrinkage of grammatical structures have been related to economy principles under which the mind favors the least amount of effort toward a cognitive task. Therefore, humans organize knowledge in their brain by dividing it into 'classes' and/or match patterns to reduce cognitive load. This means that HL speakers might resort to simpler structures that overlap in both languages (Rothman, 2015; Polinsky and Scontras, 2020) and/or develop a unique and hybrid intercultural linguistic style reflecting both their SL and their HL (Pinto and Raschio, 2007; Xiao-Desai, 2019) in order to reduce the cognitive load associated with being bilinguals, e.g., inhibition costs.

Most of the previous HL research has addressed morpho-syntactic competence of HL speakers (for an extensive overview see Montrul, 2018; Polinsky, 2018), whereas the available research on pragmatic abilities of HL speakers is still very limited and covers mainly Spanish and Russian as HLs (see Pinto and Raschio, 2007; Dubinina, 2011; Dubinina and Malamud, 2017; Xiao-Desai, 2019).

Request strategies: Differences between English and Hebrew speakers

Requests are by definition face-threatening acts. The notion of 'face' can be divided into two concepts: positive face and negative face. Positive face is the need to be accepted, liked by others, treated as a member of the group, and know that his/her wants are shared by others. Negative face is the need to be independent to have freedom of action, and not to be imposed upon by others. By making a request, the speaker impinges on the hearer's right to freedom of action and from imposition. Therefore, requests are considered to be face-threatening acts in which the negative face of the hearer is threatened, and when 'face' is threatened speakers typically act to mitigate that threat by doing 'facework' (Brown and Levinson, 1987).

Requests have been divided in the literature into three main segments: Alerters, head acts, and supportive moves. Alerters are opening elements that precede the actual request and are primarily used to get the hearer's attention. They are optional to the realization of the request and can come in the form of address terms or attention-getters (Blum-Kulka and Olshtain, 1984). Head acts are the core part of the request sequence which realize a request independently of other elements. The head acts are the actual requests and serve as an integral part of any request. A request might contain more than one head act (Bella, 2012). Supportive moves are adjuncts to the head acts used to modify the impact or force of requests. However, there are cross-linguistic and cross-cultural differences in request realizations. For example, one way in which the speaker can soften the imposition is by choosing an indirect strategy over a direct one (Blum-Kulka and Olshtain, 1984).

In both English and Hebrew, head acts of requests can be grammatically realized with imperatives, interrogatives, and declaratives (Blum-Kulka and Olshtain, 1984; Curl and Drew, 2008). However, even though in English direct imperatives are usually defined as appropriate constructions for commands or instructions, they are less appropriate for making requests (Márquez-Reiter, 2000). Searle (1975) claims that the "ordinary conversational requirements of politeness normally make it awkward to issue flat imperatives, and we therefore seek to find indirect means to our illocutionary needs." Blum-Kulka and Olshtain (1984) explain that even though imperatives in Hebrew are considered appropriate for requesting, they are the most direct and explicit level among the syntactic structures available for making requests, and therefore the least polite constructions. Olshtain and Blum-Kulka (1985) noted that in an acceptability judgment test which they developed, English speakers who were learners of L2 Hebrew (and living in Israel for a period of less than 2 years) were reluctant to accept the direct request strategy found in Hebrew native speech. For example, the L2 learners were inclined to reject the Hebrew equivalent of the declarative 'I hope you can take me back to town' when asking for a ride, whereas native Hebrew speakers had no problem accepting it. The learners have responded in accordance with the politeness norms for requesting in their L1 (the most common structure to realize a

request in English is to use an interrogative structure combined with a modal verb).

Modals can be used with both declarative and interrogative sentences (Walters, 1979). However, since linguistic expressions of modality convey speakers' claims about the permission, ability, probability, possibility, etc. of beliefs and actions, and therefore have a notion of indirectness to them (Turnbull and Saxton, 1997) there are likely to be more frequent in an English speech than in Hebrew.

Although 'please' (and its equivalents) is a universal mitigating device (Ogiermann, 2009; Murphy and De Felice, 2018; Webman-Shafran, 2019), there are cross-cultural and cross-linguistic differences with respect to its usage. Dubinina and Malamud (2017) suggest that the marker 'please' in English can be used only in conventional requests, both direct (e.g., 'Do this exercise, please!') and indirect ('Could you open the door please?'). However, it is not allowed in utterances that do not have the form conventionally used for requests, even when their form and propositional content are similar to conventional requests, and even if these are ultimately interpreted as requests [e.g., 'Are you able to open the door (#please)?']. Nonetheless, according to House (1989) this politeness marker is most appropriate in mitigating 'standard situations' where the request making and the fulfillment of it are self-evident and the function of the request is clear. However, when speakers prefer to disguise the function of the request in the form of a question, they tend to leave out the 'please' marker in order to allow the addressee to respond to the propositional content of the utterance and not reveal its conventional function.

To sum up, in English, making a request short and straightforward is considered impolite and face threatening. Therefore, in order to save 'face' English speakers prefer to use a longer and indirect version of a request by applying interrogatives and modals. However, mitigating a request by using an interrogative structure (i.e., a question) with a modal makes the use of 'please' somewhat redundant. Hebrew speakers, on the other hand, prefer a more straightforward strategy in the form of a declarative (statement). Yet, in order to mitigate 'face threatening', and as the intent of requesting is already visible, they apply the marker 'bevakasha' ('please').

Apology strategies: Differences between English and Hebrew speakers

Apologies are also considered face-threatening acts (Brown and Levinson, 1987), however, contrary to requests which are 'pre-event acts', apologies are 'post-event acts'. In other words, while requests are made to cause an event or to change one, apologies signal an event of a social norm violation that has already taken place which the speaker holds himself/herself at least partially accountable for (Leech, 1980; Blum-Kulka and Olshtain, 1984). Therefore, as opposed to requests, apologies are face-threatening acts in which the negative face of the speaker is threatened.

The speech act of apology is universal, yet its strategies and linguistic variations differ cross-culturally to a great extent (Israa, 2017). Blum-Kulka and Olshtain (1984) explain that a strategy for the act of apologizing can come in one of two basic forms (or a combination of both): (i) Direct realization of an apology *via* an explicit illocutionary force-indicating device (hereafter IFID), i.e., *via* performative expressions such as ‘sorry’, ‘apologize’, ‘forgive’, or ‘pardon’, and (ii) Indirect realization of an apology *via* the use of an utterance which contains reference to one or more elements from a closed set of four specified propositions, i.e., explaining the cause, acknowledging responsibility, offering repair, and promising forbearance.

Cohen and Olshtain (1981) investigated apologies of Hebrew speakers that learned English as an L2, and discovered that the L2-English learners with L1-Hebrew did not seem to be familiar with the accepted formulas needed for the apologies in English. To be more specific, Hebrew speakers learners of English were less likely to accept responsibility for an offense or to make an offer for repair than native English speakers. This is in line with Mills and Grainger’s (2016) claim that in Hebrew there is a tendency for directness to be evaluated positively as part of the Israeli cultural style, in contrast to Anglo-European norms which are indirect and often characterized as overly mannered. For instance, they found that Hebrew speakers were comfortable with direct statements in a way that British English speakers were not as a result of conceptual ideology about directness which Hebrew speakers perceive as signaling honesty and friendliness. Ellis and Maoz (2002) add that Hebrew speech is characterized as direct and “to the point,” and is used both within the culture and during intercultural communication. This ‘dugri’ (straightforward) style of the Israeli culture enables Hebrew speakers to use an explicitness about intentions that in other cultures could be considered offensive (Katriel, 1986). Thus, bearing in mind the directness that characterizes Hebrew, while also taking into consideration Cohen and Olshtain’s (1981) findings, it should not be surprising that Hebrew speakers prefer to use the direct strategy of expressing an explicit IFID, while English speakers prefer to use the indirect strategy of choosing one or more of the other indirect propositions.

In addition to these main strategies, which make up the speech act of the apology itself, there are ways in which the speaker can modify the apology, e.g., by performing it with different levels of intensity using intensification terms such as ‘so’, ‘very’, ‘really’, ‘terribly’, ‘extremely’, ‘totally’, ‘deeply’, and ‘highly’. Previous studies show that there are differences across speakers of different languages in the use of intensification. For example, Cohen et al. (1986) found that native Hebrew speakers who were learners of L2-English intensified their apologies significantly more than native English speakers, even though this extra intensity on the part of the Hebrew speakers was not necessarily warranted given the generally low or moderate severity of the offense. The same trend was found in Olshtain (1983) where English and Russian learners of Hebrew did not intensify apologies in a target-like manner.

To sum up, English speakers rely more on indirect strategies compared to Hebrew speakers. In regard to apologies this means that English speakers prefer the less direct strategy of propositions, while Hebrew speakers prefer the direct strategy of IFIDs. Furthermore, the few studies that have investigated the usage of adverbial intensifiers in English and in Hebrew show that Hebrew speakers tend to intensify their apologies more than English speakers.

The current study

This study aimed at providing a comprehensive overview of HL-English speakers’ linguistic behavior by comparing their realization patterns of two speech acts, requests and apologies, in both of their languages, i.e., English and Hebrew, to two groups with varying level of dominance: Hebrew-dominant speakers and English-dominant speakers. Research comparing both languages of HL speakers is limited. To the best of our knowledge, only three studies investigated both languages of HL speakers: Scontras et al. (2017) investigated the interpretation of ‘every’ in HL-Mandarin and SL-English; Kupisch et al. (2014) investigated accentedness in two different HL-SL dyads, German-French and German-Italian; and Stangen et al. (2015) investigated accentedness in HL-Turkish and SL-German. A comparison of requests and apologies in the two languages of bilingual subjects is expected to shed light on the mechanisms of cross-linguistic and cross-cultural realization of politeness under diminished input (quantitatively and qualitatively), and adds to the growing body of literature concerning politeness and speech acts in English and in Hebrew. Research on HL-English is rather limited (Polinsky, 2018), as English is mainly studied as the L2. Yet Israel offers a rare opportunity to investigate HL-English. The uniqueness of English as the HL is that it is the *de facto* international language of the modern world, on the one hand, but it also patterns with other HLs as it undergoes divergence in contact situations where it is a minority language (Meir et al., 2021).

The study aimed to answer to what extent request and apology strategies produced in English and in Hebrew by HL-English speakers differ from or resemble the ones produced by the English-dominant speakers (i.e., speakers who were born to English-speaking families and raised in an English-speaking country) and/or by the Hebrew-dominant speakers (i.e., speakers who were born to Hebrew-speaking families and raised in Israel).

For this objective, three hypotheses were formulated: (H1) HL speakers will show deviation in the production of realization patterns and carry over pragmatic and socio-linguistic behavior from their HL-English into their SL-Hebrew, (H2) HL speakers will show deviation in the production of realization patterns and carry over pragmatic and socio-linguistic behavior from their SL-Hebrew into their HL-English, and (H3) HL speakers will show a hybrid pragmatic competence, i.e., they will show evidence of developing new conventions for the production of realization

patterns, which differ from the corresponding conventions of the other two groups.

Methodology

Participants

The study comprised three groups of adult bilingual speakers between the ages of 23–30: (i) HL-English bilinguals who were dominant in Hebrew, born to English-speaking families but raised in Israel from birth or arrived in Israel before the age of five (HS); (ii) Hebrew-dominant bilinguals, born to Hebrew-speaking families and raised in Israel, and thus English is the language of their schooling (HEB-D); and (iii) English-dominant bilinguals, born to English-speaking families and immigrated to Israel from an English-speaking country after the age of 18 (ENG-D). All groups were tested in both English and Hebrew. All three groups were balanced in regard to the number of participants and the gender, i.e., ten females and ten males were tested in each group (a total of sixty participants in the study). All participants were of medium to high socio-economic status as suggested by the level of their education. The minimum age of 23 was chosen in order to make sure that the participants in the ENG-D group would have sufficient amount of proficiency in Hebrew to complete the tasks.

All participants were asked to fill in a self-report background questionnaire eliciting the participants' demographic information such as age, gender, level of education, occupation, birthplace, place of residence, etc., as well as language-related information such as age of onset (the age at which each language was acquired or learned), proficiency in all four domains of language (reading, writing, comprehending, and speaking), frequency of usage, and non-native accent ratings. The participants' demographic data are shown in Table 1, and the self-evaluated language information is shown in Table 2. Tables 1 and 2 present data on the participants per group, per language, and statistics for group differences.

In order to obtain a more direct measure of the participants' language proficiency in English and in Hebrew, the vocabulary size in both languages was tested using the Multilingual Naming Test (hereafter MINT; Gollan et al., 2012). The MINT has been validated as an objective proficiency measure for bilingual speakers who speak any combination of English, Spanish,

Mandarin, and Hebrew (Tomoschuk et al., 2019). The task included the same 68 black and white picture stimuli. In this task, the participants were asked to say out loud the name of the object they saw in the picture, once in English and once in Hebrew. Each response was coded as correct or incorrect. In line with the profile presented by the participants *via* the questionnaire, the results of the MINT task showed that the HS group paired up with the HEB-D group with respect to their vocabulary size in Hebrew, yet they showed lower proficiency than the ENG-D group in English (even though much higher than HEB-D). Table 3 presents data of the MINT scores of the participants per group, per language, and group comparisons using one-way ANOVAs.

The pragmatic task

A discourse-pragmatic task was designed for this study to elicit requests and apologies. The parallel tasks in English and in Hebrew consisted of 18 scenarios eliciting requests and 18 scenarios eliciting apologies in each language (a total of 36 scenarios in each language). The scenarios were arranged in a random order with pictures showing whether the scenario addressed a female or a male. Each request was followed by an apology that is related (for similar procedure see Márquez-Reiter, 2000). In this task, the participants were asked to say out loud what they would have said if they had been one of the participants in the actual situation while being as spontaneous as possible. In order to take into account variations of the participants' English and Hebrew literacy skills and to ensure that reading proficiency would not affect the results, the scenarios were read aloud by the experimenter (for similar procedure see Walters, 1979, 1980, 1981; Rintell, 2009 and Dubinina and Malamud, 2017). Examples of the scenarios from the task are presented in Table 4, and the entire list of stimuli is presented in Appendix A. The task manipulated 'Social Status' (i.e., the relative level of respect, honor, and deference) and 'Social Distance' (i.e., the level of familiarity between the participants), and controlled for gender.

Coding schemata

The coding schemata adopted in the current study were developed based on several former studies (Cohen and Olshtain,

TABLE 1 Participants' demographic data [Mean (SD)].

	HS (N = 20)	HEB-D (N = 20)	ENG-D (N = 20)	Group differences	Tukey HSD <i>Post hoc</i> analysis for multiple comparisons
Age (Years)	25.4 (2.4)	25.7 (2.4)	26.3 (2.8)	$F(2,57) = 0.545, p = 0.583$	n/a
Gender	F = 10 M = 10	F = 10 M = 10	F = 10 M = 10	$\chi^2 = 0, p = 1.00$	n/a
Education (Years)	15.3 (1.8)	15.5 (1.8)	14.9 (0.9)	$F(2,57) = 0.725, p = 0.489$	n/a
Immigration to Israel (Age)	1.45 (2)	0 (0)	18.3 (0.6)	$F(2,57) = 1,297, p < 0.001$	HEB-D < HS < ENG-D
Age of Onset of English	0 (0)	9 (0)	0 (0)	$F(2,57) = 180.7, p < 0.001$	(ENG-D=HS) < HEB-D
Age of Onset of Hebrew	0.8 (1.6)	0 (0)	14.4 (4.2)	$F(2,57) = 180.7, p < 0.001$	(HEB-D=HS) < ENG-D

TABLE 2 Participants' subjective ratings of proficiency [Mean (SD)].

	HS (N = 20)	HEB-D (N = 20)	ENG-D (N = 20)	Group differences	Tukey HSD <i>Post hoc</i> analysis for multiple comparisons
English proficiency in reading (1–7)	6.4 (0.5)	4.2 (0.6)	7 (0)	$F(2,57) = 163.8, p < 0.001$	HEB-D < HS < ENG-D
English proficiency in writing (1–7)	6.3 (0.5)	3.8 (0.7)	7 (0)	$F(2,57) = 186.2, p < 0.001$	HEB-D < HS < ENG-D
English proficiency in comprehending (1–7)	7 (0)	5.5 (0.4)	7 (0)	$F(2,57) = 161.4, p < 0.001$	HEB-D < (ENG-D=HS)
English proficiency in speaking (1–7)	6.6 (0.4)	4.5 (0.4)	7 (0)	$F(2,57) = 202, p < 0.001$	HEB-D < HS < ENG-D
Total English proficiency (4–28)	26.3 (1.4)	18.2 (0)	28 (0)	$F(2,57) = 249.3, p < 0.001$	HEB-D < HS < ENG-D
Hebrew proficiency in reading (1–7)	7 (0)	7 (0)	4.2 (0.8)	$F(2,57) = 209, p < 0.001$	ENG-D < (HEB-D=HS)
Hebrew proficiency in writing (1–7)	7 (0)	7 (0)	4 (0.8)	$F(2,57) = 255.4, p < 0.001$	ENG-D < (HEB-D=HS)
Hebrew proficiency in comprehending (1–7)	7 (0)	7 (0)	6.1 (0.5)	$F(2,57) = 53.07, p < 0.001$	ENG-D < (HEB-D=HS)
Hebrew proficiency in speaking (1–7)	7 (0)	7 (0)	5.3 (0.8)	$F(2,57) = 77.34, p < 0.001$	ENG-D < (HEB-D=HS)
Total Hebrew proficiency (4–28)	28 (0)	28 (0)	19.7 (2.7)	$F(2,57) = 169.8, p < 0.001$	ENG-D < (HEB-D=HS)
Degree of accentedness in English (1–7)	1.3 (0.4)	6.2 (0.6)	1 (0)	$F(2,57) = 829.2, p < 0.001$	(HS = ENG-D) < HEB-D
Degree of accentedness in Hebrew (1–7)	1 (0)	1(0)	5 (0.9)	$F(2,57) = 328.9, p < 0.001$	(HS=HEB-D) < ENG-D
Current exposure to English (0–100)	29 (7.6)	0 (0)	69 (7.6)	$F(2,57) = 579.8, p < 0.001$	HEB-D < HS < ENG-D
Current exposure to Hebrew (0–100)	71 (7.6)	100 (0)	31 (7.6)	$F(2,57) = 579.8, p < 0.001$	ENG-D < HS < HEB-D

TABLE 3 MINT task performance per group per language (Mean (SD)).

	HS (N = 20)	HEB-D (N = 20)	ENG-D (N = 20)	Group differences	Tukey HSD <i>Post hoc</i> analysis for multiple comparisons
MINT-English	57.4 (3.5)	38.6 (10.3)	64.9 (2.3)	$F(2,57) = 82.93, p < 0.001$	HEB-D > HS > ENG-D
MINT-Hebrew	58.4 (2.8)	61.4 (2.7)	40.3 (11.6)	$F(2,57) = 49.05, p < 0.001$	ENG-D > (HS=HEB-D)

TABLE 4 Examples of scenarios.

The speech act	The preamble
Request 1	You are a student. You conducted research for a seminar paper, but you do not know how to do the statistics. You know that your elderly neighbor is very good at it, and you want to ask him for an hour of his time to help you. You see him in his garden. What do you say to him?
Apology 1	You are a student. You asked your elderly neighbor to help you with statistics for an hour. However, you forgot to show up on time and you are an hour late. What do you say to him?
Request 2	You are on a bus with a child. There are plenty of seats on the bus but there are not any for two people together. You want to ask a passenger who is sitting on her own to change seats with you so that you can sit next to the child. What do you say to her?
Apology 2	A passenger has agreed to change seats with you so that you can be next to a child on the bus. While changing seats you accidentally step on her foot. What do you say to her?

1981; Blum-Kulka and Olshtain, 1984; Dubinina and Malamud, 2017; Israa, 2017).

There are syntactic, lexical, and structural variations in the production of requests and apologies across different languages. In the current study, the participants' requests were analyzed for the presence of alerters, head acts, and supportive moves. Furthermore, we also noted the syntactic structure of the head act/s (interrogative/imperative/declarative/mixed-when a request contained two head acts with different syntactic structures), the use of modals, and the use of 'please'/'bevakasha'. The participants' apologies were analyzed for expressions of apology, number of propositions added (i.e., offering explanation, taking responsibility, offering repair or

compensation, and promising forbearance), and the use of adverbial intensifiers.

Procedure

The participants in the study were recruited by word of mouth through personal social networks. Prior to the data collection, all participants who agreed to volunteer were given a recruitment letter (each in his/her dominant language) explaining the general aim of the study without revealing its specific aim. Upon agreeing to take part in this research, each participant was allocated a personal participant's code and was asked to fill in a self-report

Google Form questionnaire in his/her dominant language using that code for identification. In order to obtain a broad-based sample from a variety of geographical locations some of the participants were tested *via* Zoom. Each participant was tested individually in both English and Hebrew. Testing in each group had been counterbalanced, i.e., ten participants completed the tasks in English followed by Hebrew, and ten participants completed the tasks in Hebrew followed by English. The elicitation tasks were audio recorded for subsequent transcription and coding purposes. The administration of all tasks took approximately an hour.

Results

In order to determine to what extent request and apology strategies of HL-English speakers differ and/or resemble those of dominant speakers in HL-English and in SL-Hebrew, we fitted mixed-effects logistic regression models (Baayen et al., 2008) with lme4 package (Bates et al., 2015) since the responses were mainly coded in a binary manner (1 = present, 0 = absent), i.e., syntactic structure, the use of modals, the use of *'please'/'bevakasha'*, the use of IFIDs, and the use of adverbial intensifiers. We tested the contribution of Language (English, Hebrew) together with Group (ENG-D, HEB-D, HS); these variables and their interactions were entered as fixed effects. To account for the variability within participants and scenarios, the models included crossed random intercepts for Participant and Scenario. Fitted models were compared in terms of Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC), with reduced AIC and BIC values indicating a better model fit. This was supplemented by Likelihood ratio tests conducted to determine if the inclusion of

a predictor significantly improved the model fit. First, we examined whether the inclusion of the random effects was permitted. This was done by comparing a baseline generalized linear model without the random intercepts (null model) with a baseline mixed-effects model that only included the random intercepts. Next, we implemented a step-wise-step-up procedure for building the mixed-effects model. The significance level of the main fixed effects was obtained using the ANOVA function. We obtained the estimated marginal means (EMM) for all pairwise comparisons using Tukey's HSD adjustment for multiple comparisons. In the results subsection, we report the final models which were found to provide the best and most parsimonious fit for the data. We fitted linear models for the analysis of the number of propositions in apologies in English and in Hebrew, as the data were coded in a non-binary manner (0–4). We also ran models with 'Social Status' and 'Social Distance' as fixed variables, yet the inclusion of these effects and their interactions with Group did not improve the fit of the models. In the Limitations subsection, we outline possible reasons for that.

Requests

Alerters, head acts, and supportive moves

Table 5 presents the descriptive statistics for the use of alerters, head acts, and supportive moves coded as 1 = response with an alerter/supportive move and 0 = response without an alerter/supportive move, and 1 = response with more than one head act and 0 = response with one head act.

Table 6 presents the final model for the use of alerters and head acts. For the use of alerters, the results showed that there

TABLE 5 Alerters, head acts, and supportive moves per group per language [Mean (SD)].

	English			Hebrew		
	HS	HEB-D	ENG-D	HS	HEB-D	ENG-D
Alerters	0.71 (0.45)	0.56 (0.49)	0.43 (0.49)	0.71 (0.45)	0.58 (0.49)	0.38 (0.48)
Number of head acts	1.35 (0.47)	1.07 (0.25)	1.03 (0.19)	1.39 (0.48)	1.07 (0.26)	1.03 (0.19)
Supportive moves	0.75 (0.43)	0.54 (0.49)	0.70 (0.45)	0.72 (0.44)	0.69 (0.46)	0.60 (0.49)

TABLE 6 Final models for the use of alerters and head acts.

	Alerters				Head acts			
	Fixed effects							
	Estimate	Std. error	z Value	Pr(> z)	Estimate	Std. error	z Value	Pr(> z)
Intercept	−0.8217	0.5732	−1.434	0.151706	−3.6635	0.3373	−10.861	< 2e-16***
Group (ENG-D vs. HEB-D)	1.4199	0.7329	1.937	0.052687	0.6959	0.4137	1.682	0.0925
Group (ENG-D vs. HS)	2.5860	0.7425	3.483	0.000496***	3.0097	0.3933	7.653	1.97e-14***
Random effects: Number of observations: 2160, Participant: 60; Scenario: 18								
	Variance	Std. Dev			Variance	Std. Dev		
Participant (Intercept)	5.064	2.250			0.9365	0.9678		
Scenario (Intercept)	1.047	1.023			0.2098	0.4580		

Significance codes: 0 '***' 0.001.

was an effect of Group, yet no effect of Language and no Group*Language interaction. Follow-up pairwise Group contrasts for the use of alerters showed that there were significant differences between ENG-D and HS ($\beta = -2.59$; $SE = 0.743$; $Z = -3.483$, $p = 0.0014$), yet, there were no significant differences between ENG-D and HEB-D ($\beta = -1.42$; $SE = 0.733$; $Z = -1.937$, $p = 0.1282$) and between HEB-D and HS ($\beta = -1.17$; $SE = 0.737$; $Z = -1.581$, $p = 0.2538$).

The results for the use of head acts indicated an effect of Group, yet no effect of Language and no Group*Language interaction. Follow-up pairwise Group contrasts for the use of head acts showed that there were significant differences between ENG-D and HS ($\beta = -3.010$; $SE = 0.393$; $Z = -7.653$, $p < 0.0001$), and between HEB-D and HS ($\beta = -2.314$; $SE = 0.366$; $Z = -6.320$, $p < 0.0001$), yet, there were no significant differences between ENG-D and HEB-D ($\beta = -0.696$; $SE = 0.414$; $Z = -1.682$, $p = 0.2120$).

No Group and Language differences, and no Group*Language interaction were found in regard to supportive moves.

Syntactic structure choice of the head acts

To address the choice of a syntactic structure in requests, we coded participants' head acts as declarative, interrogative, imperative, or mixed (if there were two head acts with different syntactic structures in one request scenario). Figure 1 presents the syntactic structure of the head acts used across the three groups in English and in Hebrew.

Table 7 presents the final models for the choice of the specific syntactic structure of head acts separately for interrogative, declarative, and mixed structures respectively, coded as 1 = present and 0 = absent. It is important to note that the usage of imperatives in all groups in both languages was virtually nonexistent and therefore responses in imperative forms were not analyzed statistically. Out of 2,160 scenarios, 15 instances of imperatives were found in the English data (13 in the HEB-D group and two

in the HS group), and 18 instances of imperatives were documented in the Hebrew data (16 in the HEB-D group and two in the HS group). The ENG-D participants did not use imperatives at all in either of the languages.

For the use of interrogatives, the results showed that there was an effect of Group, yet no effect of Language and no Group*Language interaction. Follow-up pairwise Group contrasts for the use of interrogatives showed that there were significant differences between ENG-D and HEB-D ($\beta = 3.554$; $SE = 0.298$; $Z = 11.924$, $p < 0.0001$), between ENG-D and HS ($\beta = 2.812$; $SE = 0.294$; $Z = 9.580$, $p < 0.0001$), and between HEB-D and HS ($\beta = 0.742$; $SE = 0.279$; $Z = 2.662$, $p = 0.0212$).

The results for the use of declaratives indicated a significant Group*Language interaction, therefore pairwise Group comparisons within each language were conducted. The source of the interaction came from the HS group which paired up with the ENG-D group in English ($\beta = 0.664$; $SE = 0.349$; $Z = -1.901$, $p = 0.1383$), yet differed from the HEB-D group in Hebrew ($\beta = 2.604$; $SE = 0.317$; $Z = 8.220$, $p < 0.0001$). Similar to the use of interrogative, the two dominant groups differed from each other in both languages in the choice of declaratives.

For the choice of mixed strategy, the HS group differed from both dominant groups (HS vs. ENG-D: $\beta = 2.870$; $SE = 0.383$; $Z = 7.500$, $p < 0.0001$; HS vs. HEB-D: $\beta = -3.448$; $SE = 0.409$; $Z = 8.423$, $p < 0.0001$), while the two dominant groups were similarly unlikely to choose a mixed strategy as a request formation option ($\beta = 0.578$; $SE = 0.409$; $Z = -8.423$, $p = 0.3754$).

Thus, the results indicated differences between the two dominant groups: the most common syntactic structure for making requests among ENG-D was the interrogative structure while for HEB-D it was the declarative. Both dominant groups transferred this strategy from their dominant language into their weaker language. Interestingly, HS usage of both interrogatives and declaratives was between the two dominant groups in both English and Hebrew. Moreover, the pattern of

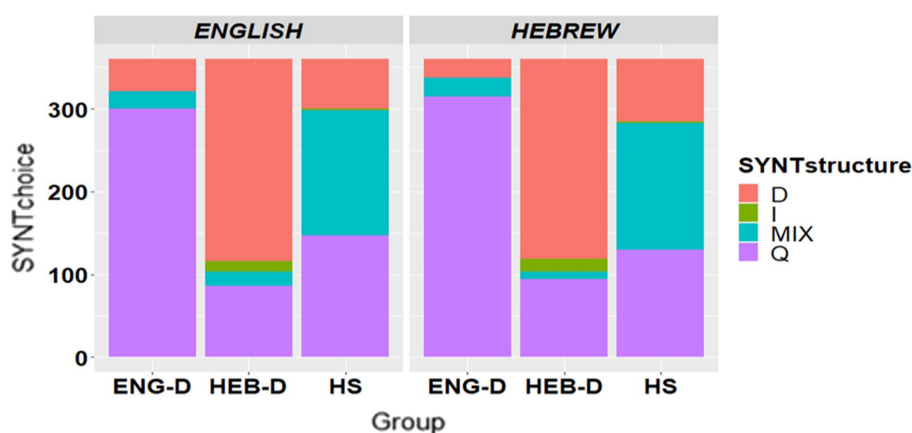


FIGURE 1

The choice of syntactic structure in request formation per group per language. D=declarative, I= imperative, Q= question (interrogative), MIX=combination of two sentence types.

TABLE 7 Final models for the syntactic structure choice.

	Interrogatives				Declaratives				Mixed			
	Estimate	Std. error	z Value	Pr(> z)	Estimate	Std. error	z Value	Pr(> z)	Estimate	Std. error	z Value	Pr(> z)
	Fixed effects				Fixed effects				Fixed effects			
Intercept	2.1842	0.2998	7.286	.18e-13***	-2.6872	0.3464	-7.758	8.66e-15***	-3.2622	0.3283	-9.937	< 2e-16***
Group (ENG-D vs. HEB-D)	-3.5543	0.2981	-11.924	< 2e-16***	3.6571	0.3466	10.550	< 2e-16***	0.5781	0.4329	-1.336	0.182
Group (ENG-D vs. HS)	-2.8120	0.2935	-9.580	< 2e-16***	0.6642	0.3494	1.901	0.05727	2.8697	0.3826	7.500	6.36e-14***
Language												
Group HEB-D:					-0.6789	0.2904	-2.338	0.01941*				
Language Hebrew					0.6323	0.3393	1.864	0.06237				
Group HS:												
Language Hebrew					1.0214	0.3596	2.840	0.00451**				
Random Effects: Number of observations: 2160, Participant: 60, Scenario: 18												
	Variance	Std. Dev			Variance	Std. Dev			Variance	Std. Dev		
Participant (Intercept)	0.6078	0.7796			0.6026	0.7763			0.9711	0.9855		
Scenario (Intercept)	0.7833	0.8850			0.8849	0.9407			0.2898	0.5383		

Significance codes: 0 '***', 0.001 '**', 0.01 '*', 0.05.

mixed syntactic structure was the most preferred among HS and more common compared to the other two groups in both languages. In other words, HS diverged from both dominant groups in their overuse of mixed structure in both the HL-English and the SL-Hebrew. While ENG-D speakers preferred the interrogative structure and HEB-D preferred the declarative, HS relied on a mixed strategy that contained the interrogative and the declarative structures in one scenario (i.e., two head acts in one request).

The use of modals

Figure 2 presents the use of modals across the three groups in English and in Hebrew.

Table 8 presents the final model for the use of modals on a binary scale, coded as 1 = response with a modal and 0 = response without a modal. The results indicated no effect of Language, yet an effect of Group and a significant Group*Language interaction. Therefore, pairwise Group comparisons within each language were conducted. The results showed that ENG-D usage of modals was significantly higher than HEB-D usage of modals in English ($\beta = 2.385$; $SE = 0.292$; $Z = 8.176$, $p < 0.0001$) and in Hebrew ($\beta = 1.453$; $SE = 0.291$; $Z = 4.987$, $p < 0.0001$). However, the interaction came from the HS group which behaved differently in each language. The HS group paired up with the ENG-D group in English ($\beta = 0.436$; $SE = 0.308$; $Z = 1.417$, $p = 0.3322$) and with the HEB-D in Hebrew ($\beta = -0.451$; $SE = 0.267$; $Z = -1.691$, $p = 0.2087$).

To sum up, the two dominant groups were significantly different in their usage of modals, i.e., ENG-D usage of modals was significantly higher than that of HEB-D speakers. However, contrary to both dominant groups who mirrored their strategy to both their languages, the HS were parallel to the ENG-D when speaking English and to the HEB-D when speaking Hebrew.

The use of 'please'/'bevakasha'

Figure 3 presents the use of 'please'/'bevakasha' across the three groups in English and in Hebrew.

Table 8 presents the final model for the use of 'please'/'bevakasha' coded as 1 = response with 'please'/'bevakasha' and 0 = response without 'please'/'bevakasha'. The results indicated no effect of Language, yet an effect of Group and a Group*Language interaction. The follow-up analyses showed that HEB-D usage of 'please'/'bevakasha' was significantly higher than that of ENG-D in both English ($\beta = -4.343$; $SE = 0.405$; $Z = -10.730$, $p < 0.0001$) and Hebrew ($\beta = -4.234$; $SE = 0.408$; $Z = -10.384$, $p < 0.0001$). However, the HS group differed from both dominant groups in both languages by being somewhere in between, i.e., HS usage of 'please'/'bevakasha' was higher than ENG-D ($\beta = -0.982$; $SE = 0.378$; $Z = -2.600$, $p = 0.0253$) and lower than HEB-D ($\beta = -3.361$; $SE = 0.391$; $Z = -8.595$, $p < 0.0001$) in English, as well as higher than ENG-D ($\beta = -1.679$; $SE = 0.399$; $Z = -4.206$, $p = 0.0001$) and lower than HEB-D ($\beta = 2.556$; $SE = 0.374$; $Z = 6.840$, $p < 0.0001$) in Hebrew.

To sum up, the findings for the usage of 'please'/'bevakasha' showed that HEB-D resorted to the use of 'please'/'bevakasha'

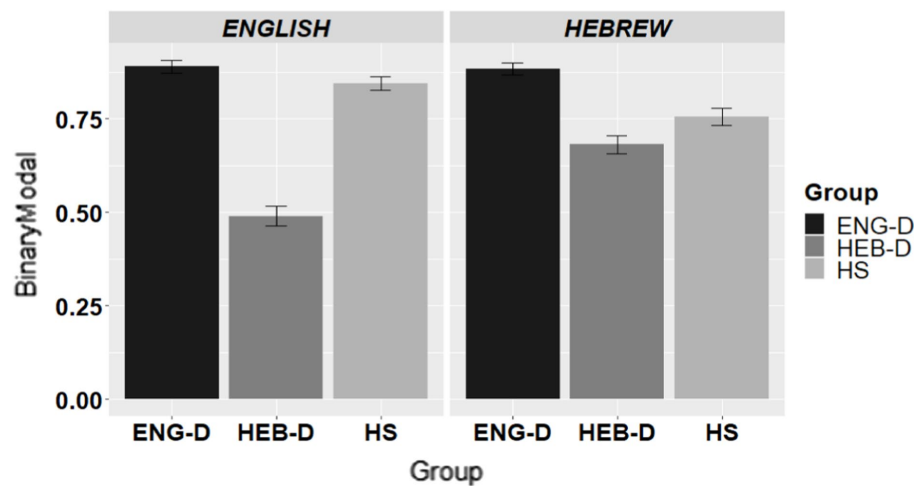


FIGURE 2
The use of modals in request formation per group per language.

TABLE 8 Final models for the use of modals and 'please'/'bevakasha'.

	The use of modals				The use of 'please'/'bevakasha'			
	Fixed effects							
	Estimate	Std. error	z Value	Pr(> z)	Estimate	Std. error	z Value	Pr(> z)
Intercept	2.33706	0.25168	9.286	< 2e-16***	−1.9945	0.3155	−6.321	2.59e-10***
Group (ENG-D vs.HEB-D)	−2.38490	0.29169	−8.176	2.93e-16***	4.3432	0.4048	10.730	< 2e-16***
Group (ENG-D vs. HS)	−0.43609	0.30780	−1.417	0.15654	0.9822	0.3778	2.600	0.00932**
Language Hebrew	−0.05917	0.23893	−0.248	0.80442	−0.8082	0.2459	−3.287	0.00101**
Group HEB-D:	0.93164	0.28840	3.230	0.00124**	−0.1087	0.3316	−0.328	0.74317
Language Hebrew								
Group HS:	−0.56647	0.31061	−1.824	0.06819	0.6967	0.3030	2.300	0.02147*
Language Hebrew								
	Random effects: Number of observations: 2160, Participant: 60; Scenario: 18							
	Variance	Std. Dev			Variance	Std. Dev		
Participant (Intercept)	0.4006	0.6330			1.0033	1.0017		
Scenario (Intercept)	0.1881	0.4338			0.4038	0.6354		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

significantly more than ENG-D in forming requests, while the HS were in between the two dominant groups in both languages.

Apologies

The use of propositions

To address the issue of the use of propositions, we coded the number of propositions used in each apology from 0 to 4 (i.e., offering explanation, taking responsibility, offering repair or compensation, and promising forbearance) giving 1 point for each proposition. Figure 4 presents the number of propositions across the three groups in English and in Hebrew.

The results in Figure 4 show that while the ENG-D group had the highest frequency of propositions, the HEB-D group had the lowest. This trend was observed in both languages. The HS group was in between these two dominant groups in both English and Hebrew. This was confirmed by the statistical analysis in which we fitted linear regression models, as the data were coded in a non-binary manner (0–4), with Group, Language, and the interaction between them as fixed variables. Table 9 presents the final model for the number of propositions used, coded as 0 = response without a proposition, 1 = response with one proposition, 2 = response with two propositions, 3 = response with three propositions, and 4 = response with four propositions. The results showed an effect of Group, an effect of Language, and a significant Language*Group

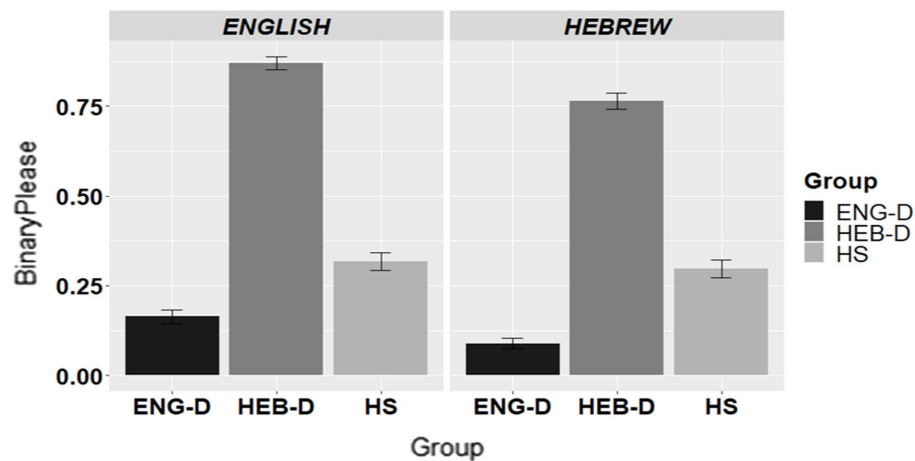


FIGURE 3
The use of 'please'/'bevakasha' in request formation per group per language.

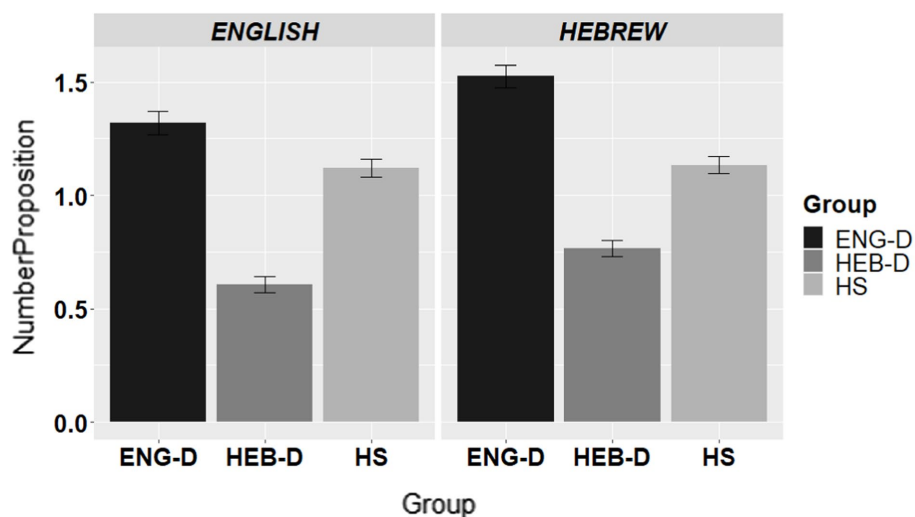


FIGURE 4
The use of propositions in apology formation per group per language.

TABLE 9 Estimate parameters for the use of propositions.

	Est.	25%	75%	<i>t</i> value	<i>p</i>
Intercept	1.32	1.29	1.35	31.96	0.00
Group (ENG-D vs. HEB-D)	−0.71	−0.75	−0.67	−12.23	0.00
Group (ENG-D vs. HS)	−0.20	−0.24	−0.16	−3.43	0.00
Language	0.21	0.17	0.24	3.52	0.00
Group HEB-D: Language	−0.04	−0.10	0.01	−0.54	0.59
Group HS: Language	−0.19	−0.25	−0.14	−2.32	0.02

interaction. Follow-up pair-wise contrasts indicated that the groups differed from each other in both languages (all comparisons at $p < 0.001$).

To sum up, the findings for the usage of propositions showed that the ENG-D group adhered to the use of propositions significantly more than the HEB-D group in forming apologies,

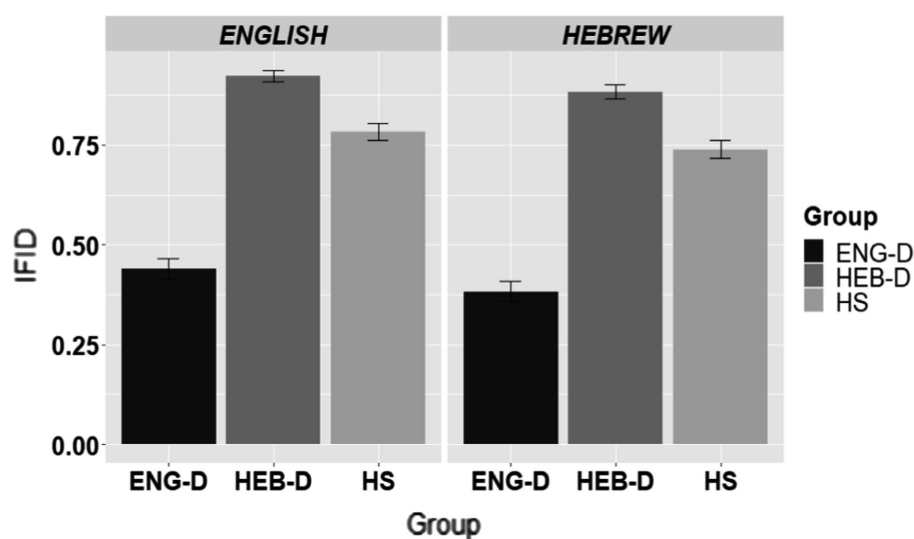


FIGURE 5

The use of illocutionary force-indicating devices (IFIDs) in apology formation per group per language.

while the HS were in between the two dominant groups in both languages.

The use of IFIDs

Figure 5 presents the use of IFIDs across the three groups in English and in Hebrew.

The results in Figure 5 show that IFIDs were frequently used by HEB-D speakers while less so by ENG-D speakers in both languages. The HS were in between these two dominant groups in both English and Hebrew. This was confirmed by the statistical analysis shown in Table 10 which presents the final model of the use of IFIDs, coded as 1 = response with an IFID and 0 = response without an IFID. The results indicated an effect of Group, yet no effect of Language and no significant Group*Language interaction. The results showed that the HEB-D's usage of IFIDs was significantly higher than that of ENG-D ($\beta = -3.48$; $SE = 0.408$; $Z = -8.535$, $p < 0.0001$). The HS group differed from both dominant groups in both languages by being somewhere in between, i.e., HS usage of IFIDs was higher than ENG-D ($\beta = -2.13$; $SE = 0.390$; $Z = -5.475$, $p < 0.0001$) and lower than HEB-D ($\beta = 1.35$; $SE = 0.406$; $Z = 3.317$, $p = 0.0026$).

To sum up, the findings of the usage of IFIDs showed that HEB-D adhered to the use of IFIDs significantly more than ENG-D, while the HS were in between the two dominant groups in both languages.

The use of adverbial intensifiers

Figure 6 presents the use of adverbial intensifiers across the three groups in English and in Hebrew.

To investigate the usage of intensifiers, we also used a binary coding scheme (1 = if an adverbial intensifier was present, and 0 = if it was absent in an apology response). The results in Figure 6 show that the use of adverbial intensifiers was the highest in the HEB-D

group, while it was the lowest in the ENG-D group in both languages. As for the participants in the HS group, their usage of adverbial intensifiers was between these two dominant groups in Hebrew, yet on par with the HEB-D participants in English. This was confirmed by the statistical analysis shown in Table 10. The results indicated no effect of Language, yet an effect of Group and a significant Group*Language interaction. Therefore, pairwise Group contrasts within each language were conducted. The results showed that the HEB-D usage of adverbial intensifiers was significantly higher than that of ENG-D in English ($\beta = -1.731$; $SE = 0.372$; $Z = -4.653$, $p < 0.0001$) and in Hebrew ($\beta = -2.586$; $SE = 0.380$; $Z = -6.802$, $p < 0.0001$). The source of the Group*Language interaction came from the HS group which paired up with the HEB-D in English ($\beta = 0.134$; $SE = 0.370$; $Z = 0.363$, $p = 0.9301$) and differed from both dominant groups in Hebrew by being somewhere in between (both comparisons were $p < 0.001$).

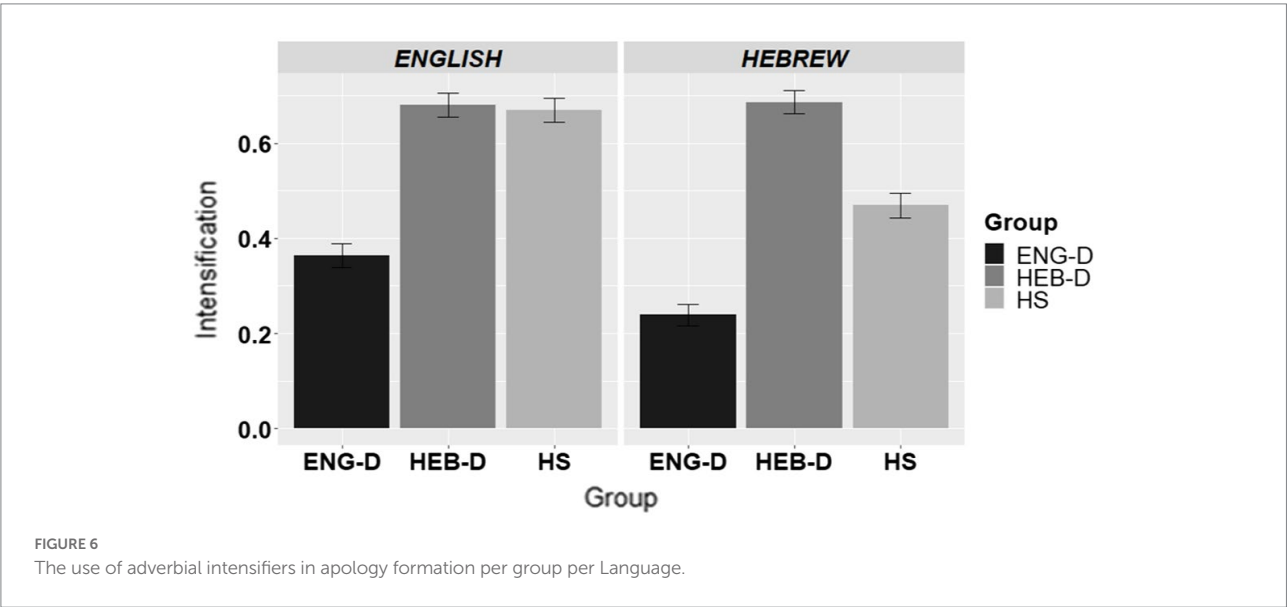
To sum up, the findings of the usage of adverbial intensifiers showed that HEB-D adhered to the use of adverbial intensifiers significantly more than ENG-D. The HS paired up with the HEB-D group in English, while they were in between the two dominant groups in Hebrew.

Table 11 presents examples of prototypical requests and apologies produced by ENG-D, HEB-D, and HS. A prototypical request produced by an ENG-D participant had an interrogative form and included a modal, yet no 'please'/'bevakasha', while a prototypical request produced by a HEB-D participant was formulated as a declarative and included 'please'/'bevakasha', yet no modals. A prototypical HS request included a mixed strategy containing both the declarative and the interrogative structures, and an 'in-between' 'please'/'bevakasha' strategy in both HL-English and SL-Hebrew. HS strategy with regard to modals paired up with the ENG-D in English and with the HEB-D in Hebrew.

TABLE 10 Final models for the use of IFIDs and adverbial intensifiers.

The use of IFIDs					The use of adverbial intensifiers			
					Fixed Effects			
	Estimate	Std. error	z Value	Pr(> z)	Estimate	Std. error	z Value	Pr(> z)
Intercept	−0.2944	0.3713	−0.793	0.42785	−0.7055	0.3290	−2.144	0.0320*
Group (ENG-D vs.HEB-D)	3.4799	0.4077	8.535	< 2e-16***	1.7312	0.3720	4.653	3.27e-06***
Group (ENG-D vs. HS)	2.1332	0.3896	5.475	4.38e-08***	1.5971	0.3694	4.323	1.54e-05***
Language HEBREW	−0.3737	0.1211	−3.086	0.00203**	−0.8207	0.1929	−4.255	2.09e-05***
Group HEB-D:					0.8548	0.2659	3.215	0.0013**
Language HEBREW								
Group HS:					−0.2279	0.2588	−0.881	0.3786
Language HEBREW								
Random effects: Number of observations: 2160, Participant: 60; Scenario: 18								
	Variance	Std. Dev			Variance	Std. Dev		
Participant (Intercept)	1.29	1.136			1.0337	1.017		
Scenario (Intercept)	1.10	1.049			0.7175	0.847		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.



A prototypical apology produced by an ENG-D participant included propositions, yet no IFIDs and no adverbial intensifiers, while a prototypical apology produced by a HEB-D participant included IFID/IFIDs and adverbial intensification, yet no propositions. A prototypical HS apology reflected a mixed strategy, i.e., it included both IFIDs and propositions, and an ‘in-between’ adverbial intensification strategy in Hebrew, while on par with HEB-D in English.

Discussion

The overall aim of this study was to investigate HL-English speakers’ pragmatics *via* request and apology realization patterns in both of their languages, i.e., English and Hebrew, as compared to two

dominant groups: Hebrew-dominant speakers (HEB-D) and English-dominant speakers (ENG-D). The results indicated that dominant speakers of Hebrew and English use different strategies in request and apology formation. Furthermore, the results showed that dominant speakers transfer strategies from their dominant language into their weaker one. As for the HL-English speakers (HS), the results showed a complex picture. In some cases, they paired with dominant speakers, yet sometimes they favored a different strategy. These novel strategies suggest that HL pragmatics is a hybrid system which embodies a mixture of the HL and the SL pointing at a bi-directional cross-linguistic transfer. This hybridity enables HL speakers to draw on pragmatic patterns from their two languages in order to accommodate both languages. Our study shows that the nature of HL pragmatics can be studied when considering the pragmatic competence in both languages of HL speakers.

TABLE 11 Examples of prototypical requests and apologies per group per language (Scenario 2).

Requests		
ENG-D	English (P53)	May I use your computer just for a few minutes?
	Hebrew (P6)	אני יכולה להשתמש במחשב שלך לדקה, שתי דקות? Ani yexola lehishtamesh bamaxshev shelxa ledaka, shtei dakot? <i>Can I use your computer for a minute, two minutes?</i>
HEB-D	English (P24)	I need to use your computer for a sec please.
	Hebrew (P16)	אני יודעת שממש חשוב לך להשתמש במחשב, אבל אשמח לאפשרות לקבל אותו לכמה דקות בבקשה. Ani yoda'at shemamash xashuv lexa lehishtamesh bamaxshev, aval esmax le'efsharut lekabel oto lekama dakot bevakasha. <i>I know that it is extremely important for you to use the computer, but I'll be happy to have the opportunity to get it for a few minutes please.</i>
HS	English (P43)	Hi honey. I need your computer. Please, it's important. Can I have it? Just for a few minutes.
	Hebrew (P33)	היי חומר. המחשב שלי לא נדלק אז אני אשמח לקחת את שלך לרגע. יש מצב שאתה נותן לי אותו לכמה דקות? Hi xomed. Hamaxshev sheli lo nidlak az ani esmax lakaxat et shelxa lerega. Yesh matzav she'ata noten li oto lekama dakot? <i>Hi hon. My computer is not turning on so I'll be happy to take yours for a second. Is there a chance you give it to me for a few minutes?</i>
Apologies		
ENG-D	English (P59)	Oh my goodness. I do not know how it happened. I'll do another essay and write your name on it.
	Hebrew (P38)	אופס, זה לא היה בכוונה. אולי אני יכולה לעזור לך לכתוב את זה עוד פעם? Oops, ze lo haya bexavana. Ula'i ani yexola la'azor lexa lixtov et ze od pa'am? <i>Oops, it wasn't on purpose. Maybe I can help you rewrite it again?</i>
HEB-D	English (P15)	Ohh, I'm very very sorry for deleting your essay. I do not have any words to beg your pardon.
	Hebrew (P15)	וואו, אני מאוד מצטער על מחיקת התיכון. אני מבקש את סליחתך. Wow, ani me'od mitzta'er al mexikat haxibur. Ani mevakesh et slixatxa. <i>Wow, I'm so sorry about the essay deletion. I request your forgiveness.</i>
HS	English (P1)	Listen, I accidentally deleted your file. It wasn't on purpose. I'm really sorry.
	Hebrew (P50)	בטעות נמחק לי התיכון שלך. אני ממש מצטער. אם אתה רוצה שאני אעזור לך לשחזר אני איתך. Beta'ut nimxak li haxibur shelxa. im ata rotze she'ani e'ezor lexa leshaxzer ani itxa. <i>By mistake your essay has been deleted to me. I'm really sorry. If you want me to help you rconstruct it I'm with you.</i>

Requests strategies

For the analysis of request strategies, we compared the choice of a syntactic structure of head acts, the use of modals, and the use of 'please'/'bevakasha' across the three groups in both of their languages. These measures of analyses were chosen as they were hypothesized to reflect differences in request formation in English and in Hebrew (see the studies reviewed in the Introduction subsection).

Starting with the choice of the syntactic structure, differences were found between the two dominant groups (ENG-D and HEB-D) reflecting differences in the cultural perception of appropriateness of English-dominant and Hebrew-dominant speakers. However, before exploring the preferred syntactic structure in each of the languages, it is important to mention that even though requests in both English and Hebrew can be grammatically realized with imperatives, interrogatives, and

declaratives (Blum-Kulka and Olshtain, 1984; Curl and Drew, 2008), imperatives were hardly ever used by Hebrew-dominant speakers, even less so by HL-English speakers, and not even once by English-dominant speakers. In a continuation to directness ideas, a request in the form of an imperative is the most direct and explicit, and therefore is considered to be the least polite (Blum-Kulka and Olshtain, 1984). This indicates that usage of imperative structure requires much effort to modify, and therefore tends to be avoided. Blum-Kulka et al. (1989) and Leech (1980) noted that since interrogatives are most often requests for permission they increase the degree of optionality, and therefore are perceived as being more polite and indirect than declaratives. As expected from the literature review summarized in the Introduction, and as can be seen from the results of the current study, the preferred structure for requests among English-dominant speakers was the interrogative, while for the Hebrew-dominant speakers it was the declarative. The HL-English speakers had at their disposal the

strategies of two languages; however, they favored a mixed strategy containing the interrogative and the declarative structures in the same request. This hybrid strategy allowed the HL-English speakers to transfer the same strategy between both their languages.

Differences in the usage of modals were also found between the two dominant groups. English-dominant speakers used modals significantly more frequently than Hebrew-dominant speakers in their dominant language and in their weaker one. This is consistent with [Turnbull and Saxton's \(1997\)](#) proposal that English speakers use expressions of modality to do 'facework' since they convey the notion of permission, ability, probability, possibility, etc., and therefore further emphasize the indirectness of the utterance. The HL-English speakers' usage of modals was found to be on par with the English-dominant speakers in English and with the Hebrew-dominant speakers in Hebrew. Contrary to the trend shown for the choice of the syntactic structure, in the usage of modals, the HS group did not develop a unique and hybrid strategy, but rather adopted the customary behavior of each language.

Differences in the usage of *'please'/'bevakasha'*, as predicted, were also found between the two dominant groups: Hebrew-dominant speakers resorted to the use of *'please'/'bevakasha'* significantly more frequently than English-dominant speakers. This matches [House's \(1989\)](#) idea that the politeness marker *'please'* is most appropriate in mitigating situations where the function of the request is clear, and less so with interrogatives since it might reveal the true nature of the request. The HL-English speakers, on the other hand, were found to be in between these two dominant groups in both languages.

The results for request formation showed that HL-English speakers developed a unique and hybrid intercultural linguistic style reflecting strategies of both languages (their HL-English and their SL-Hebrew): In both languages, HL-English speakers adhered to a mixed strategy containing the interrogative and the declarative structures in the same request, and their usage of *'please'/'bevakasha'* was in between the two dominant groups. This is in line with [Taguchi and Roever's \(2017\)](#) suggestion that HL pragmatics is a hybrid system reflecting norms of both HL and SL that develops in blended social contexts (i.e., social interactions with both languages' communities) and is mediated by bi-directional cross-linguistic influence. Furthermore, the results were in line with [Pinto and Raschio's \(2007\)](#) findings showing that when HL-Spanish speakers came into contact with English both pragmatic systems were affected. However, we also see that in some cases HL speakers adopt the customary behavior of each language, as it is the case for the use of modals.

Apology strategies

For the analysis of apology strategies, we compared usage of apology expressions, number of propositions added (i.e., offering explanation, taking responsibility, offering repair or compensation, and promising forbearance), and usage of adverbial intensifiers

across the three groups in their both languages. These measures of analyses were chosen as they were hypothesized to reflect differences in apology formation in English and in Hebrew (see the studies reviewed in the Introduction subsection).

In order to interpret the results of the usage of propositions and the usage of IFIDs, it is important to consider them together, since, as we shall see, the results of the HL-English speakers' usage of both are connected. In line with previous studies, there were significant differences between the two dominant groups (ENG-D and HEB-D) with respect to both usage of propositions and usage of IFIDs. As [Blum-Kulka and Olshtain \(1984\)](#) noted there are two options for apologizing: the first one is direct and explicit and involves the use of an IFID, while the second one is indirect and involves 'going around' by offering one or more of four propositions. However, a speaker might also choose to incorporate both strategies within one apology. Since Hebrew is reported to be more direct and straightforward than English ([Mills and Grainger, 2016](#)), it was not surprising to find that Hebrew-dominant speakers preferred to apply the direct strategy in the form of an IFID, while English-dominant speakers tended to apply the indirect strategy in the form of propositions to their apologies. In fact, this trend was so salient that it looked as if the Hebrew-dominant speakers believed that an apology must comprise an IFID as a compulsory component, optionally followed by the other strategy, while the English-dominant speakers believed that propositions such as explaining, taking responsibility or offering repair were more appropriate than IFIDs. The HL-English speakers, however, did not replicate either one of the dominant groups' strategies. Instead, they seemed to develop their own strategy of apologies by combining both IFIDs and propositions. This hybridity was applied by HS in both their languages. In other words, the HL-English speakers were found to be in between the two dominant groups in the use of both IFIDs and propositions.

Differences in the usage of intensifiers, again as expected, were found between the two dominant groups; Hebrew-dominant speakers favored the use of intensifiers significantly more than English-dominant speakers. However, it is important to note here that this study focused on adverbial intensifiers expressions only (lexical, and not phrasal), such as 'so', 'very', 'really', 'terribly', 'extremely', 'totally', 'deeply', 'highly' etc., and disregarded other intensifying expressions. Since English is less direct than Hebrew, and English native speakers' usage of IFIDs is reduced as compared to native Hebrew speakers, it is logical to assume that native English speakers might choose to incorporate less direct intensifying strategies in their apologies. For example, they might choose expressions that convey concern for the hearer, which are external to the IFID or the other strategies used such as 'Have you been waiting long?'. However, this trend was not checked in the current research. The usage of adverbial intensifiers among the HS group was between the two dominant groups in Hebrew, yet, on par with the Hebrew-dominant speakers in English. Future studies should expand the research on intensification by looking into the usage of intensifiers in all their forms.

Thus, the picture for apology formation in HL-English speakers was similar to that of request realization. In some aspects, HL-English speakers adhered to the strategy of dominant speakers of the languages, as is the case for the use of adverbial intensifiers in English, yet in other cases their realization patterns of apologies reflected a blended pragmatic system which suited both languages, as is the case of IFIDs and propositions.

Heritage language pragmatics: Economy principle/dual identity/intercultural style hypothesis

The results indicated that the dominant groups had different strategies for making requests and apologies which they systematically transferred from their dominant language into their weaker one, confirming the cross-cultural and cross-linguistic differences between request and apology strategies in English and in Hebrew. As for HL-English speakers, new blended conventions of request and apology were detected.

The HL-English speakers' pragmatic hybridity might be explained in the light of the 'cognitive economy principle'. The HL speakers' proficiency in two languages enables them to combine their knowledge into one strategy and use it for both languages. The principle of economy has been proposed to influence the restructuring of HL grammars. We speculate that the driving source of the hybrid nature of pragmatics in HL-Speakers in their HL-English and SL-Hebrew might be related to the proposed economy principles. Blended new conventions, formed as a result of a bi-directional cross-linguistic transfer, might be less cognitively costly as compared to the storage and retrieval of two separate systems. However, we agree that 'cognitive economy' is an elusive concept (Westergaard, 2021), and therefore call for future studies to further investigate this possibility.

Alternatively, the HL-English speakers' hybridity might also be connected to issues of dual identity as it fulfills not just linguistic but also identity needs. As Val and Vinogradova (2010) suggested, the core identity of HL speakers involves the process of constant negotiation and self-positioning within a bilingual and bicultural environment. Previous studies investigating the identity of HL speakers note their complex identities. For example, Kang (2013) showed that HL-Korean speakers residing in the USA perceived themselves as different from both Koreans and "mainstream Americans." The identity perception of HL-English speakers residing in Israel was demonstrated for preschool children (see Altman et al., 2021). The authors showed that English-Hebrew bilingual children residing in Israel gave similar ratings to Societal/Israeli and Home/American identities, pointing to the existence of bicultural identity already in young children. It is highly plausible that the pragmatic competence of adult HL-English speakers residing in Israel in the current study reflects their multiple sociolinguistic identities. Future research should address how the sociolinguistic identity is related to the pragmatic competence of an HL speaker, i.e., whether there are differences

between HL speakers who value their HL identity higher versus those who value their SL identity higher. Yet, our research cannot support or rule out this hypothesis, and future studies also need to incorporate data on the identity of HL-speakers to test whether linguistic hybridity reflects HL-speakers' complex dual identity.

Our findings for the hybrid/blended pragmatic conventions highlight the importance of analyzing bi-directional interaction in pragmatic development and might also be related to the 'Intercultural Style Hypothesis'. Intercultural style has been shown to develop when speakers master proficiency in two languages or more. Since bilinguals/multilinguals are exposed to different ways of achieving pragmatic competence in different languages, they could use an underlying conceptual base and develop an intercultural style which explains the similarities of their realization patterns in all their languages. Monolinguals do not need to use these strategies since their realizations correspond to their experience in one single language (Cenoz, 2008).

Limitations and future studies

Despite the fact that the study contributes to the understanding of the existing literature on politeness and language maintenance among adult HL speakers, it is not without limitations. The results showed no effects of Social Status and Social Distance parameters which was rather surprising. One possible explanation is that the design of this study did not control for the severity (imposition for requests, and offense for apologies) and the settings of the situations. We believe that future studies should control for situational severity and situational settings in order to detect Social Status and Social Distance effects in a more rigorous manner. Furthermore, the focus of this study regarding intensifiers was limited to adverbial intensifier expressions only (lexical, and not phrasal), and disregarded other intensifying expressions such as expressions that convey concerns for the hearer which are external to the IFID or other strategies used. This might have caused a partial picture of the usage of intensifications. In future studies, it might be worthwhile to look into the usage of intensifiers in a more comprehensive way in order to get a fuller and more accurate picture. Future studies might also want to distinguish between different types of declaratives as they behave differently with respect to pragmatics. Finally, our recommendations are to expand the investigation of requests to other linguistic categories following Blum-Kulka and Olshtain (1984) such as strategy type (direct, conventionally indirect, non-conventionally indirect), point of view operation, downgraders, etc. as well as to other speech acts and/or languages.

Conclusions

The study adds to the existing literature on politeness and language maintenance among bilingual speakers. The design of the current study, which included the investigation of both languages of three groups of bilinguals, has provided valuable

insights into the pragmatics of dominant speakers, L2 learners, and HL speakers. From a theoretical perspective, the study sheds light on the pragmatic competence of HL speakers in language contact situations by examining cross-linguistic and cross-cultural differences in order to provide a greater understanding of the mechanisms responsible for shaping speech act realizations. The results indicate that dominant speakers of Hebrew and English adhere to different strategies for making requests and apologies and that they systematically transfer these strategies from their dominant language into their weaker one, confirming the cross-cultural and cross-linguistic differences between request and apology strategies in English and in Hebrew. For the HL-English speakers, the picture was more complex: in some cases, strategies of HL-English speakers paired up with dominant speakers in HL-English and/or SL-Hebrew, while in other cases HL-English speakers developed a unique and hybrid linguistic style reflecting pragmatic conventions of both their languages, HL-English and SL-Hebrew. From a pedagogical perspective, the current study contributes to the field of teaching pragmatic skills to HL speakers and L2 learners, helping educators develop research-supported curricula that facilitate appropriate politeness strategies.

We believe that the main strength of the current study lies in its methodology: testing both languages of three groups with different levels of dominance. This design enabled us to investigate the two linguistic systems simultaneously and draw conclusions about their nature. Despite the assumption that HL speakers diverge in their HL and perform on par with dominant speakers in their SL, the current study shows that subtle differences may be observed in both languages. Thus, we highlight here the importance/advantages of investigating both languages of HL speakers in future studies in order to obtain a fuller picture of this unique bilingual group.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

References

- Achugar, M. (2006). Writers on the borderlands: constructing a bilingual identity in Southwest Texas. *J. Lang. Identity Educ.* 5, 97–122. doi: 10.1207/s15327701jlie0502_1
- Altman, C., Burstein-Feldman, Z., Fichman, S., Armon-Lotem, S., Joffe, S., and Walters, J. (2021). Perceptions of identity, language abilities and language preferences among Russian-Hebrew and English-Hebrew bilingual children and their parents. *J. Multiling. Multicult. Dev.*, 1–16. doi: 10.1080/01434632.2021.1974462
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Bates, D., Machler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bella, S. (2012). Length of residence and intensity of interaction: modification in Greek L2 requests. *Pragmatics* 22, 1–39. doi: 10.1075/prag.22.1.01bel
- Berard, T. J. (2005). On multiple identities and educational contexts: remarks on the study of inequalities and discrimination. *J. Lang. Identity Educ.* 4, 67–76. doi: 10.1207/s15327701jlie0401_4
- Blum-Kulka, S. (1990). You don't touch lettuce with your fingers: parental politeness in family discourse. *J. Pragmat.* 14, 259–288. doi: 10.1016/0378-2166(90)90083-P
- Blum-Kulka, S. (1991). "Interlanguage pragmatics: the case of requests," in *Foreign/second Language Pedagogy Research*. eds. R. Phillipson, E. Kellerman, L. Selinker, M. Sharwood Smith and M. Swain (Clevedon: Multilingual Matters), 255–272.
- Blum-Kulka, S., House, J., and Kasper, G. (1989). *Cross-cultural Pragmatics: Requests and Apologies*. Norwood, NJ: Ablex.
- Blum-Kulka, S., and Olshtain, E. (1984). Requests and apologies: a cross-cultural study of speech act realization patterns (CCSARP). *Appl. Linguis.* 5, 196–213. doi: 10.1093/applin/5.3.196
- Blum-Kulka, S., and Sheffer, H. (1993). "The metapragmatic discourse of American-Israeli families at dinner," in *Interlanguage Pragmatics*. eds. G. Kasper and S. Blum-Kulka (Oxford: Oxford University Press), 196–223.
- Brown, P., and Levinson, S. (1987). *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Ethics statement

The current study was reviewed and approved by Bar Ilan University. The participants provided their written informed consent to participate in this study.

Author contributions

SBO and NM conceptualized the study and wrote the manuscript. SBO developed the tasks, collected and processed the data. NM was involved in data analysis. All authors contributed to the article and approved the submitted version.

Funding

The study was partially supported by the Israel Science Foundation (ISF) No. 552/21 "Towards Understanding Heritage Language Development: The Case of Child and Adult Heritage Russian in Israel and the USA" granted to NM.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Cenoz, J. (2003). "The intercultural style hypothesis: L1 and L2 interaction in requesting behaviour," in *Effects of the Second Language on the First*. ed. V. Cook, 62–80.
- Cenoz, J. (2008). "The acquisition of pragmatic competence and multilingualism in foreign language contexts," in *Intercultural Language Use and Language Learning* (Dordrecht: Springer), 123–140.
- Cohen, A. D., and Olshtain, E. (1981). Developing a measure of socio-cultural competence: the case of apology. *Lang. Learn.* 31, 113–134. doi: 10.1111/j.1467-1770.1981.tb01375.x
- Cohen, A. D., Olshtain, E., and Rosenstein, D. S. (1986). Advanced EFL apologies: what remains to be learned? *Int. J. Sociol. Lang.* 1986, 51–74. doi: 10.1515/ijsl.1986.62.51
- Curl, T. S., and Drew, P. (2008). Contingency and action: a comparison of two forms of requesting. *Res. Lang. Soc. Interact.* 41, 129–153. doi: 10.1080/08351810802028613
- Dubinina, I. (2011). How to ask for a favor: A pilot study in Heritage Russian pragmatics. In *Instrumentarium of linguistics: Sociolinguistic approaches to non-standard Russian*. eds. A. Mustajoki, E. Protassova and N. Vakhtin (Helsinki: University of Helsinki), 418–431.
- Dubinina, I., and Malamud, S. (2017). Emergent communicative norms in a contact language: indirect requests in heritage Russian. *Linguistics* 55, 67–116. doi: 10.1515/ling-2016-0039
- Elabbas, B., Montrul, S., and Polinsky, M. (2013). Defining an 'ideal' heritage speaker: theoretical and methodological challenges. Reply to peer commentaries. *Theor. Linguist.* 39, 3–4. doi: 10.1515/tl-2013-0018
- Ellis, D. G., and Maoz, I. (2002). Cross-cultural argument interactions between Israeli-Jews and Palestinians. *J. Appl. Commun. Res.* 30, 181–194. doi: 10.1080/00909880216583
- Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., and Cera, C. M. (2012). Self-ratings of spoken language dominance: a multilingual naming test (MINT) and preliminary norms for young and aging Spanish-English bilinguals. *Biling. Lang. Cogn.* 15, 594–615. doi: 10.1017/S1366728911000332
- House, J. (1989). "Politeness in English and German: the functions of 'please' and 'bitte' in Cross-cultural Pragmatics: Requests and Apologies. eds. S. Blum-Kulka, J. House and G. Kasper (New Jersey: Ablex), 96–119.
- Israa, Q. (2017). Politeness study of requests and apologies as produced by Saudi Hijazi, EFL learners, and British English university students. PhD, London: University of Roehampton.
- Kang, H. S. (2013). Korean American college students' language practices and identity positioning: "not Korean, but not American," *J. Lang. Educ.* 12, 248–261. doi: 10.1080/15348458.2013.818473
- Kasper, G., and Blum-Kulka, S. (1993). "Interlanguage pragmatics: an introduction," in *Interlanguage Pragmatics* (Oxford: Oxford University Press), 3–17.
- Katriel, T. (1986). *Talking Straight: Dugri Speech in Israeli Sabra Culture*. Cambridge: Cambridge University Press.
- Kecskes, I., and Papp, T. (2000). *Foreign Language and Mother Tongue*. Mahwah, NJ: Erlbaum.
- Kupisch, T., Barton, D., Hailer, K., Klaschik, E., Stangen, I., Lein, T., et al. (2014). Foreign accent in adult simultaneous bilinguals. *Heritage Lang. J.* 11, 123–150. doi: 10.46538/hlj.11.2.2
- Leech, G. (1980). *Explorations in Semantics and Pragmatics*. Amsterdam: John Benjamins.
- Márquez-Reiter, R. (2000). *Linguistic Politeness in Britain and Uruguay: A Contrastive Study of Requests and Apologies*. Amsterdam: John Benjamins Publishing.
- Meir, N., Joffe, S., Shabtaev, R., Walters, J., and Armon-Lotem, S. (2021). "Heritage languages in Israel: the multilingual tapestry with Hebrew threads," in *The Cambridge Handbook of Heritage Languages and Linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 129–155.
- Mills, S., and Grainger, K. (2016). *Directness and Indirectness Across Cultures*. Berlin: Springer.
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.
- Montrul, S. (2018). Heritage language development: connecting the dots. *Int. J. Biling.* 22, 530–546. doi: 10.1177/136700691665436
- Murphy, M. L., and De Felice, R. (2018). Routine politeness in American and British English requests: use and non-use of please. *J. Politeness Res.* 15, 77–100. doi: 10.1515/pr-2016-0027
- Ogiermann, E. (2009). Politeness and in-directness across cultures: a comparison of English, German, Polish and Russian requests. *Journal of politeness research language behaviour. Culture* 5, 189–216. doi: 10.1515/JPLR.2009.011
- Olshtain, E. (1983). "Sociocultural competence and language transfer: the case of apology," in *Language Transfer in Language Learning*. eds. S. Gass and L. Selinker (Rowley, MA: Newbury House), 232–249.
- Olshtain, E., and Blum-Kulka, S. (1985). Cross-cultural pragmatics and the testing of communicative competence. *Lang. Test.* 2, 16–30. doi: 10.1177/026553228500200103
- Pinto, D., and Raschio, R. (2007). A comparative study of requests in heritage speaker Spanish, L1 Spanish, and L1 English. *Int. J. Biling.* 11, 135–155. doi: 10.1177/13670069070110020101
- Polinsky, M. (2018). *Heritage Languages and Their Speakers*. Cambridge: Cambridge University Press.
- Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245
- Rintell, E. (2009). *Sociolinguistic Variation and Pragmatic Ability: A Look at Learners*. Berlin: De Gruyter Mouton.
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814
- Rothman, J. (2015). Linguistic and cognitive motivations for the typological primacy model (TPM) of third language (L3) transfer: timing of acquisition and proficiency considered. *Biling. Lang. Cogn.* 18, 179–190. doi: 10.1017/S136672891300059X
- Scontras, G., Polinsky, M., Tsai, C. Y. E., and Mai, K. (2017). Cross-linguistic scope ambiguity: when two systems meet. *Glossa* 2, 1–28. doi: 10.5334/gigl.198
- Searle, J. (1975). "Indirect speech act," in *Syntax and Semantic*. eds. P. Cole and J. Morgan (New York, NY: Academic Press), 59–82.
- Stangen, I., Kupisch, T., Ergün, A. L., and Zielke, M. (2015). "Foreign accent in heritage speakers of Turkish in Germany," in *Transfer Effects in Multilingual Language Development*. ed. H. Peukert (New York, NY: Oxford University Press), 87–108. Amsterdam/Philadelphia, PA: John Benjamins.
- Su, I. R. (2010). Transfer of Pragmatic competences: A bi-directional perspective. *The Modern Language Journal*, 94, 87–102.
- Taguchi, N., and Roeper, C. (2017). *Second language pragmatics*. Oxford: Oxford University Press, (pp. 241–276).
- Tomoschuk, B., Ferreira, V. S., and Gollan, T. H. (2019). When a seven is not a seven: self-ratings of bilingual language proficiency differ between and within language populations. *Biling. Lang. Cogn.* 22, 516–536. doi: 10.1017/S1366728918000421
- Turnbull, W., and Saxton, K. L. (1997). Modal expressions as facework in refusals to comply with requests: I think I should say 'no' right now. *J. Pragmat.* 27, 145–181. doi: 10.1016/S0378-2166(96)00034-3
- Val, A., and Vinogradova, P. (2010). *What is the Identity of a Heritage Language Speaker? Heritage Briefs*. Center for Applied Linguistics: Washington, DC.
- Wallace, K. R. (2004). Situating multiethnic identity: contributions of discourse theory to the study of mixed heritage students. *J. Lang. Identity Educ.* 3, 195–213. doi: 10.1207/s15327701jlie0303_2
- Walters, J. (1979). Strategies for requesting in Spanish and English. Structural similarities and pragmatic differences. *Lang. Learn.* 29, 277–293. doi: 10.1111/j.1467-1770.1979.tb01069.x
- Walters, J. (1980). Grammar, meaning, and sociocultural appropriateness in second language acquisition. *Can. J. Psychol./Revue canadienne de psychologie* 34, 337–345. doi: 10.1037/h0081107
- Walters, J. (1981). Variation in the requesting behavior of bilingual children. *Int. J. Sociol. Lang.* 1981, 77–92. doi: 10.1515/ijsl.1981.27.77
- Webman-Shafran, R. (2019). Level of directness and the use of please in requests in English by native speakers of Arabic and Hebrew. *J. Pragmat.* 148, 1–11. doi: 10.1016/j.pragma.2019.05.020
- Westergaard, M. (2021). Microvariation in multilingual situations: the importance of property-by-property acquisition. *Second. Lang. Res.* 37, 379–407. doi: 10.1177/0267658319884116
- Xiao-Desai, Y. (2019). "Heritage learner pragmatics," in *The Routledge Handbook of Second Language Acquisition and Pragmatics* (London: Routledge), 462–478.

Appendix A: List of stimuli

Social status	Social distance	Gender	Speech act	The situation
S>H	L-SD	F	R	A parent asks his/her daughter to plan her birthday party on her own.
S>H	L-SD	F	A	The parent forgets about it and does not show up for the party.
S>H	L-SD	M	R	An uncle/aunt asks a young nephew to use his computer for a couple of minutes.
S>H	L-SD	M	A	The uncle/aunt deletes the essay he was writing.
S>H	M-SD	F	R	A manager asks an employee to take notes during an important meeting.
S>H	M-SD	F	A	The manager spills coffee on the employee's dress.
S>H	M-SD	M	R	A school principal asks a teacher to come to his/her office at four o'clock to discuss a school project.
S>H	M-SD	M	A	The school principal has kept the teacher waiting for an hour because of an unexpected meeting.
S>H	H-SD	F	R	A senior manager asks a new trainee to borrow her brand-new laptop.
S>H	H-SD	F	A	The senior manager accidentally smashes the laptop screen.
S>H	H-SD	M	R	A university professor asks a new student to give his presentation a week earlier than scheduled.
S>H	H-SD	M	A	The university professor realizes that the original date was the correct one.
S<H	L-SD	F	R	The speaker asks his/her mother to borrow her car.
S<H	L-SD	F	A	The speaker drives the car into a tree.
S<H	L-SD	M	R	The speaker asks his/her father to borrow his boat for a date.
S<H	L-SD	M	A	The speaker breaks the steering wheel.
S<H	M-SD	F	R	A teacher asks the school principal to leave early so he/she can go to a conference.
S<H	M-SD	F	A	The teacher finds out that the conference is scheduled for the following week.
S<H	M-SD	M	R	The speaker asks his elderly neighbor for help in statistics.
S<H	M-SD	M	A	The speaker forgets to show up on time and is an hour late.
S<H	H-SD	F	R	A new student asks a lecturer to borrow her book.
S<H	H-SD	F	A	The student forgets to return the book.
S<H	H-SD	M	R	An employee asks a new manager for a loan.
S<H	H-SD	M	A	The employee returns the money later than agreed.
S=H	L-SD	F	R	A sibling asks his/her sister to be first in the shower.
S=H	L-SD	F	A	The sibling realizes that he/she used up all the hot water.
S=H	L-SD	M	R	A friend asks another friend to make use of his house in the countryside.
S=H	L-SD	M	A	The friend spills ink on an expensive carpet.
S=H	M-SD	F	R	An employee asks a coworker from a different department to cover for him/her for an hour while going on a personal errand.
S=H	M-SD	F	A	The employee returns after 3 hours.
S=H	M-SD	M	R	The speaker asks his neighbor for help to move a bookcase out of the apartment.
S=H	M-SD	M	A	The speaker accidentally closes a drawer on the neighbor's hand.
S=H	H-SD	F	R	The speaker asks a bus passenger to swap seats with her.
S=H	H-SD	F	A	The speaker steps on the passenger's foot.
S=H	H-SD	M	R	The speaker asks someone on the street to make a quick call from his cellphone.
S=H	H-SD	M	A	The speaker realizes that he had his/her cellphone in his/her pocket all along.

S=Speaker, H=Hearer, L-SD=Low Social Distance, M-SD=Medium Social Distance, H-SD=High Social Distance, F=Female, M=Male, R=Request, A=Apology.



OPEN ACCESS

EDITED BY

Maki Kubota,
UiT The Arctic University of Norway,
Norway

REVIEWED BY

Irini Mavrou,
Nebrija University,
Spain
Sarah Surrain,
University of Texas Health Science Center
at Houston, United States

*CORRESPONDENCE

My V. H. Nguyen
mvhnguyen@uh.edu

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 03 September 2022

ACCEPTED 14 November 2022

PUBLISHED 11 January 2023

CITATION

Nguyen MVH, Serafini EJ, Leeman J and
Winsler A (2023) Factors predicting
secondary school language course
enrollment and performance among U.S.
heritage speakers of Spanish.
Front. Psychol. 13:1035716.
doi: 10.3389/fpsyg.2022.1035716

COPYRIGHT

© 2023 Nguyen, Serafini, Leeman and
Winsler. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Factors predicting secondary school language course enrollment and performance among U.S. heritage speakers of Spanish

My V. H. Nguyen^{1*}, Ellen J. Serafini², Jennifer Leeman²
and Adam Winsler³

¹Department of Psychology, University of Houston, Houston, TX, United States, ²Department of Modern and Classical Languages, George Mason University, Fairfax, VA, United States, ³Department of Psychology, George Mason University, Fairfax, VA, United States

Introduction: While a growing body of research indicates that Spanish language courses can promote Spanish maintenance and lead to overall improved educational outcomes among heritage speakers, there is little empirical or longitudinal evidence of factors that shape their enrollment in Spanish language courses at the secondary level. To address this issue, the current study takes a large-scale, longitudinal approach to investigate rates of enrollment in secondary school (6th–12th grade) Spanish and other non-English language courses, as well as factors that predict heritage speakers' enrollment and performance in non-English language courses.

Method: We analyzed subsample data from the Miami School Readiness Project (MSRP), a large-scale, longitudinal study consisting of 17,341 heritage speakers of Spanish (47% female, 95.4% Hispanic/Latino, 82.8% received free/reduced-price lunch, and 18.3% with a disability) who were followed from 4 years old until the end of high school.

Results: In general, Heritage speakers enrolled in Spanish language courses at a higher rate than other non-English language courses (52.2 and 25.3%, respectively). Enrollment patterns varied across different type of languages and grade level. Student-level factors including disability status, poverty status, early behavioral problems, and prior academic achievement significantly predicted students' enrollment in Spanish and performance in non-English language courses.

Discussion: Findings shed light on the long-term patterns of language study of this growing segment of the US school population with implications for future research and school policies that seek to improve heritage language learning and maintenance as well as equitable access to language education for language-minority students.

KEYWORDS

heritage language learners, Spanish, language courses, language learning, language education, educational equity

Introduction

In the United States, speakers of languages other than English face many challenges in maintaining their home language and/or passing it on to their children. Among the most significant obstacles is the lack of sustained educational opportunities to build proficiency and academic literacy in those languages, which is linked to state and national language policies that promote English monolingualism (Cummins, 2005; Wiley and García, 2016; Ennser-Kananen and King, 2018; Fuller and Leeman, 2020). While local community-based schools historically have played a key role in supporting minority language maintenance and linguistic diversity (Fishman, 2001), since the early 20th century, the dominant language-in-education paradigm in the U.S. has been English-medium schooling. This policy is buoyed by pervasive English-only ideologies and discourses that portray proficiency in non-English languages as a threat to national unity and an impediment to the acquisition or mastery of English (Crawford, 2000; Wiley, 2000).

This lack of consistent educational support, coupled with a “language as problem” (Ruiz, 1984) orientation to language in policy and planning, contributes to the dominant pattern of a shift to English language dominance/use over time within immigrant families and the loss of the heritage language, typically by the third generation (Veltman, 1983; Fishman, 2001; Alba, 2004; Rumbaut, 2009). As the most commonly spoken non-English language in the U.S. with deep sociohistorical roots and continuous immigration patterns, some have suggested that Spanish might be more immune to language shift and loss than other minoritized languages. However, while research indicates that Spanish is sometimes maintained beyond the third generation, this is typically only in areas with a high density of Spanish-speakers and continued immigration, and it remains the exception rather than the norm (Alba, 2004; Mora et al., 2005; Villa and Rivera-Mills, 2009).

Heritage speakers of Spanish – that is, individuals who grow up in Spanish-speaking homes and who have some degree of proficiency in Spanish (Valdés, 2001) – face several hurdles within the K-12 U.S. public schooling system that limit their opportunities to maintain their home language and further develop their bilingual and bicultural identities. One key barrier is the persistent emphasis on some form of English-immersion schooling and lack of bilingual schooling options. The limited bilingual education programs that are available in the U.S. (Redford, 2018) are often subtractive in that they prioritize children’s acquisition and use of English and are designed to transition students to mainstream English classrooms as quickly as possible without support for the home language, which does not align with evidence that knowledge and skills transfer from the home language to the school language (Cummins, 2000).

On the other hand, dual language immersion (DLI) models are additive models designed to promote bilingualism, biliteracy, academic achievement, and intercultural competence among both majority-and minority-language children (Howard et al., 2018). A

growing body of longitudinal empirical evidence demonstrates the superiority of additive DLI models in terms of promoting faster English acquisition and stronger academic achievement in the long-term for language-minority students (Thomas and Collier, 2002; Marian et al., 2013; Umansky and Reardon, 2014; Steele et al., 2017; Serafini et al., 2020). Interestingly, recent research also reports that being bilingual and being enrolled in bilingual education at the primary level may impact later language study. For instance, Nguyen and Winsler (2021) found that early bilingualism in kindergarten was a significant predictor of later foreign language course enrollment and performance in middle or high school, and this held true after controlling for demographic factors, school readiness skills, and early academic achievement. Another recent study found that both Spanish-speaking and non-Spanish-speaking students who were enrolled in DLI programs continued on to advanced language study at the secondary level with high levels of performance (Padilla et al., 2022).

In spite of this evidence demonstrating the benefits of additive bilingual programs, such programs remain few and far between. Although, thirty-nine states and the District of Columbia reported offering dual language education programs during the 2012–2013 school year with Spanish as the most commonly reported partner language followed by Chinese (Boyle et al., 2015), these programs serve just a small minority of English language learners. According to a nationally representative study of the 2010–2011 school year (Redford, 2018), only 8% of kindergarteners participating in English language programs received dual language education as the primary type of instruction while 60% received English as a Second Language instruction as the primary type of instruction. Moreover, 57% of kindergarteners in English language programs did not receive *any* academic instruction in their native language (Redford, 2018).

The dominance of subtractive models of English language instruction and the lack of school-based support for minoritized languages reflects the impact of dominant monolingual ideologies, including the staffing challenges that result from a lack of pre-service training for dual language teachers. Further, scholars have recently argued that many of the DLI programs that do exist have moved away from prioritizing the dynamic bilingualism and cultural identities of Hispanic/Latina/o/x and other minoritized communities and toward serving the needs of non-Hispanic White students and students from English-speaking and/or affluent homes (Valdez et al., 2016; Flores and García, 2017; Flores et al., 2021), thus under-serving minority-language populations.

Educational opportunities to study languages other than English as a subject become more common in the secondary school context middle school (6th–8th grade) and high school (9th–12th grade). However, such opportunities most often come in the form of “Spanish as a ‘foreign’ language” classes and are typically offered as an elective for a few hours per week. In a national report (American Councils for International Education, 2017), just 20% of the school-age population enroll in world/foreign language elective courses in a given school year

(2014–2015 for this source). In other words, only around 1 in 5 U.S. public school students study a world or ‘foreign’ language at all, with an overwhelming majority of these enrolled in Spanish. Moreover, children in poverty and children of color are far less likely to have access to L2 language courses, thus providing them with less opportunity to develop multilingual skills and negatively impacting their chances for attending higher education (Darling-Hammond, 2001; Baggett, 2016). However, to date, large-scale research has not investigated which factors, such as socioeconomic status or home language, impact the language(s) that heritage speakers study in the secondary setting.

Importantly, despite the fact that a quarter of the public school-age population identifies as Hispanic/Latina/o/x (National Center for Education Statistics, 2015), the Spanish curricula in these “foreign” language courses are typically designed to meet the needs of monolingual second language (L2) learners exposed to Spanish exclusively in formal contexts, rather than draw on or develop the bilingual and bicultural knowledge, experiences, and abilities of students who acquired Spanish in home or community settings (Beaudrie et al., 2014; Parra, 2020; Leeman and Serafini, 2021). For example, heritage speakers tend to have strong listening comprehension and oral fluency skills with the ability to naturally converse about a range of daily topics. As a result, heritage speakers of Spanish may be bored in such classes or alternatively find them too difficult because they focus on metalinguistic knowledge more typical of L2 learners (Potowski, 2002). Moreover, Spanish courses often focus on so-called standard varieties spoken by monolingual elites, which may differ from the varieties that heritage speakers learn at home. Finally, there is a pervasive discourse that heritage speakers take Spanish to earn an “easy A,” which may impact their sense of belonging and deter them from studying it in a formal setting, sometimes leading them to choose a different “foreign” language (Leeman and Serafini, 2021).

This misalignment in curricular and learner needs, together with Civil-Rights-era calls for more equitable and inclusive educational policies and practices, spurred the development of the field of Spanish heritage language education (SHL; Valdés, 1981). In contrast to Spanish as a foreign language, SHL courses and programs start from the premise that students arrive in class with some prior knowledge of Spanish, whether receptive or productive knowledge. Heritage speakers’ linguistic and cultural knowledge and experiences are integrated into the curriculum (Beaudrie, 2015). Like Spanish as a foreign language course, some SHL courses also reproduce the standard language ideology, but appreciation of linguistic variation is increasingly seen as a key goal (Beaudrie et al., 2014), and in some cases, students’ critical awareness and understanding of relationships between language and power are also core objectives (Leeman, 2005, 2018; Leeman and Serafini, 2016; Holguín Mendoza, 2018; Beaudrie and Loza, in press; Holguín Mendoza, 2022). Recent evidence has demonstrated that Spanish courses and programs designed specifically for heritage speakers are linked to improved educational and emotional outcomes as well as stronger language

maintenance and student retention (Amezcu, 2019; Jang and Brutt-Griffler, 2019; Prada et al., 2021). While no comprehensive national survey data currently exist for heritage courses and program offerings or heritage student enrollment at K-12 level by state, research indicates that only a small minority of U.S. public school students have access to heritage language courses at the secondary level (Prada et al., 2021).

Motivation for the current study

The predominance of English-dominant models of schooling in the U.S. and a lack of access to both bilingual education and heritage language courses and programs, coupled with the dominance of Spanish courses geared toward L2 learners potentially make it less likely that heritage speakers of Spanish will elect to study it at the secondary level and beyond. However, to date, little to no large-scale longitudinal research has investigated what factors impact heritage student enrollment and performance in Spanish or non-English language courses in middle school or high school. One recent study conducted in a community college setting sheds some initial light on this question in a community college context (Nagano et al., 2019). In their analysis of data from a nationwide survey of heritage student enrollment in heritage languages courses versus studying a third language (L3) in community college ($N = 1,756$), almost half, or 42%, identified as heritage speakers and slightly less than half of these students (45%) reported studying an L3. In contrast, slightly over half (55%) were studying their heritage language. The authors found that “the primary reason for HL speakers not studying their HL is the lack of modern language courses offered in their HL” (p. 324), particularly for less commonly taught languages. However, given that Spanish was the most frequently offered language at the community colleges included in the sample, this did not explain the relatively high number of Spanish heritage speakers who chose to study an L3 (36%) instead of Spanish (64%), which was linked to differences in type of motivation (integrative vs. instrumental).

While Nagano et al. (2019) study offers an initial look at this issue in higher education, there is no research exploring the language course enrollment patterns of Spanish heritage speakers in the K-12 setting or what factors predict whether students choose to study Spanish in middle school/junior high (grades 6–8) or high school (grades 9–12). In general, we know that students of color have less access to and are significantly underrepresented in world language courses (Baggett, 2016). However, there is “no known literature that has reported Latino/a student enrollment patterns” (p. 163). Exploring this question empirically could not only lead to a better understanding of the language enrollment patterns of heritage speakers, but also shed light on whether heritage speakers choose to study the heritage language, a third language, or none at all. While the examination of enrollment patterns cannot not tell us about their underlying motivations or motivational profiles (see Stewart-Strobelt and Chen, 2003; Thompson, 2017), such large-scale data would be useful for

identifying “specific gateways for student enrollment, including policies regarding tracking, and school personnel that may make recommendations related to enrollment, such as guidance counselors, language teachers, and language department administrators” (Baggett, 2016, p. 175).

With these gaps in mind, the current study takes a large-scale longitudinal approach drawing on data from the Miami School Readiness Project (MSRP; Winsler et al., 2008, 2012, 2014; Serafini et al., 2020; Nguyen and Winsler, 2021). Previously, we examined foreign language learning and third language learning (L3) in the larger MSRP sample that included monolingual English students and students who spoke many other languages at home ($N=32,779$; Nguyen and Winsler, 2021). There, we found that 59.4% of all students enrolled in some type of non-English language course, and with 47.7% enrolled in Spanish, and 19.9% enrolled in other non-English language courses. Here, we follow a subsample of low-income ($N=17,341$) Spanish-speaking heritage students—those for whom parents reported Spanish to be their primary home language at school entry (kindergarten or 1st grade)—from age four through high school. Here, we investigate: (a) what percentage of heritage speakers of Spanish enroll in Spanish and other non-English language courses in middle and high school, (b) when and how long they took these courses, and (c) what demographic and early academic factors predict their enrollment and performance in Spanish courses. By providing a longitudinal look at Spanish heritage students’ secondary language study within a particular sociopolitical context, this work takes a significant step toward understanding the secondary school enrollment patterns of heritage students in Spanish (and other languages) and which factors predict their academic success in those courses.

The current study addresses the following questions. First, we ask the preliminary, descriptive (but still important) question of (1) At what rate do heritage speakers of Spanish enroll in Spanish and other non-English language courses in secondary school, and when, and for how long, do they take such language courses? Then we ask our primary research questions: (2) What factors predict Spanish heritage speakers’ enrollment in Spanish courses? and (3) what factors predict Spanish heritage speakers’ performance in Spanish (and other non-English language) courses?

Materials and methods

Context and participants

The sample in the current paper is drawn from the Miami School Readiness Project (MSRP; Winsler et al., 2008, 2014; Nguyen and Winsler, 2021). The MSRP is a cohort-sequential, longitudinal study that followed five cohorts of 4-year-old children from school entry until the end of high school. The first of the five cohorts entered kindergarten in 2002, and the last in 2007. The

children in the study were either enrolled in public school pre-K programs or qualified to receive childcare subsidies for low-income families at age 4 (Winsler et al., 2008). In the year before each cohort of children entered kindergarten, school readiness assessments were administered to evaluate their cognitive, language, socio-emotional, and motor skills. School information (including grades, courses taken, and standardized test scores) was collected every year (Winsler et al., 2008, 2012, 2014).

For the purpose of the current study, the sample included only students for whom (a) Spanish was listed as their home language at school entry (as reported by parents), (b) home language data were present in kindergarten or first grade, and (c) school transcript data were available in 6th grade or later. Our total sample was $N=17,341$ heritage speakers of Spanish (see Table 1). We use data obtained through academic year 2016–2017, during which some students in our sample were still completing high school. Cohort A and B had completed 12th grade ($n=2,838$ [16.4%]; and $n=3,413$, [19.7%], respectively), while the other three cohorts had only completed 11th grade (cohort C; $n=3,924$; 22.6%), 10th grade (cohort D; $n=3,871$; 22.3%), or 9th grade (cohort E; $n=3,295$; 19%). Students who were retained in grade (repeated a year) were also included in the sample.

The current study took place in Miami Dade County, Florida, United States. This is a linguistically and ethnically diverse area, with 70% Hispanic/Latino, 17% Black/African American, 14% White, and 75% reported speaking a language other than English at home (U.S. Census Bureau, 2020). Notably, in schools, the dominant instructional language is English, despite the prevalence of Spanish in the environment (66.3% of the population report speaking Spanish at home; U.S. Census Bureau, 2020). In the current context, there were various bilingual education programs offered in elementary school, varying from transitional bilingual education models (i.e., Mainstream-Inclusion Core/Basic Subject Areas, Mainstream-Inclusion English/Language Arts, Sheltered Core/Basic Subject Areas, Sheltered English/Language Arts) to DLI programs (i.e., One-Way Immersion, Dual Language or Two-Way Immersion; Serafini et al., 2022); however we do not have child-level data on who experienced which type of bilingual education in elementary school. In terms of the type of Spanish elective courses offered during secondary school, we know that the school system offered both “Spanish 1, 2, 3” courses as well as “Spanish for Spanish Speakers” courses. Unfortunately, we do not have child-level data on which type of Spanish course the heritage speakers in our study took. However, based on analyses for the entire MSRP sample – which includes English monolinguals and speakers of other languages (Nguyen and Winsler, 2021), we know that 55% of all students who took any type of Spanish courses were heritage speakers of Spanish. Further, group-level preliminary analyses show about 23% of all high school students studying were enrolled in a “Spanish for

TABLE 1 Demographic.

Total sample	<i>N</i> = 17,341
Has 6th grade data	<i>n</i> = 16,738
Has 7th grade data	<i>n</i> = 16,550
Has 8th grade data	<i>n</i> = 15,922
Has 9th grade data	<i>n</i> = 13,408
Has 10th grade data	<i>n</i> = 10,250
Has 11th grade data	<i>n</i> = 6,032
Has 12th grade data	<i>n</i> = 2,623
Gender	
Male	9,186 (53%)
Female	8,155 (47%)
Ethnicity	
Hispanic	16,538 (95.4%)
Black	341 (2%)
White/Asian/other	462 (2.7%)
Poverty status (6th grade)	
Received free/reduced-price lunch	13,886 (82.8%)
Did not receive free/reduced-price lunch	2,890 (17.2%)
Disability status (6th grade)	
Has a disability	3,004 (18.3%)
Non-disabled	13,438 (81.7%)
School readiness skills (nat. percentiles)	<i>M</i> (<i>SD</i>)
<i>LAP-D</i> (1–99 scale)	
Cognitive skills	50.12 (29.79)
<i>DECA</i> (1–99 scale)s	
Socio-emotional skills	59.69 (27.55)
Behavioral concerns	44.58 (29.13)
5th grade elementary academic achievement	
FCAT math (1–5 scale)	2.16 (1.32)
FCAT reading (1–5 scale)	2.17 (1.29)
GPA (0–4 scale)	3.25 (0.51)

Spanish Speakers” course; at the middle school level, 40% of students studying Spanish took Spanish courses designed for Spanish speakers. Thus, we estimate that about half of the current sample of heritage Spanish speakers took heritage language courses designed for Spanish speakers.

Students classified as dual language learners by the public-school system receive English for Speakers of Other Languages (ESOL) services and must complete a yearly English proficiency test. Once students reach the highest ESOL level determined by the school system (5), they are considered English proficient and exit the program. Thus, our sample consists of students with varying degree of proficiency in English across different elementary school years but who are all proficient in English by secondary school.

Finally, students in our sample have sufficient access to non-English language courses in secondary school, with 92.5% of middle schools and 100% of high schools in the study offering Spanish (and other) language courses (Nguyen and Winsler, 2021).

According to the school system, foreign language is not required for students to graduate high school with a ‘standard’ diploma so it is technically correct for us to use the term ‘elective’ language courses when speaking about these language courses. However, 2 years of high school foreign language classes (or demonstration of foreign language proficiency on a test at a level equal to 2 years of high school foreign language) is required for several of the more advanced college prep diploma types, and is required for application to 4-year state universities (Miami-Dade County Public Schools, n.d.).

Measures

Language learning outcomes

The MSRP tracked five cohorts of students from 2002 to 2016. The first cohort of students reached middle school (6th grade) in 2009. By 2013, all five cohorts had reached 6th grade and were presented with the opportunity to take “foreign language” courses. Students can, of course, take non-English language courses multiple times across secondary school. Thus, in the present study, students’ enrollment and performance in these classes were determined at each grade level and then combined to create overall variables capturing if students had *ever* taken any Spanish and other non-English language courses, as well as their *average* performance in these courses across all instances/years of taking language courses (for details, see Nguyen and Winsler, 2021).

Enrollment

Students were coded for *Spanish enrollment* if they had ever taken a Spanish course any time between 6th and 12th grade (1 = yes, 0 = no). Students were coded for *other non-English language enrollment* if they had ever taken a language course in another non-English language any time between 6th and 12th grade (1 = yes, 0 = no). However, as explained above, it is important to note that we do not have information about what the Spanish courses were called. Other foreign language courses that appeared on the transcripts included French, German, Chinese, Russian, Latin, Italian, Greek, Japanese, Portuguese, and Hebrew.

Performance

Students received a grade for each language course they took (original performance obtained in ordinal letter grades—A, B, etc. and converted to a 0–4 scale). Grades across all language courses were averaged to create an overall, roughly continuous, performance variable. Students’ overall averaged performance across all Spanish courses and across all other non-English language courses was calculated between 6th and 12th grade.

Student-level predictors

Demographics

Demographic information was obtained using school records. The variables of interest include: gender (1 = male, 0 = female),

students' free/reduced-price lunch status (a proxy for poverty 1 = yes, 0 = no), and disability status (1 = student has at least one of the following exceptionalities according to the district: autism, visual impairment, deafness, brain injury, learning disability, intellectual disability, speech/language disorder, emotional disturbance, or other health impairment; 0 = no; See Table 1). Notably, in addition to Hispanic students (95.4%) whose parents reported the home language is Spanish, the current sample also included a small number of Black, White, Asian, and other race/ethnicities (4.7%) who reported a home language of Spanish. Ethnicity is reported in Table 1 for demographic purposes, but due to (a) very small numbers of White, Asian, and Black individuals in this heritage Spanish sample, and (b) the fact that race was unrelated to the outcomes of interest, and results for the other variables did not change when race was included/excluded in models, race/ethnicity was not included in the regression models.

School readiness

Children were assessed for school readiness at age 4. Specifically, *cognitive skills* were measured by the cognitive subscale The Learning Accomplishment Profile-Diagnostic (LAP-D; Nehring et al., 1992) at the beginning (September/October) and end (April/May) of pre-kindergarten year. The LAP-D is a national norm-referenced developmental assessment, reliable and valid for diverse populations, with four domains: cognitive, language, fine motor, and gross motor (Winsler et al., 2008). Assessments were given in either English or Spanish based on child's strongest language as determined by the assessor and their teachers, and children were assessed individually (Winsler et al., 2008). Percentile scores from the cognitive subscale were used to measure children's cognitive skills at age 4.

In addition, parents and preschool teachers filled out the Devereux Early Childhood Assessment (DECA; LeBuffe and Naglieri, 1999) at the beginning and end of the pre-kindergarten year to measure children's *socio-emotional skills* and *behavior problems*. Higher scores for the subscales correspond to better socio-emotional skills and more behavior problems, respectively. The DECA is a nationally standardized assessment available in English and Spanish and is frequently used to measure socio-emotional skills in early childhood (Stewart-Brown and Edmunds, 2003). The DECA has 37 items in four subscales: initiative, attachment, self-control, and behavior concerns. DECA scores were determined as two main constructs: total socioemotional protective factors (TPF; 27 combining the initiative, attachment, and self-control scales) and behavioral concerns (10 items). Notably, the scale retains its integrity in linguistically and ethnically diverse and low-income children, which is important for the current sample of interest. In the MSRP, internal consistency alpha ranges from 0.71 to 0.94 (Crane et al., 2011) and does not vary by language of form (English, Spanish; see Nguyen and Winsler, 2021).

Prior/elementary school academic achievement

Prior academic achievement consists of 5th grade standardized test scores and teacher-assigned letter grades (grade

point average [GPA]) in 5th grade. Students completed the high-stakes, state-wide, Florida Comprehensive Assessment Test (FCAT) beginning in third grade to assess achievement in reading and math (Florida Department of Education, 2019a). Both a standard score and a proficiency category were given to students, with proficiency ranging from 1 (little success with the challenging content) to 5 (success with the most challenging content). In the 2010–2011 school year, the state changed the test from the FCAT to the very different FCAT-2 (Florida Department of Education, 2019a). Thus, students in our sample would have taken only one of these tests in 5th grade but since this study includes 5 cohorts, some took different versions of the test. Due to this discrepancy, FCAT proficiency ordinal scores (1–5 scale) were included in our analyses instead of the standard scores (which were on different scales; Florida Department of Education, 2019a). The FCAT (English) reading score in fifth grade was used as a covariate measure of prior academic language performance. In addition, we conducted additional analyses replacing the reading score with the math score. Theoretically, the FCAT math score would demonstrate student general ability, while the reading score would be more influenced by student English language skills (for further details, see Nguyen and Winsler, 2021). Finally, 5th grade GPA consist of the overall average teacher-assigned letter grades across all subject areas (converted into a 5-point scale: 4.0 = A, 3.0 = B, 2.0 = C, 1.0 = D, 0.0 = F). Student GPA in 5th grade was used as a covariate measure of overall prior academic performance.

Analyses

Descriptive statistics were used to answer the first question which focused on the percentage of heritage speakers who enrolled in Spanish and other non-English language courses, as well as the timing and length of that enrollment. For the last two questions, which investigated predictors of heritage speakers' enrollment and performance in language classes, data were analyzed using hierarchical multiple regression (logistic regression for enrollment, linear regression for performance). The first block included demographic variables (gender, poverty, and disability status); the second block included early school readiness skills at age 4 (cognitive skills, social skills, and behavior concerns); and the final block included prior academic achievement (fifth grade GPA, and test scores).

Missing data

Due to the longitudinal and school-based, real-world nature of the study, there were missing data on some predictors as well as attrition in the sample over time. Since our inclusion criteria required that students have at least some 6th grade or later data, and all students had a chance to reach 9th grade based on their age/cohort, we defined longitudinal attrition as middle school students who did not have any high school data (grade 9 or above) meaning that the student would have left the public school system before 9th grade and did not return. Across the

full sample, 14.2% did not have any high school data (left the public school system). Given the large sample size, we only note correlations between missingness and relevant variables that are greater than $r=0.10$. Missing high school information moderately correlated with several predictors and outcomes, including disability ($r=-0.11$), FCAT reading and math ($r=0.35$ and $r=0.34$, respectively), Spanish enrollment ($r=0.20$), Spanish and other non-English language course grade ($r=0.19$ and $r=0.12$, respectively). In sum, students missing high school data were less likely to have a disability and had higher initial achievement (in general and for middle school Spanish/language classes) than students who remained in the sample.

Missing data for predictors were less than 17% of cases, with the exception of cognitive skills (39.7% missing) and 5th grade GPA (23.7% missing). Some correlations between the missingness on predictors and the outcome were moderate, including disability ($r=0.30$), cognitive skills ($r=-0.24$), social skills ($r=-0.14$), FCAT reading and math ($r=-0.30$ and $r=-0.29$, respectively), Spanish enrollment ($r=-0.20$), and Spanish and other non-English languages performance ($r=-0.20$ and $r=-0.13$, respectively). Thus, students missing a predictor variable were more likely to have a disability, tended to score lower on school readiness skills, and were less likely to take (and performed poorer in) Spanish and other language courses compared to students with no missing predictors.

We first ran the set of analyses described above using listwise deletion in IBM SPSS Statistics, then conducted additional analyses in R (<https://www.R-project.org/>) using the lavaan package (Rosseel, 2012) to use full information maximum likelihood (FIML) to adjust for missing data on the predictors.

Results

RQ1: Enrollment in Spanish and other non-English language courses

The first research question was answered using descriptive statistics. Frequencies were used to analyze the number of students taking Spanish and other non-English language courses in each grade level (Table 2). Of the 17,341 heritage Spanish speakers in our sample, 11,414 (65.8%) enrolled in some type of language course at least once from grade 6 to grade 12. More specifically, 52.2% enrolled in Spanish courses and 25.3% enrolled in other non-English language courses; it should be noted that these categories are not mutually exclusive, as some students took both types of language courses. In general, the percentage of enrollment in heritage Spanish courses is higher, sometimes over twice the percentage of students enrolling in other non-English language courses. In terms of enrollment patterns, within the 11,414 students mentioned above who took some type of language course in secondary school, 61.5% ($n=7,019$) of heritage speakers only enrolled in Spanish courses, 20.8% ($n=2,368$) only enrolled in other non-English language courses, and 17.7% ($n=2,012$) enrolled in both types of courses in secondary school.

Table 3 depicts the grade at which heritage Spanish students first took a non-English language course. Among heritage Spanish speakers enrolled in non-English languages courses in secondary school, most students first took Spanish in sixth (27%), ninth (20.1%), or tenth (21.8%) grade. Similarly, most students who took other non-English languages courses also began in these grades (32, 28.3, and 18.4% respectively). Sixth grade is the first grade of middle school and 9th grade the first year of high school in this district, suggesting that students tend to take heritage language courses most when they enter the next level of schooling, and it is less common to start taking languages later.

In terms of Spanish enrollment patterns, of the heritage Spanish speakers who took some type of Spanish course, 14.6% ($n=1,208$) enrolled only in middle school, 57.2% ($n=4,723$) enrolled only in high school, and 28.2% ($n=2,335$) enrolled at least once in both middle school and high school. Similarly, of the students who took some type of other non-English language course, 23.7% ($n=944$) enrolled only in middle school, 58.4% ($n=2,321$) enrolled only in high school, and 17.9% ($n=711$) enrolled in both middle school and high school.

Table 4 shows the total number of grades in which students took non-English language courses. Of the 9,043 heritage students who took Spanish courses, most students only enrolled in one (39.2%) or two (34.9%) Spanish courses/years. Of the 4,383 heritage students who took some type of other non-English language course, the largest group of students also only enrolled in one (43.8%) or two (32.5%) other non-English language courses/years. Interestingly, a small number of students enrolled in these courses for all 7 years of secondary school (23 and 20 in Spanish and other non-English language courses, respectively).

RQ2: What factors predict heritage Spanish speakers' enrollment in Spanish courses?

For this question, a hierarchical logistic regression was conducted with the entire sample ($N=17,341$). Table 5 shows the results where odds ratios (OR) are provided. An OR greater than 1 indicates an increase in the odds of taking Spanish courses, and an OR less than 1 indicates a decrease in the odds of taking Spanish courses. For categorical variables, the OR is a function of being on one level of the variable (i.e., male) compared to the other (female). For continuous variables, OR indicates the increase/decrease in odds of FL enrollment with a 1-point increase in the predictor.

Overall, in model 1 with only demographic predictors, poverty status and disability status, but not gender, uniquely predicted Spanish courses enrollment. Specifically, poverty appeared to hinder heritage speakers' enrollment in Spanish courses, as those who receive free or reduced-price lunch had lower odds of taking a Spanish course than those who did not ($OR=0.712$, $p<0.001$). Similarly, heritage speakers with disabilities had significantly lower odds of Spanish course enrollment compared to those without disabilities ($OR=0.504$, $p<0.001$).

TABLE 2 Enrollment in non-English language courses of heritage Spanish students courses in secondary school by year and type of course.

	Ever in middle school		Ever in high school		Ever in secondary school		Grade 6		Grade 7		Grade 8		Grade 9		Grade 10		Grade 11		Grade 12	
Total	17,178		13,567		17,329		16,738		16,550		15,922		13,408		10,250		6,032		2,623	
Type of course	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Any non-English	5,667	33%	9,485	69.9%	11,414	65.8%	3,413	20.4%	3,472	21%	3,486	21.9%	4,958	37%	5,973	58.3%	3,257	54%	1,239	47.2%
Spanish	4,242	24.7%	7,137	52.6%	9,043	52.2%	2,351	14%	2,397	14.5%	2,573	16.2%	3,241	24.2%	4,178	40.8%	2,370	39.3%	977	37.2%
Other non-English/Spanish	2,034	11.8%	3,060	22.6%	4,383	25.3%	1,354	8.1%	1,331	8%	1,088	6.8%	1,779	13.3%	1,881	18.4%	917	15.2%	274	10.4%

Course categories are not mutually exclusive; a student can enroll in both Spanish and other non-English/Spanish courses. Percentage is calculated using the total sample for that grade/time period.

In model 2, when student school readiness skills were added, only early behavior concerns predicted later Spanish course enrollment for heritage speakers. A one-point increase in behavior concerns at age 4 was associated with 0.003 decrease in enrollment odds. In other words, a student in the 50th percentile compared to a student in the 25th percentile (a 25-point difference) in behavior problems at school entry would have a 7.5% increased chance (25×0.003) of enrolling in Spanish courses in secondary school. Poverty and disability status remained significant negative predictors even after controlling for school-entry skills.

In the last block, student 5th grade achievement was entered. GPA and test scores in 5th grade significantly predicted the odds of taking at least one Spanish course in middle or high school. A 1-point increase in GPA (e.g., moving from a C to a B) increased the odds of enrolling in Spanish courses by 53.5%. Similarly, a 1-point increase in test scores (on a 1–5 point scale) increased the odds of enrollment by about 15% for both math and reading. Interestingly, lunch status, disability status, and early behavior concerns remained significant. Additional analyses were conducted with a maximum likelihood estimator to account for missing data. Findings remained similar, with only minor differences. Poverty status was no longer a significant predictor. In the model with FCAT reading, cognitive skill became a significant predictor ($OR=1.000$, $p=0.016$); in the model with FCAT math, gender was a significant predictor ($OR=0.978$, $p=0.003$). Overall, heritage students who had a disability or had more behavior problems were less likely to enroll in Spanish than their counterparts. In addition, those with higher prior GPAs and standardized test scores were more likely to enroll in Spanish than those with lower grades and scores.¹

RQ3: What factors predict heritage Spanish speakers' performance in language courses?

Spanish language course performance

For this research question, two hierarchical multiple regression analyses were conducted for each language type (Spanish vs. other non-English language courses). Table 6 shows results for the analyses concerning Spanish course performance. The model significantly predicted average performance (GPA) across all Spanish courses taken by students, $F(3,4,656)=109.396$, $p<0.001$, $R^2_{\text{adjusted}}=0.065$. On

¹ The same analyses were conducted to predict other non-English language course enrollment. Similar to analyses for Spanish courses, disability status and elementary academic achievement were significant predictors. However, poverty did not predict non-English language course enrollment. In addition, in the first two steps, gender was a significant predictor, such that male students were less likely to enroll in other non-English language courses than female students, a result not seen for Spanish enrollment.

TABLE 3 The grade at which heritage Spanish students first took a non-English language course.

Grade	Spanish course		Other non-English/ Spanish courses	
	N	%	N	%
6	2,351	27.0%	1,354	32.0%
7	1,028	11.8%	436	10.3%
8	800	9.2%	226	5.3%
9	1,750	20.1%	1,197	28.3%
10	1,893	21.8%	779	18.4%
11	694	8.0%	214	5.1%
12	184	2.1%	29	0.7%
Total	8,700		4,235	

TABLE 4 Total number of grades heritage Spanish students took non-English language courses.

Number of grades taken	Spanish courses		Other non-English/ Spanish courses	
	N	%	N	%
1	3,545	39.2%	1,918	43.8%
2	3,158	34.9%	1,424	32.5%
3	1,491	16.5%	598	13.6%
4	594	6.6%	255	5.8%
5	176	1.9%	104	2.4%
6	56	0.6%	64	1.5%
7	23	0.3%	20	0.5%
Total	9,043		4,383	

TABLE 5 Logistic regression predicting Spanish enrollment in secondary school ($n=7,382$).

	Model 1		Model 2		Model 3	
	OR	SE(B)	OR	SE(B)	OR	SE(B)
Demographics						
Male	0.960	0.049	0.995	0.050	1.063 ^c	0.051
Lunch (poverty)	0.712***	0.068	0.747***	0.069	0.848*	0.070
Special education	0.504***	0.104	0.552***	0.106	0.671***	0.108
School readiness at age 4						
LAP-D cognitive skills			1.002	0.001	1.000 ^c	0.001
DECA social skills			1.001	0.001	1.000	0.001
DECA behavior concerns			0.997**	0.001	0.998*	0.001
Elementary academic performance						
GPA in 5th grade					1.535***	0.055
^a Reading 5th grade					1.144***	0.020
^a Math in 5th grade					1.155***	0.020

* $p < 0.05$, ** $p < 0.01$, *** $p \leq 0.001$. ^aMath and reading scores were run in different models to avoid multicollinearity. ^cResults are significant when analyzed with FIML.

average, male students, students in poverty, and students with a disability had lower grades than their counterparts ($B_{\text{male}} = -0.349$, $B_{\text{lunch}} = -0.245$, and $B_{\text{disability}} = -0.375$, all $p < 0.001$, respectively).

In step 2, early school readiness skills were added to the model, and the model significantly predicted student performance in Spanish language courses ($F(6,4,653) = 71.56$, $p < 0.001$, $R^2_{\text{adjusted}} = 0.083$), and significantly improved the prediction compared to model 1 ($R^2_{\text{change}} = 0.018$, $p < 0.001$). In general, students with higher cognitive and social skills at age 4 outperformed students with lower scores in these skills 7+ years later ($B = 0.003$, $p < 0.001$; and $B = 0.001$, $p = 0.043$, respectively). In addition, students with lower behavior problems in preschool had higher average scores in Spanish courses compared to those with higher early behavior concerns ($B = -0.002$, $p < 0.001$). Demographic factors remained significant.

The last model included the influence of 5th grade achievement on secondary school Spanish course performance. Model 3 significantly predicted students' performance in Spanish courses ($F(8, 4,651) = 413.297$, $p < 0.001$, $R^2_{\text{adjusted}} = 0.196$), above and beyond compared to the previous model ($R^2_{\text{change}} = 0.113$, $p < 0.001$). Notably, GPA ($B = 0.66$,

$p < 0.001$), but not math or reading scores predicted Spanish language performance for heritage Spanish students, such that higher 5th grade GPA was associated with better grades later in Spanish classes in secondary school. Interestingly, only gender and poverty status remained significant predictors in this model with 5th grade performance added. The same analyses were conducted with FIML to account for missing data. Findings remained similar, although disability status ($b = -0.103$, $p = 0.001$) and behavior concerns ($b = -0.001$, $p = 0.032$) became significant predictors as well. Overall, male students, students in poverty, students with a disability, and those with more behavior problems than their peers had lower grades in Spanish courses. In addition, students with a higher GPA in 5th grade displayed higher Spanish course performance.

Other non-English language course performance

Similar patterns were found for other non-English/Spanish language course performance across secondary school. Overall, model 1 significantly predicted performance and all the same predictors were associated with student grades in other

TABLE 6 Multiple regression predicting Spanish course performance in secondary school ($n=4,660$).

	Model 1			Model 2			Model 3		
	<i>B</i>	<i>SE(B)</i>	β	<i>B</i>	<i>SE(B)</i>	β	<i>B</i>	<i>SE(B)</i>	β
Demographics									
Male	−0.349***	0.024	−0.207	−0.322***	0.024	−0.191	−0.239***	0.023	−0.142
Lunch (poverty)	−0.245***	0.031	−0.112	−0.197***	0.031	−0.090	−0.084**	0.029	−0.038
Special education	−0.375***	0.061	−0.088	−0.286***	0.061	−0.067	−0.077 ^c	0.058	−0.018
School readiness at age 4									
LAP-D cognitive skills				0.003***	0.000	0.093	0.000	0.000	0.016
DECA social skills				0.001*	0.001	0.033	0.000	0.001	0.006
DECA behavior concerns				−0.002***	0.000	−0.066	−0.001 ^c	0.000	−0.028
Elementary academic performance									
GPA in 5th grade							0.660***	0.027	0.367
^a Reading 5th grade							0.005	0.009	0.008
^a Math in 5th grade							0.009	0.009	0.014
R^2	0.065			0.083			0.196		
R^2 change				0.018***			0.113***		

* $p < 0.05$, ** $p < 0.01$, *** $p \leq 0.001$. ^aMath and reading scores were run in different models to avoid multicollinearity. ^cResults are significant when analyzed with FIML.

non-English courses. In model 2, only cognitive skills and behavior concerns predicted other non-English/Spanish language course performance—children with higher cognitive skills and lower behavior concerns at age 4 achieved higher grades in other non-English language courses over 7 years later. However, in model 3, school readiness skills were no longer significant predictors. Similar to findings about Spanish courses, only gender and poverty status remained significant demographic predictors in this model. Notably, unlike the findings about Spanish performance where only GPA was a significant predictor, GPA and test scores ($B_{\text{reading}}=0.048$, $B_{\text{math}}=0.44$, all $p < 0.001$) significantly predicted performance in other non-English language courses, with higher grades and test scores associated with higher grades in these courses. The final model accounted for 25.5% of variance in other non-English language course performance of heritage Spanish students (Table 7). Additional analyses were conducted using FIML to account for missing data. Findings remained the same. Overall, male students and students in poverty had lower grades in other non-English language courses compared to their counterparts, and students with higher 5th-grade achievement performed better in these courses.

Discussion

To the knowledge of the authors, the current paper is the first of its kind to longitudinally explore the language study of heritage Spanish speakers in Spanish or non-English language courses in secondary school at a large scale in the United States. We found that the predominantly low-income, heritage Spanish speakers in this community enrolled in language courses at a high rate in every grade (20.4–58.3%) and in total (66%) in secondary school, and heritage speakers enrolled in Spanish at almost twice the rate as other

non-English language courses. Nationally for all K-12 students, the estimate is that only about 20% of students take non-English language courses in school (American Councils for International Education, 2017) in a given school year. This level of interest in learning non-English languages is notable and likely reflects not only serious commitment on the part of the students to maintain Spanish and master an L3, but also demonstrates the dedication of the school system, within this context, to provide students with opportunities to access and pursue language learning throughout the secondary years. Given that 2 years of second language courses is required in high school for students to receive the more advanced, college bound diploma types, these high rates of language course enrollment may also reflect high educational aspirations for Spanish heritage students.

The preference to study Spanish observed here is also consistent with national trends showing that enrollment in Spanish is more than three times the total enrollment in other major languages including French, Arabic, Chinese, German, Japanese, Latin, and Russian (American Councils for International Education, 2017). Enrollment patterns varied across grade levels and types of language, with some students taking multiple languages or continuously enrolling in the same type of languages (Spanish or other non-English languages) over several years. Our findings also provide large-scale longitudinal evidence as to which student background variables significantly predict enrollment and performance in these language courses. Overall, findings ssp contribute to better understanding the broader picture of Spanish-speaking heritage students’ language enrollment patterns in secondary school. Moreover, the current research has key implications for understanding and supporting heritage students’ continued language learning and long-term language maintenance, and the crucial importance of offering equal opportunities for students to access heritage courses specifically designed to meet their needs.

TABLE 7 Multiple regression predicting other non-English language course performance in secondary school ($n=2,325$).

	Model 1			Model 2			Model 3		
	<i>B</i>	<i>SE(B)</i>	β	<i>B</i>	<i>SE(B)</i>	β	<i>B</i>	<i>SE(B)</i>	β
Demographics									
Male	−0.470***	0.043	−0.219	−0.429***	0.043	−0.200	−0.294***	0.040	−0.137
Lunch/poverty	−0.410***	0.057	−0.145	−0.330***	0.057	−0.116	−0.136*	0.053	−0.045
Special education	−0.369**	0.127	−0.058	−0.251***	0.127	−0.040	−0.057	0.115	−0.009
School readiness at age 4									
LAP-D cognitive skills				0.004***	0.001	0.118	0.001	0.001	0.023
DECA social skills				0.002	0.001	0.046	0.000	0.001	0.010
DECA behavior concerns				−0.002*	0.001	−0.048	0.000	0.001	−0.009
Elementary academic performance									
GPA in 5th grade							0.988***	0.048	0.419
^a Reading 5th grade							0.048***	0.015	0.058
^a Math in 5th grade							0.044***	0.017	0.056
R^2		0.073			0.094			0.255	
R^2 change					0.021***			0.161***	

* $p < 0.05$, ** $p < 0.01$, *** $p \leq 0.001$. ^aMath and reading scores were run in different models to avoid multicollinearity.

Enrollment of heritage speakers in non-English language courses

To date, there has been little research on the non-English language course-taking of Spanish heritage language speakers at the secondary school level in the United States. Indeed, prior research primarily focused on students in higher education (Potowski, 2002; Brown and Thompson, 2018; Looney and Lusin, 2019; Nagano et al., 2019), even though evidence suggests that heritage learners' proficiency in high school positively predicts college academic attainment (Jang and Brutt-Griffler, 2019), and starting language courses earlier (i.e., middle school as opposed to high school) leads to higher motivation for language learning in students (Kissau et al., 2015). There are no current national statistics regarding the rate of enrollment for heritage students in Spanish or other non-English language courses, even though this is one of the fastest growing groups of students in the U.S. (National Center for Education Statistics, 2018).

In the current sample, 65.8% of Spanish heritage students enrolled in some type of language courses in secondary school; within this group, around 62% enrolled in Spanish, 20% enrolled in other non-English language courses, and 18% enrolled in both types of language courses. Notably, the rate of general non-English language course enrollment is slightly higher in heritage students compared to the larger sample of students in the MSRP ($n = 33,247$, 59.4%), especially in the rate of Spanish enrollment (47.7% in full MSRP sample, 62% in the heritage student sample; Nguyen and Winsler, 2021). This is consistent with prior research suggesting that students who speak multiple languages may be more inclined to enroll in additional language courses at the secondary school level (Nguyen and Winsler, 2021). Although we do not know which type

of Spanish courses individual students took, we know that about 30% of the courses were "Spanish for Spanish Speakers" classes designed specifically to meet the needs of heritage speakers. These are promising numbers as heritage students are known to experience a range of benefits from home language study in heritage language courses, including ethnolinguistic pride, heritage language maintenance, and increased student motivation and persistence/retention (Carreira, 2000; Leeman et al., 2011; Amezcua, 2019; Prada et al., 2021; Serafini, 2021; Holguín Mendoza, 2022).

In addition to general enrollment rates, we explored the timing of enrollment in Spanish and other non-English courses for Spanish heritage students. Students usually began enrollment at the beginning of middle school (6th grade), or the first 2 years of high school (9th and 10th grades). It is possible that the bump seen in language course enrollment at the beginning of middle and high school is due to guidance counselors emphasizing the importance of language courses during these school transitions. Notably, across all types of language courses, most students enrolled in language courses in high school rather than middle school, with a sizeable minority enrolled continuously in middle school and high school. Given that college entrance sometimes requires demonstration of other language proficiency, perhaps students are encouraged to begin taking language courses at the high school level given its relevance, in line with national trends and prior research showing greater language enrollment in high school than in middle school (Pufahl and Rhodes, 2011; American Councils for International Education, 2017; National Center for Education Statistics, 2018; Nguyen and Winsler, 2021).

Heritage students who enrolled either in Spanish or other non-English language courses in middle school might have a higher level of interest and motivation in maintaining and improving their home language or learning novel languages given

their experience being bilingual (Nguyen and Winsler, 2021). This “cyclical bilingualism,” in which adolescents seek to reacquire or further develop the heritage language(s) they spoke in childhood, is well documented (Silva-Corvalán, 1994; Villa and Rivera-Mills, 2009). In addition, those who enrolled in language courses in both middle school and high school may be uniquely different than their peers, and more likely to experience both short and long term cultural and social advantages associated with the continued pursuit of language learning (Wight, 2015). When students did take language courses, they usually enrolled for one to 2 years rather than longer, and very few students enrolled for as long as 7 years. This is reasonable as students also must take many other subjects across the middle school and high school years; as language courses are usually electives, they may not be high on the priority of courses to be taken continuously, unlike math and literature. Indeed, most states do not require foreign language study for graduation (AICE, 2017; Met and Brandt, 2017), and language courses were not mandatory in the state of Florida at the time of the present study (Florida Department of Education, 2019b).

Predicting heritage speakers’ enrollment in Spanish courses

In addition to enrollment patterns in Spanish and other language courses, we explored student-specific factors that may predict heritage speakers’ enrollment in Spanish courses. Findings reveal that students in poverty, students who had a disability, and students who displayed more behavior problems early on had a lower likelihood of enrolling in Spanish courses compared to their peers, while those with better elementary school achievement were more likely to take these courses. Indeed, students in poverty may not consider elective language courses to be a high priority as they are faced with additional responsibilities in the household or experience additional stressors that prevent them from devoting time to schoolwork (Jensen, 2009). Similarly, students with a disability may actively choose not to take language courses even when it is their home language due to the belief that they lack the academic skills to succeed and feel less positive about language learning despite wanting to learn (Sparks et al., 1993), and they may also be discouraged to enroll in them due to false assumptions of teachers and counselors about students’ abilities (Sparks, 2016). Sustained opportunities to build literacy in and maintain one’s home language can be framed not only as an individual ‘right’, but also as contributing to strengthening a collective ‘resource’ (Peyton et al., 2001). Further, language courses are beneficial to all students as they can provide pragmatic, cognitive, and cultural gains (Sparks, 2016); thus, educators have an ethical responsibility to encourage heritage Spanish students with a disability to take Spanish courses and to advocate for more systematic heritage course offerings at all levels of education.

Another factor associated with a lower likelihood of Spanish enrollment for heritage Spanish speakers was greater preschool behavioral problems. Prior research conducted in the same

population showed that behavior problems were linked to slower English attainment in dual language learners (Winsler et al., 2014) and poorer academic performance later on in general (Ricciardi et al., 2021). Relatedly, students with higher 5th grade GPA and math and reading test scores were more likely to enroll in Spanish courses than their peers, which is consistent with prior findings in the larger population about language course enrollment more generally (Nguyen and Winsler, 2021). Students who perform well in school may be more likely to be encouraged to take different kinds of electives including languages as ‘enrichment’ or they have more freedom to choose additional electives instead of having to take remedial courses and extra study halls as is often required for students struggling academically.

In sum, Spanish heritage students’ background and achievement appear to influence whether they enroll in Spanish courses, given that access to these courses is not a major issue (92.5% of middle schools and 100% of high schools in our sample offered foreign language courses including Spanish; Nguyen and Winsler, 2021). Our findings are consistent with prior research about general language learning for all students in the current context (Nguyen, 2020) and provide additional understanding for researchers and educators of the factors predicting heritage students’ Spanish course enrollment at the secondary level.

One important perspective for interpreting factors associated with enrollment is that access to Spanish classes and language maintenance are equity issues. The type of students who tend not to enroll in Spanish language courses are students who may lack necessary resources available to them (those in poverty, those with disabilities) and might be the most likely to benefit from such courses. It is well-documented that minority language populations such as Spanish heritage speakers are underserved within the U.S. education system due, at least in part, to a lack of resources. As previously discussed, a key obstacle is the gentrification of bilingual education models, particularly those that are known to be most effective in closing the ‘gap’ among language majority and language minority children (Serafini et al., 2022). That is, schools that do offer DLI programs have been critiqued for catering to the needs of non-Hispanic White students and students from English-speaking and/or affluent homes (Flores and García, 2017; Flores et al., 2021). Thus, certain groups of students in our sample may be disadvantaged in multiple ways, being a Spanish heritage speaker while also being in poverty or having a disability. It is crucial that educators be aware of these intersectional structural inequities to better serve these students.

Performance of heritage speakers in non-English language courses

Beyond enrollment, we were also interested in performance of heritage students in both Spanish and other non-English language courses. Findings across the different language types are

similar; male students, students in poverty, students with a disability, and students with more preschool behavior problems had lower grades than their counterparts, while those with higher 5th grade GPA outperformed their peers in language courses. This is consistent with enrollment findings as well as previous findings about language course performance in the same population including all students (Nguyen and Winsler, 2021). Certain demographic effects went away when 5th-grade achievement was included as predictors (i.e., poverty, disability, school readiness), suggesting that performance in elementary school was the strongest predictor of performance in secondary school language courses.

An interesting difference between the findings of Spanish performance and other non-English language performance lies in the effect of 5th-grade test scores. Specifically, standardized test scores did not predict heritage student performance in Spanish courses but did predict performance in other non-English language courses. It is possible that heritage students who are learning Spanish, their home language, in a classroom environment are more engaged and invested than when in other courses, and performance in Spanish is less linked to traditionally assessed skills such as math and reading. Another possibility is that Spanish language courses (or *Spanish for Spanish Speakers* courses) do not tap into the type of knowledge assessed in standardized high-stakes testing. Performance in courses for which students have little prior knowledge (such as other non-English languages for Spanish speakers) appears more associated with general learning abilities such as reading and math.

Limitations, implications, and future directions

The current study filled a gap in the literature concerning factors that impact heritage student enrollment and performance in Spanish or non-English language courses in middle school and high school. A limitation, however, is that the findings may not generalize to other settings and populations outside of Miami that have different ethnicity distributions and less ethnolinguistic vitality and sociolinguistic support in the community for Spanish language use and maintenance. We also lacked child-level information about whether students took “foreign language” Spanish courses vs. courses designed specifically for heritage Spanish learners. In spite of these limitations, the current paper offers novel descriptive information on Spanish and other non-English course enrollment patterns in secondary school for heritage Spanish speakers, and new understandings of factors related to heritage students’ enrollment and performance in Spanish and other non-English language courses. In addition, our longitudinal study design provided rich, robust evidence which allowed us to characterize trends of Spanish enrollment over time in heritage students, which have not been investigated previously. While our findings clearly do not inform us as to *why* heritage

Spanish speakers chose to take or not take Spanish courses, prior research suggested that the discrepancy between students’ knowledge and course design, as well as educator beliefs may contribute to heritage student course selections (Potowski, 2002; Beaudrie et al., 2014; Parra, 2020; Leeman and Serafini, 2021). Future qualitative studies are needed to enhance our understanding of the lived experiences of heritage Spanish speakers, and their decision-making processes and motivations when it comes to selecting elective courses in secondary school.

Conclusion

Heritage Spanish students are among the fastest growing group of K-12 students in the United States. Supporting and maintaining their home language is important concern not only at an individual level, but as a societal level as well. In general, there is a lack of Spanish language courses and programs designed to meet the need of these students. Specifically, English is typically the only language of instruction in the school system (Cummins, 2005; Wiley and García, 2016; Ennser-Kananen and King, 2018; Fuller and Leeman, 2020), and DLI courses are not usually available or may have moved away from prioritizing the dynamic bilingualism and cultural identities of Latinx and other minoritized communities (Flores and García, 2017; Flores et al., 2021). Further, Spanish courses at the secondary level often come in the form of Spanish as a *foreign* language rather than *heritage* language (Potowski, 2002).

The current paper contributes to the literature by describing Spanish course taking among heritage Spanish-speaking students and identifying factors related to heritage student enrollment and performance in Spanish or other non-English language courses in middle school or high school. Overall, heritage students enrolled at high rates in Spanish and other non-English language courses, with enrollment patterns varying across the grade levels, similar to the general population of K-12 students in the United States. Notably, student background and early achievement can predict both enrollment and performance of heritage students in these language courses. Thus, student motivation and goals may not be the only important component that leads to Spanish heritage speakers’ choices to pursue advanced language study as is often assumed. Our findings emphasize the need to apply a critical lens to the individual and social implications of U.S. language education, and provide useful insights for informing language education policy and underscoring the need for more systematic efforts to advocate for the needs, rights, and resources of language minority students.

Data availability statement

The data analyzed in this study are subject to the following licenses/restrictions: Data are not available upon request because they belong to the public school system. The data use agreement

in play does not allow other researchers to access the data. Questions about this should be directed to AW awinsler@gmu.edu.

Ethics statement

The studies involving human participants were reviewed and approved by the original participating school system's Institutional Review Board (IRB; Miami-Dade County Public Schools IRB 09141-01) and the participating university's IRB (George Mason University #477930-9). Written informed consent to participate in this study was provided by the participant's legal guardian/next of kin.

Author contributions

MN and AW conceived of the presented idea. Data were previously collected in AW's laboratory. MN performed all analyses. ES and JL contributed to the background theory and

framework as well as the parameters for data analysis. All authors discussed the results and contributed to the final manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Alba, R. (2004). Language Assimilation Today: Bilingualism Persists More Than in the Past, But English Still Dominates. Lewis Mumford Center for Comparative Urban and Regional Research. Available at: http://mumford.albany.edu/children/reports/language_assimilation/language_assimilation01.htm (Accessed July 15, 2022).
- American Councils for International Education. (2017). The National K-12 Foreign Language Enrollment Survey Report. Available at: <https://www.americancouncils.org/> (Accessed July 15, 2022).
- Amezcu, A. (2019). An analysis of Spanish language maintenance motivation in a heritage learning classroom. *Span. Port. Rev.* 5, 73–85.
- Baggett, C. (2016). Student enrollment in world languages: L'Egalité des Chances? *Foreign Lang. Ann.* 49, 162–179. doi: 10.1111/flan.12173
- Beaudrie, S. M. (2015). Approaches to language variation: goals and objectives of the Spanish heritage language syllabus. *Herit. Lang. J.* 12, 1–21. doi: 10.46538/hlj.12.1.1
- Beaudrie, S. M., Ducar, C., and Potowski, K. (2014). *Heritage Language Teaching: Research and Practice*. Columbus, OH: McGraw.
- Beaudrie, S. M., and Loza, S. (in press). Insights into SHL program direction: student and program advocacy challenges in the face of ideological inequity. *Lang. Aware.*, 1–19. doi: 10.1080/09658416.2021.1949333
- Boyle, A., August, D., Tabaku, L., Cole, S., and Simpson-Baird, A. (2015). *Dual Language Education Programs: Current State Policies and Practices* Office of English Language Acquisition, US Department of Education.
- Brown, A. V., and Thompson, G. L. (2018). *The Changing Landscape of Spanish Language Curricula: Designing Higher Education Programs for Diverse Students* Georgetown University Press.
- Carreira, M. (2000). Validating and promoting Spanish in the United States: lessons from linguistic science. *Biling. Res. J.* 24, 423–442. doi: 10.1080/15235882.2000.10162776
- Crane, J., Mincic, M., and Winsler, A. (2011). Parent-teacher agreement and reliability on the Devereux early childhood assessment (DECA) in English and Spanish for ethnically diverse children in poverty. *Early Educ. Dev.* 22, 520–547. doi: 10.1080/10409289.2011.565722
- Crawford, J. (2000). *At War with Diversity: US Language Policy in an Age of Anxiety*. Multilingual Matters.
- Cummins, J. (2000). *Language, Power, and Pedagogy. Bilingual Children in the Crossfire* Multilingual Matters.
- Cummins, J. (2005). A proposal for action: strategies for recognizing heritage language competence as a learning resource within the mainstream classroom. *Mod. Lang. J.* 89, 585–592.
- Darling-Hammond, L. (2001). "Inequality in teaching and schooling: how opportunity is rationed to students of color in America" in *The Right Thing to Do, the Smart Thing to Do: Enhancing Diversity in Health Professions -- Summary of the Symposium on Diversity in Health Professions in Honor of Herbert W. Nickens, M.D.* eds. B. D. Smedley, A. Y. Stith, L. Colburn and C. H. Evans (Washington, D.C.: National Academies Press)
- Ennsner-Kananen, J., and King, K. A. (2018). "Heritage languages and language policy" in *The Encyclopedia of Applied Linguistics*. ed. C. A. Chapelle (John Wiley & Sons), 1–6.
- Fishman, J. (2001). "300-plus years of heritage language education in the United States" in *Heritage Languages in America: Preserving a National Resource*. eds. J. K. Peyton, D. A. Ranard and S. McGinnis (Washington, DC: Center for Applied Linguistics & Delta Systems), 81–98.
- Flores, N., and García, O. (2017). A critical review of bilingual education in the United States: from basements and pride to boutiques and profit. *Annu. Rev. Appl. Linguist.* 37, 14–29. doi: 10.1017/S0267190517000162
- Flores, N., Tseng, A., and Subtirelu, N. (2021). *Bilingualism for all: raciolinguistic perspectives on dual language education in the United States* Multilingual Matters.
- Florida Department of Education (2019a). FCAT Historical. Available at: <http://www.fldoe.org/accountability/assessments/k-12-student-assessment/archive/fcat/> (Accessed July 15, 2022).
- Florida Department of Education (2019b). World Languages (Foreign Languages). Available at: <http://www.fldoe.org/core/fileparse.php/7583/urlt/06-05-152015-CELLA-FAQs.pdf> (Accessed July 15, 2022).
- Fuller, J. M., and Leeman, J. (2020). *Speaking Spanish in the US: The sociopolitics of language* (2nd edition). Bristol: Multilingual Matters.
- Holguín Mendoza, C. (2018). Critical Language Awareness (CLA) for Spanish Heritage Language Programs: Implementing a Complete Curriculum. *Int. Multiling. Res. J.* 12, 65–79. doi: 10.1080/19313152.2017.1401445
- Holguín Mendoza, C. (2022). "Sociolinguistic justice and student agency in language education: towards a model for critical sociocultural linguistics literacy" in *Heritage Language Teaching: Critical Language Awareness Perspectives from Research and Pedagogy*. eds. S. Loza and S. Beaudrie (London, UK: Routledge), 138–156.
- Howard, E. R., Lindholm-Leary, K. J., Rogers, D., Olague, N., Medina, J., Kennedy, B., et al. (2018). *Guiding Principles for Dual Language Education*. 3rd edn. Washington, DC: Center for Applied Linguistics.
- Jang, E., and Brutt-Griffler, J. (2019). Language as a bridge to higher education: a large-scale empirical study of heritage language proficiency on language minority students' academic success. *J. Multiling. Multicult. Dev.* 40, 322–337. doi: 10.1080/01434632.2018.1518451

- Jensen, E. (2009). How Poverty Affects Behavior and Academic Performance. Teaching with Poverty in Mind. Association for Supervision and Curriculum Development.
- Kissau, S., Adams, M. J., and Algozzine, B. (2015). Middle school foreign language instruction: a missed opportunity? *Foreign Lang. Ann.* 48, 284–303. doi: 10.1111/flan.12133
- LeBuffe, P. A., and Naglieri, J. A. (1999). The Devereux early childhood assessment (DECA): a measure of within-child protective factors in preschool children. *NHSA Dialog* 3, 75–80. doi: 10.1207/s19309325nhsa0301_10
- Leeman, J. (2005). Engaging critical pedagogy: Spanish for native speakers. *Foreign Lang. Ann.* 38, 35–45. doi: 10.1111/j.1944-9720.2005.tb02451.x
- Leeman, J., Rabin, L., and Román-Mendoza, E. (2011). Identity and activism in heritage language education. *Mod. Lang. J.* 95, 481–495. doi: 10.1111/j.1540-4781.2011.01237.x
- Leeman, J., and Serafini, E. J. (2016). “Sociolinguistics for heritage language educators and students: A model for critical translanguing competence” in *Innovative Strategies for Heritage Language Teaching*. eds. M. Fairclough and S. M. Beaudrie (Georgetown University Press), 56–79.
- Leeman, J. (2018). “Critical language awareness and Spanish as a heritage language: challenging the linguistic subordination of US Latinxs” in *The Routledge Handbook of Spanish as a Heritage Language*. ed. K. Potowski (New York: Routledge), 345–358.
- Leeman, J., and Fuller, J. M. (2021). Hablar Español en Estados Unidos: La Sociopolítica del Lenguaje. Multilingual Matters.
- Leeman, J., and Serafini, E. J. (2021). “It’s not fair”: discourses of deficit, equity, and effort in mixed Heritage and Second Language Spanish classes. *J. Lang Identity Educ.* 20, 425–439. doi: 10.1080/15348458.2020.1777866
- Looney, D., and Lusin, N. (2019). Enrollments in Languages other than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016. Modern Language Association. Available at: <https://www.mla.org/content/download/83540/2197676/2016-Enrollments-Short-Report.pdf> (Accessed July 15, 2022).
- Marian, V., Shook, A., and Schroeder, S. (2013). Bilingual two-way immersion programs benefit academic achievement. *Biling. Res. J.* 36, 167–186. doi: 10.1080/15235882.2013.818075
- Met, M., and Brandt, A. M. (2017). “Foreign language learning in K-12 classrooms in the USA” in *Second and Foreign Language Education*. eds. N. Deussen-Scholl and S. May. 3rd Ed. ed (Springer International Publishing), 357–370.
- Miami-Dade County Public Schools (n.d.). Required Courses of Study: Graduation Requirements. Available at: <http://mkhs.dadeschools.net/Curriculum/Files/Course/Req.htm> (Accessed July 15, 2022).
- Mora, M. T., Villa, D. J., and Dávila, A. (2005). Language maintenance among the children of immigrants: a comparison of border states with other regions of the U.S. *Southwest J. Linguist.* 24, 127–144.
- Nagano, T., Ketcham, E., and Funk, A. (2019). Why do heritage language speakers opt out of their own heritage language? A survey-based study of heritage language learners at community colleges. *Heritage Lang. J.* 16, 318–339. doi: 10.46538/hlj.16.3.3
- National Center for Education Statistics. (2015). Enrollment and Percentage Distribution of Enrollment in Public Elementary and Secondary Schools, by Race/Ethnicity and Region: Selected Years, Fall 1995 through Fall 2025. Digest of Education Statistics. Available at: https://nces.ed.gov/ipeds/data/digest/d15/tables/dt15_203.50.asp (Accessed July 15, 2022).
- National Center for Education Statistics. (2018). Digest of Education Statistics. Available at: https://nces.ed.gov/ipeds/data/digest/d17/tables/dt17_204.20.asp (Accessed July 15, 2022).
- Nehring, A.D., Nehring, E.F., Bruni, J.R., and Randolph, P.L. (1992). *Learning Accomplishment Profile—Diagnostic Standardized Assessment*. Kaplan Press, Lewisville, NC.
- Nguyen, M. V. H. (2020). Early Bilingualism and Foreign Language Learning in Secondary School (Master Thesis), George Mason University. Mason Archival Repository Service. Available at: <http://hdl.handle.net/1920/11932> (Accessed July 15, 2022).
- Nguyen, M. V. H., and Winsler, A. (2021). Early bilingualism predicts enhanced later foreign language learning in secondary school. *Dev. Psychol.* 57, 1926–1942. doi: 10.1037/dev0001248
- Padilla, A. M., Chen, X., Swanson, E., Peterson, M., and Peruzzaro, A. (2022). Longitudinal study of Spanish dual language immersion graduates: Secondary school academic and language achievement. *Foreign Language Annals* 55, 408–434.
- Parra, M. L. (2020). Working with diversity in the Spanish heritage language classroom: a critical perspective. *Porta Ling.* 34, 1–24. doi: 10.30827/portalin.v0i34.16730
- Peyton, D., Ranard, A., and McGinnis, S. (2001). “Charting a new course: Heritage language education in the United States”, in *Heritage Languages in America: Preserving a National Resource. Language in Education: Theory and Practice*. eds. J. K. Peyton, D. A. Ranard and S. McGinnis (Washington, DC: Center for Applied Linguistics), 3–27.
- Potowski, K. (2002). Experiences of Spanish heritage speakers in university foreign language courses and implications for teacher training. *ADFL Bull.* 33, 35–42. doi: 10.1632/adfl.33.3.35
- Prada, J., and Pascual, D. y Cabo, D. (2021). Towards an understanding of the relationship between heritage language programs and Latinx student retention and graduation: an exploratory study. In M. Bowles (Ed.), *Outcomes of University Spanish Heritage Language Instruction*. Georgetown University Press.
- Pufahl, I., and Rhodes, N. (2011). Foreign language instruction in U.S. schools: results of a national survey of elementary and secondary schools. *Foreign Lang. Ann.* 44, 258–288. doi: 10.1111/j.1944-9720.2011.01130.x
- Redford, J. (2018). English Language Program Participation among Students in the Kindergarten Class of 2010–11: Spring 2011 to Spring 2012. National Center for Education Statistics. Available at: <https://nces.ed.gov/pubs2018/2018086.pdf> (Accessed July 15, 2022).
- Ricciardi, C., Hartman, S., Manfra, L., Dinehart, L., Bleiker, C., and Winsler, A. (2021). School readiness skills at age 4 predict academic achievement through grade 5. *Early Child. Res. Q.* 57, 110–120. doi: 10.1016/j.ecresq.2021.05.006
- Rossee, Y. (2012). Lavaan: an R package for structural equation modeling. *J. Stat. Softw.* 48, 1–36. doi: 10.18637/jss.v048.i02
- Ruiz, R. (1984). Orientations in language planning. *NABE J.* 8, 15–34. doi: 10.1080/08855072.1984.10668464
- Rumbaut, R. G. (2009). “Pigments of our imagination: On the racialization and racial identities of ‘Hispanics’ and ‘Latinos’ in *How the US racializes Latinos: White hegemony and its consequences*. (Boulder, CO: Paradigm), 15–36.
- Serafini, E. J., Rozell, N., and Winsler, A. (2020). Academic and English language outcomes for DLLs as a function of school bilingual education model: The role of two-way immersion and home language support. *Int. J. Bilingual Educ. Bilingualism*, 1–19. doi: 10.1080/13670050.2019.1707477
- Serafini, E. J. (2021). “Assessing students through a critical language awareness framework” in *Heritage Language Teaching: Critical Language Awareness Perspectives for Research and Pedagogy*. eds. S. Loza and S. M. Beaudrie (New York: Routledge), 80–97.
- Serafini, E. J., Rozell, N., and Winsler, A. W. (2022). Academic and English language outcomes for DLLs in Miami as a function of school bilingual education model: the role of two-way immersion and home language support. *Int. J. Biling. Educ. Biling.* 25, 552–570. doi: 10.1080/13670050.2019.1707477
- Silva-Corvalán, C. (1994). *Language Contact and Change: Spanish in Los Angeles*. Oxford: Oxford University Press.
- Sparks, R. L. (2016). Myths about foreign language learning and learning disabilities. *Foreign Lang. Ann.* 49, 252–270. doi: 10.1111/flan.12196
- Sparks, R. L., Ganschow, L., and Javorsky, J. (1993). Perceptions of low and high risk students and students with learning disabilities about high school foreign language courses. *Foreign Lang. Ann.* 26, 491–510. doi: 10.1111/j.1944-9720.1993.tb01183.x
- Steele, J., Slater, R., Zamarró, G., Miller, T., Li, J., Burkhauser, S., et al. (2017). Effects of dual-language immersion programs on student achievement: evidence from lottery data. *Am. Educ. Res. J.* 54, 282S–306S. doi: 10.3102/0002831216634463
- Stewart-Brown, S., and Edmunds, L. (2003). Assessing emotional and social competence in preschool and primary school settings: a review of instruments. *Perspect. Educ.* 21, 17–40.
- Stewart-Strobel, J., and Chen, H. (2003). Motivations and attitudes affecting high school students’ choice of foreign language. *Adolescence* 38, 161–170.
- Thomas, W. P., and Collier, V. P. (2002). A National Study of School Effectiveness for Language Minority Students’ Long-Term Academic Achievement. Center for Research on Education, Diversity, and Excellence – University of California at Santa Cruz.
- Thompson, A. S. (2017). Language learning motivation in the United States: an examination of language choice and multilingualism. *Mod. Lang. J.* 101, 483–500. doi: 10.1111/modl.12409
- U.S. Census Bureau. (2020). 2020 American Community Survey 5-Year Estimates. Available at: <https://www.census.gov/quickfacts/miamidadecountyflorida> (Accessed July 15, 2022).

- Umansky, I. M., and Reardon, S. F. (2014). Reclassification patterns among Latino English learner students in bilingual, dual immersion, and English immersion classrooms. *Am. Educ. Res. J.* 51, 879–912. doi: 10.3102/0002831214545110
- Valdés, G. (1981). “Pedagogical implications of teaching Spanish to the Spanish-speaking in the United States” in *Teaching Spanish to the Hispanic Bilingual: Issues, Aims, and Methods*. eds. G. Valdés, A. G. Lozano and R. García-Moya (New York: Teachers College Press), 3–20.
- Valdés, G. (2001). “Heritage language students: profiles and possibilities” in *Heritage Languages in America: Preserving a National Resource*. eds. J. K. Peyton, D. A. Ranard and S. McGinnis (Washington, DC: Center for Applied Linguistics & Delta Systems), 37–80.
- Valdez, V. E., Freire, J. A., and Delavan, M. G. (2016). The gentrification of dual language education. *Urban Rev.* 48, 601–627. doi: 10.1007/s11256-016-0370-0
- Veltman, C. (1983). *Language shift in the United States* De Gruyter Mouton.
- Villa, D. J., and Rivera-Mills, S. (2009). An integrated multi-generational model for language maintenance and shift: the case of Spanish in the southwest. *Span. Context* 6, 26–42. doi: 10.1075/sic.6.1.03vil
- Wight, M. C. S. (2015). Students with learning disabilities in the foreign language learning environment and the practice of exemption. *Foreign Lang. Ann.* 48, 39–55. doi: 10.1111/flan.12122
- Wiley, T. G. (2000). “Continuity and change in the function of language ideologies in the United States” in *Ideology, Politics and Language Policies: Focus on English*. ed. T. Ricento (Philadelphia: John Benjamins), 67–85.
- Wiley, T. G., and García, O. (2016). Language policy and planning in language education: legacies, consequences, and possibilities. *Mod. Lang. J.* 100, 48–63. doi: 10.1111/modl.12303
- Winsler, A., Hutchison, L., De Feyter, J. J., Manfra, L., Bleiker, C., Hartman, S., et al. (2012). Child, family, and childcare predictors of delayed school entry and kindergarten retention among linguistically and ethnically-diverse children. *Dev. Psychol.* 48, 1299–1314. doi: 10.1037/a0026985
- Winsler, A., Kim, Y. K., and Richard, E. R. (2014). Socio-emotional skills, behavior problems, and Spanish competence predict the acquisition of English among English language learners in poverty. *Dev. Psychol.* 50, 2242–2254. doi: 10.1037/a0037161
- Winsler, A., Tran, H., Hartman, S. C., Madigan, A. L., Manfra, L., and Bleiker, C. (2008). School readiness gains made by ethnically diverse children in poverty attending center-based childcare and public school pre-kindergarten programs. *Early Child. Res. Q.* 23, 314–329. doi: 10.1016/j.ecresq.2008.02.003



OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway,
Norway

REVIEWED BY

Ethan Kutlu,
The University of Iowa,
United States
Mike Putnam,
The Pennsylvania State University (PSU),
United States

*CORRESPONDENCE

Begoña Arechabaleta Regulez
✉ begoar@uchicago.edu

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 23 November 2022

ACCEPTED 15 February 2023

PUBLISHED 20 April 2023

CITATION

Arechabaleta Regulez B and Montrul S (2023)
Production, acceptability, and online
comprehension of Spanish differential object
marking by heritage speakers and L2 learners.
Front. Psychol. 14:1106613.
doi: 10.3389/fpsyg.2023.1106613

COPYRIGHT

© 2023 Arechabaleta Regulez and Montrul.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Production, acceptability, and online comprehension of Spanish differential object marking by heritage speakers and L2 learners

Begoña Arechabaleta Regulez^{1*} and Silvina Montrul²

¹Department of Romance Languages and Literatures, University of Chicago, Chicago, IL, United States,

²Department of Spanish and Portuguese, University of Illinois at Urbana-Champaign, Urbana, IL, United States

We analyzed the production, acceptability and online comprehension of Spanish differential object marking (DOM) by two groups of bilingual speakers living in the U.S.: heritage speakers and L2 learners. DOM is the overt marking of direct objects that are higher on the animacy and referentiality scales, such as animate and specific objects in Spanish, marked by the preposition *a* (*Juan ve a María* 'Juan sees DOM María'). Previous studies have reported variability and high omission rates of obligatory DOM in bilingual situations where Spanish is in contact with a non-DOM language. Our study combined different methodologies to tap knowledge of DOM in the two groups. The results showed that heritage speakers and L2 learners (1) exhibited variability with DOM in production (in two oral tasks), comprehension (in an acceptability judgement task), and processing (in an eye-tracking reading task); (2) can integrate DOM into their production, judgments and processing, but they do so inconsistently, and (3) type of task and type of sentence each have an effect on speakers' use of DOM.

KEYWORDS

differential object marking, Spanish, variation, L2 speaker, heritage speaker, production, acceptability, processing

Introduction

Inflectional morphology is an area of significant variability in some bilingual grammars. It is still not known whether this variability is due to problems at the level of linguistic representations in the weaker, or non-dominant language or whether it is access to linguistic representations for comprehension, production and processing that is at the root of such variability. Both second language (L2) learners of Spanish and heritage speakers of Spanish have been shown to have difficulty with differential object marking (DOM), the overt morphological marking of animate, specific direct objects with the preposition "a" (Farley and McCollam, 2004; McCollam Wiebe, 2004; Montrul, 2004, 2010; Guijarro-Fuentes and Marinis, 2007; Bowles and Montrul, 2008, 2009; Guijarro-Fuentes, 2012; Montrul and Sánchez-Walker, 2013; Arechabaleta-Regulez, 2014). These studies have found high rates of omission of DOM in bilingual situations, where Spanish is in contact with a non-DOM language. In such situations, speakers omit DOM with animate specific objects, as in *Caperucita Roja visitó la abuelita* 'Little Red Riding Hood visited \emptyset her grandmother' (Montrul and Sánchez-Walker, 2013). Such omission of DOM in obligatory contexts has been reported in U.S. Spanish in contact with English (Montrul, 2004; Montrul and Sánchez-Walker, 2013) and in Peru Spanish in contact with Quechua (Sánchez,

2003). Although it is possible that DOM omission in these cases may be related to the fact that the other language does not exhibit DOM, DOM omission has also been reported in some monolingual contexts, as in Dominican (Lunn, 2002; Bullock and Toribio, 2004) and Cuban Spanish (Alfaraz, 2011). So, the nature of this variability is still begging for an explanation.

Spanish differential object marking

Spanish is similar to many other languages including Romanian, Hindi or Turkish in that overt case-marking happens to mark differentially some but not all objects by prepositions or postpositions. This phenomenon is known as Differential Object Marking (DOM). The object that is marked is semantically prominent and is distinguished from subjects by overt marking (Aissen, 2003). In Spanish, animate and specific (definite) objects are marked with DOM. For example, sentence (1) shows that because the direct object is [+animate] and [+specific] (definite)¹, DOM is required. When the direct object is [+animate] and [–specific], DOM is not required (2) and DOM can either be used or omitted. However, when the direct object is [–animate] and [+specific] (3) or [–animate] and [–specific] (4) DOM is not used.

- (1) *Mario vio a la doctora* ‘Mario saw the- DOM doctor [+animate] and [+specific]
- (2) *Mario vio (a) una doctora* ‘Mario saw a (DOM) doctor [+animate] and [–specific]
- (3) *Mario vio el carro* ‘Mario saw the car’ [–animate] and [+specific]
- (4) *Mario vio un carro* ‘Mario saw a car’ [–animate] and [–specific]

Even though animate objects are typically marked and inanimate objects are not, there exists some variation in the use of DOM in both monolingual and bilingual contexts. For example, several Spanish varieties in Latin America appear to show a slight tendency to overextend DOM to inanimate objects. A sentence like (5) in Rioplatense Spanish or (6) in Mexican Spanish are acceptable for some speakers in those varieties, while the same sentences are ungrammatical in other varieties, such as Peninsular Spanish. Moreover, in other Spanish varieties, the opposite development has been observed: DOM retraction. DOM retraction refers to the omission of DOM in contexts where DOM is usually used. The omission of DOM with animate and specific (definite) objects has been observed in some monolingual contexts (Lunn, 2002) as well as in some bilingual contexts (Montrul and Sánchez-Walker, 2013). This study focusses on this DOM omission by bilingual speakers (heritage speakers and L2 learners) living in the US.

- (5) *Chocó al coche* (Sánchez and Zdrojewski, 2013 Rioplatense Spanish)

¹ The notions of definiteness and specificity are both discourse related. The notion of definiteness refers to the state of knowledge shared between the speaker and hearer (or writer and reader), while the notion of specificity refers to the state of knowledge known to the speaker (writer) only (Ionin et al., 2004).

‘He hit the car’.

- (6) *Cosecharon al maíz* (von Heusinger and Kaiser, 2005 Mexican Spanish)

‘They harvested-DOM the corn’.

Heritage speakers and L2 learners of Spanish

Most Spanish language classrooms in the U.S. consist of both English-speaking students learning Spanish as an L2 (L2 learners of Spanish) as well as students who were raised hearing Spanish spoken at home (heritage speakers of Spanish). Heritage speakers are typically simultaneous and early successive bilinguals who are exposed to a minority language at home since birth and to a majority language in the community since birth or in childhood (Valdés, 2001; Montrul, 2004, 2016). As adults, heritage speakers tend to be dominant in the majority language and weaker in their heritage language, as assessed by both self-reports (Montrul, 2022), independent measure of proficiency (Montrul, 2016), and linguistic tasks (Montrul and Ionin, 2010). L2 learners, on the other hand, are usually sequential bilinguals who grow up exposed to the majority language and only begin acquiring an L2 during or after puberty. L2 learners and heritage speakers’ experience with the weaker language is different. Table 1 summarizes the main features of the two types of acquisition (heritage language and L2) from which differences and similarities between heritage speakers and L2 learners can be drawn. While heritage speakers are exposed to Spanish during childhood, typically through an aural medium and in a naturalistic context (home), L2 learners are exposed to Spanish during or after puberty in a formal context (classroom) with a strong emphasis on reading and writing activities as well as structured grammar explanations, activities and feedback. Therefore, L2 learners, but not heritage speakers, tend to be very literate in their L2 and have highly developed metalinguistic knowledge of the target language. Metalinguistic knowledge is typically defined as the explicit and declarative knowledge the speakers have about the language. Heritage speakers, on the other hand, usually have less developed literacy skills and less metalinguistic knowledge of their heritage language than their majority language. Motivation to learn and maintain the language is another important difference between these two types of speakers. The main motivations for

TABLE 1 Characteristics of heritage language and L2 acquisition.

Time	Early exposure	Late exposure (during or after puberty)
Setting	Naturalistic (home)	Instructed (classroom)/ study-abroad
Mode	Aural Input	Aural and Written Input
Errors	Developmental and transfer errors	Developmental and transfer errors
Fossilization	Typical	Typical
Motivation	Yes	Yes
Outcome	Variable	Variable

heritage speakers to regain their language are to maintain their heritage language, strengthen family connections, and reinforce their identity (Reynolds et al., 2009). In contrast, L2 learners usually seek to improve their grammatical skills (Mikulski, 2006) and be able to communicate with people who can speak the target language (Reynolds et al., 2009). They also seek professional opportunities (Beaudrie and Ducar, 2005; Alarcon, 2010; Carreira and Kagan, 2011). Finally, heritage speakers may not want to use their heritage language due to the social stigma attached to their Spanish which debilitates their view of themselves as Spanish speakers (Kutlu and Kircher, 2021).

Despite these differences in language experience, heritage speakers and L2 learners also share many similarities. For example, when using the target language, both types of speakers tend to show morphological variability due to the influence of the majority language. Previous research comparing heritage speakers and L2 learners has suggested that age of acquisition alone cannot explain the main differences between the two groups (Au et al., 2002; Benmamoun et al., 2010). According to the notion “earlier is better,” heritage speakers should always outperform L2 learners because they are exposed to the language at an earlier age. However, this is simply not the case (Au et al., 2002; Montrul et al., 2008), because language experience shapes their knowledge as well and this is manifested in different tasks and the modality in which the language is tested. While heritage speakers usually outperform L2 learners in oral tasks of morphosyntax, results vary and often depend on the type of task. Heritage speakers tend to have an advantage with tasks that tap language implicitly and minimize metalinguistic knowledge (Bowles, 2011); L2 learners, by contrast have an advantage with tasks that focus on explicit knowledge of the language and are more metalinguistic. The fact that heritage speakers and L2 learners perform differently depending on the degree of explicitness or implicitness of the task suggests that performance is heavily influenced by language experience (Bowles, 2011). That is why in order to understand the nature of their linguistic knowledge, it is important to use tasks that tap into participants’ explicit and implicit knowledge.

It has been common to test implicit knowledge *via* oral tasks because language production unfolds over time. However, analyzing participants’ free production in oral tasks is often insufficient to measure implicit knowledge accurately. For instance, participants still have opportunities to resort to their explicit knowledge in oral tasks, especially when the task is untimed and participants can monitor and repair their performance (Jiang, 2004).

The use of online processing techniques are essential to offer evidence of implicit knowledge. Unlike offline tasks, online tasks tap into individuals’ implicit knowledge by analyzing the actual processing mechanisms that are being used during comprehension or production in real time (Field, 2004). Thus, these online tasks measure implicit real-time behavior/reactions as opposed to measuring potential ‘learned’ knowledge of heritage language or the L2. Properly examining access to implicit knowledge is important because, according to certain language processing theories, implicit knowledge can only be accessed if one is exposed to the language early in life [e.g., The Declarative/Procedural Model (Ullman, 2004); The Shallow Structure Hypothesis (Clahsen and Felser, 2006)]. Moreover, access to implicit knowledge is thought to be central to acquiring native-like competence in both L1 and L2 acquisition (Krashen, 1982). If heritage speakers do not show the same advantages over L2 learners when tested with online processing tasks

as they do when tested with oral tasks, this would suggest that early exposure in a naturalistic context is not enough to achieve a high level of implicit knowledge in that language. If this is the case, limited use and exposure to the language in late childhood and adolescence may be affecting their competence.

Omission of DOM by L2 learners and heritage speakers

Previous studies have consistently shown that both heritage speakers and L2 learners omit the *a*-marker with animate objects (Farley and McCollam, 2004; McCollam Wiebe, 2004; Montrul, 2004; Guijarro-Fuentes and Marinis, 2007; Bowles and Montrul, 2008, 2009; Montrul, 2010; Guijarro-Fuentes, 2012; Montrul and Sánchez-Walker, 2013; Arechabaleta-Regulez, 2014). For example, Montrul (2010) compared heritage speakers and L2 learners on the acquisition of DOM. Montrul investigated whether age of onset of acquisition and/or influence from their dominant language, English, was preventing heritage speakers and L2 learners from fully acquiring DOM. Heritage speakers ($n=67$) and L2 learners ($n=72$) were divided into three groups depending on their Spanish proficiency: advanced (Heritage Speakers = 32, L2 = 25), intermediate (Heritage Speakers = 26, L2 = 25) and low (HS = 13, L2 = 22). Heritage speakers and L2 learners were compared to a group of monolingually-raised native speakers from different Spanish-speaking countries. Participants completed two main tasks: an oral narrative task (Montrul, 2004) and an acceptability judgment task. Results for the oral narrative task showed that heritage speakers and L2 learners at all proficiency levels omitted DOM with animate objects, while the native speakers did not. However, the L2 learners produced almost twice the amount of omissions (46.9%) as the heritage speakers (26.5%). Moreover, results also showed that advanced heritage speakers did not differ significantly from the native speaker control group, which suggests that proficiency is an important factor when comparing heritage speakers to monolingually-raised native speakers. As for the AJT, results showed that, overall, heritage speakers and L2 learners accepted sentences with DOM omission and animate objects, but the control group did not. In this task, L2 learners behaved more like the native speakers, as heritage speakers, regardless of proficiency, accepted sentences with DOM omission and animate objects significantly more often. Therefore, the two groups differed significantly from the native speakers, but the L2 learners outperformed the heritage speakers, especially at lowest levels of proficiency. Montrul (2010) concluded from the results of the two tasks that DOM is subject to incomplete acquisition or attrition for both heritage speakers (Montrul and Bowles, 2009, 2010) and L2 learners (McCollam Wiebe, 2004; Bowles and Montrul, 2009). Montrul also noted the importance of using different tasks when comparing heritage speakers and L2 learners. In the oral task, the heritage speakers showed an advantage over the L2 learners, but in the written task, the L2 learners showed an advantage over the heritage speakers. Finally, Montrul suggested that DOM omission can easily be attributed to transfer from English. Spanish, unlike English, is a language with rich inflection, and rich agreement co-occurs with the possibility of non-canonical word order. In those cases, Spanish relies on case marking to indicate thematic roles. Thus, DOM is crucial to understand *who* is doing *what*, especially when the

object is animate. English word order, on the other hand, is relatively fixed. Thus, word order usually conditions thematic interpretations in English.

In fact, the omission of case marking is heavily influenced by the word order flexibility of the language. To test the correlation between word order and case marking in a language, Fedzechkina et al. (2015) exposed learners to two miniature artificial languages. Both languages contained case marking, but while one language had flexible word order, the other had fixed word order. Results showed that learners who were exposed to the language with flexible word order used case marking more often than the learners who were exposed to the language with fixed word order. The learners made changes to the artificial languages that are compatible with language universals; that is, grammatical patterns that are prone to happen crosslinguistically. In cases where speakers have grammatical cues that are highly informative (e.g., word order), other cues become redundant and are thus omitted (e.g., case marking). Indeed, Lunn (2002) has suggested that DOM is disappearing from Dominican Spanish because of another innovation occurring in this dialect: Dominican native speakers appear to use a stricter SVO word order, and thus direct objects are expected to appear after the verb. Therefore, using DOM to disambiguate thematic roles is becoming uninformative. The tradeoff between word order and case marking as a cue to thematic roles is also discussed in the Unified Competition Model (UCM; MacWhinney, 2005; see also the Competition Model of Bates and MacWhinney, 1987).

The omission of DOM with animate objects that has been observed in both heritage speakers and L2 learners is compatible with these language universals. Perhaps, DOM retraction may be a consequence of a change in the word order possibilities of Spanish in contact with English. In other words, the Spanish of the United States may be acquiring a more fixed SVO word order similar to Dominican Spanish. Still, a major question remains: Is DOM disappearing across the board in these varieties or only in contexts where case marking may be less informative (sentences with canonical word order)?

The aim of this study is to investigate whether heritage speakers and L2 learners, who often omit DOM in production and grammaticality judgments, do not process DOM during sentence processing. Unlike previous studies that have mostly focused on SVO sentences, the present study examines whether omission of DOM occurs with canonical and/or non-canonical word order sentences. The majority of studies on heritage speakers and L2 learners have not examined the interaction between word order and DOM. However, heritage speakers and L2 learners may show omission of DOM only in contexts where case marking is less informative, as in SVO sentences. If tested in contexts where DOM is critical for comprehension (sentences with non-canonical word order), heritage speakers and L2 learners may not show the same extent of DOM omission. Previous research on the processing of DOM by heritage speakers and L2 learners of Spanish suggests that DOM omission is reflected in speakers' processing mechanisms. When exposed to ungrammatical sentences with unmarked animate objects, neither heritage speakers nor L2 learners show any sensitivity to ungrammaticality (Jegerski, 2015, 2018).

Arechabaleta-Regulez (2016), investigated heritage speakers' processing of DOM in sentences with canonical (SVO) and non-canonical (VSO) word order. Results of an eye-tracking during

reading task demonstrated that heritage speakers were more sensitive to DOM omission with non-canonical VSO word order than with canonical SVO sentences. This suggests that heritage speakers rely on word order and ignore case marking with canonical word order sentences, possibly due to transfer from their dominant language (English). However, with non-canonical word order sentences, heritage speakers appeared to utilize DOM as an informative cue to word order. Therefore, omission of DOM was evident in their processing of canonical word order sentences but not in their processing of non-canonical word order sentences. Building on Arechabaleta-Regulez (2016), this study examines whether L2 learners behave like heritage speakers in their processing of DOM. We predicted that their different language learning experiences regarding timing (before vs. after the critical period) and context of acquisition (naturalistic vs. formal environment), may affect their processing. However, we also test production and judgments of DOM because it is well known that heritage speakers and L2 learners tend to show DOM omission (e.g., Montrul and Sánchez-Walker, 2013). In this study, the tasks provide comprehensive information related to participants' production, acceptance and online comprehension of DOM. The importance of analyzing bilinguals' productive and receptive knowledge is to understand potential dissociations and asymmetry between speakers' production, acceptability and processing.

Participants completed the reading task with eye-tracking first, followed by the oral tasks, first the narrative task and then the elicitation task, and finally the AJT. After completing these tasks, participants also completed the background questionnaire and a written Spanish Proficiency test. Proficiency scores were included as covariates to assess the extent to which proficiency affected participants' production, acceptability and online comprehension of DOM. The following sections describe each task in greater detail, including the corresponding research questions, hypotheses and results. Rather than following the exact order in which participants completed the tasks, the discussions are arranged so that the most innovative findings are discussed last.

Methodology

Participants

Thirty-five heritage speakers and 42 L2 learners were recruited. All participants were between the ages of 18 and 25 (average age 21.3). In order to participate in the study, heritage speakers were required to: (1) have been born in the U.S. (they were all second generation); (2) have been exposed primarily to Spanish in early childhood or to both Spanish and English and (3) be of Mexican origin to the greatest extent possible (either one parent or both were from Mexico). L2 learners were required to: (1) have been born in the U.S.; (2) have been exposed to Spanish in a formal context but not earlier than the age of 10 (L2 speakers reported that they had been exposed to various Spanish dialects depending on their teachers) and (3) not speak any other second language besides Spanish. We were primarily interested in testing heritage speakers and L2 learners with an intermediate to high proficiency in Spanish. Heritage speakers and L2 learners completed a background questionnaire to determine whether they met all of these

requirements and an adapted version of the DELE (Diploma of Spanish as a Foreign Language) proficiency test as an independent measure of proficiency in Spanish (see Table 2).

When comparing the results obtained in the DELE test, there was a significant effect ($\beta = -12.66$, $SE = 1.46$, $p < 0.0001$) as heritage speakers scored significantly higher than the L2 learners. Moreover, as Figure 1 shows, the dispersion of the scores varied. While most of the heritage speakers scored above 35 points, most of the L2 learners scored between 20 and 30 points out of a maximum of 50 points. Before testing the participants' language processing, it is also important to test their production and judgments of DOM. Therefore, participants completed two oral tasks and an acceptability judgment task (AJT). No study has used all these methodologies to examine oral production, judgment and sentence processing during reading in the two groups. Analyzing bilinguals' productive and receptive knowledge is critical to understand potential dissociations and asymmetries between their production, acceptability and processing.

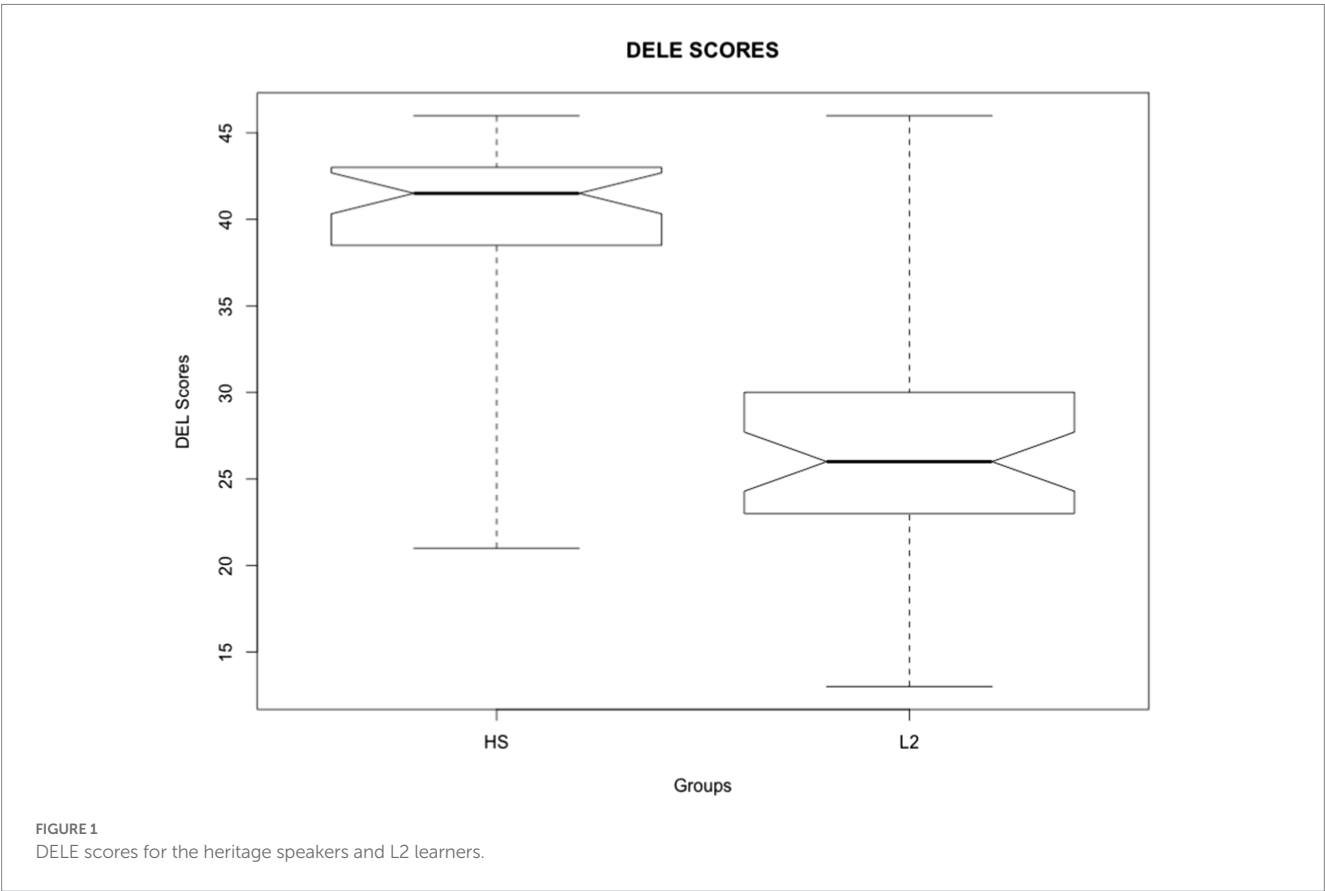
TABLE 2 Background questionnaire information.

Participants	N	Age	AoA of Spanish	AoA of English	DELE scores
Heritage Speakers	35	19.3 (18–22)	Birth	2.2(0–4)	39.76 (21–46)
L2 Learners	42	20.2(18–24)	12.2 (10–14)	Birth	26.62 (13–46)

Procedure

Participants arrived at the laboratory where they first read and signed a consent form. Then, they began the study by completing the reading task with eye-tracking, for which a portable eye-tracker (Eye Link SR Research, Ltd.; Ottawa, Canada) with remote desktop camera sampling at 500 Hz was used. The eye-tracker was used in a diagnostic manner because it recorded and analyzed participants' eye position while reading sentences. Subjects were seated 50 cm from the monitor with their chin/head rest. It is important to use a chin/head rest to increase accuracy of measurement (Carter and Luke, 2020). Sentences were presented in 18-point Courier font, left-aligned on the display. Before the task began, a calibration procedure was carried out to accurately track participants' eye-movements. During this initial process, participants were instructed to fix their gaze on a set of nine fixation points (black dots) displayed on the screen at known locations. While they were doing this, the positions of their eyes were recorded. If there were no errors when the calibration was performed, the computer then "validated" the information before subjects could begin the actual test. A calibration was accepted if average error was less than 1 degree of visual angle and calibration was as necessary during the experiment.

Next, participants completed a practice session, which consisted of 8 trials, following the same procedure as the actual study to familiarize participants with the eye-tracker and the response controller. The structure of each trial was as follows: first, a white screen with a black dot, the central fixation point, appeared in the left



middle of the screen. Participants were told to look at this point immediately prior to pressing a button on a controller, which prompted a sentence to appear on the screen. After reading the sentence, participants pressed the button again to continue to a comprehension question related to the sentence they had previously seen. Participants used one of two buttons to respond ‘yes’ or ‘no’ to the comprehension questions after each trial. After the practice session, participants were instructed to move their head as little as possible during the experiment to ensure accurate tracking of their eye movements. Participants were also informed that they would be allowed to take three breaks during the experiment. If participants decided to take a break, and thus, moved their chin, recalibration was performed again. The eye-tracker machine recorded all movements of each participant’s right eye between the appearance of the white screen with the black point, indicating the beginning of a new trial, and the disappearance of the sentence, when a participant pressed the button to proceed to the comprehension question. In total, this task lasted between 30 and 45 min.

After the reading task with eye-tracking, participants completed the oral task in two parts: first the narrative task, then the elicitation task. Participants were seated in front of a laptop computer and their answers were recorded by the same laptop for both portions. For the narrative task, participants were asked to narrate the story in Spanish based on the pictures with as many details as possible. They advanced through the presentation at their own pace while their narration was continuously recorded. This task did not take longer than 10 min. The participants then completed the elicited production task, which took less than 10 min.

After the two oral tasks, participants completed the acceptability judgment task (AJT) using the same laptop they used for the oral tasks. Before starting the AJT, participants were told to read the sentences as carefully and as quickly as possible and to rely on their first instinct. The sentences were presented visually, and participants had as much time as they wanted to read and judge the sentences. They were instructed to rate the sentences on a scale of 1 to 5 by pressing a button on the computer, with 1 indicating completely unacceptable and 5 totally acceptable. A rating of 3 represented ‘undecided’. Participants completed the task within 30 to 40 min. Finally, participants completed the background questionnaire, which took about 20 to 30 min. In total, it took participants between 1.5 to 2 h to complete all of the tasks. Thus, all participants completed the

most implicit tasks first (i.e., the reading task with eye-tracking) and the most explicit tasks last (i.e., the AJT).

The following sections describe each task in greater detail, including the corresponding research questions, hypotheses and results. Rather than following the exact order in which participants completed the tasks, the discussions are arranged so that the most innovative findings are discussed last.

Oral tasks: Narrative task and elicited production task

First, we asked to what extent heritage speakers and L2 learners omit DOM in obligatory contexts in oral production, and whether their performance depended on the implicit or metalinguistic nature of the task, as found in previous studies. Two oral tasks—an oral narrative task and an elicited production task—measured participants’ oral production of Spanish DOM. For the narrative task, participants narrated the children’s story ‘Little Red Riding Hood’ (from Montrul, 2004). Participants were provided with 14 colorful pictures of the story *via* a PowerPoint slideshow and were asked to narrate the story using the preterit tense while providing as much detail as possible based on the pictures. The pictures contained many animate and inanimate referents as objects. Because participants are usually more concerned with *what* to say (meaning of the story) rather than *how* to say it (grammar) when completing narrative tasks, this task provides semi-spontaneous data, perhaps comparable to what one can elicit with sociolinguistic interviews.

In the elicitation task, participants were presented with a picture with a verb and animate and inanimate NPs as subjects and objects on a computer screen and were asked to produce a sentence describing the picture using the verb and NPs given (see Figure 2). Participants were told to conjugate the verb in the preterite tense, so the presence or absence of DOM could be perceived. In total, participants were presented with 24 pictures: 12 with animate objects and 12 with inanimate objects. Another 12 pictures were included as fillers. The fillers prompted participants to use different constructions (e.g., sentences with *gustar*-type verbs). We believe that in this task participants have less freedom to produce spontaneous speech as they are given some of the words they need to use.



FIGURE 2

Sample of items used in the oral elicitation task: (A) shows the picture used for the verb *saludar* ‘to greet’; (B) shows the picture used for the verb *escuchar* ‘to listen’ [reproduced with permission from Arechabaleta-Regulez and Montrul (2021)].

Both heritage speakers and L2 learners were expected to show a significant rate of DOM omission in their production. However, following previous research on the production of DOM, overall, heritage speakers were expected to show less DOM omission than the L2 learners; especially because proficiency an important factor (Montrul, 2010). Participants with a higher proficiency were expected to show fewer ungrammatical unmarked animate objects. Overall, participants were not expected to extend DOM to inanimate objects. With respect to task effects, the L2 learners were expected to show more DOM omission in the narrative task than in the elicitation task. The elicitation task is more explicit, and thus participants may rely more on their explicit knowledge and use their metalinguistic knowledge while completing this task. As for the heritage speakers, they were expected to show the opposite pattern; namely, more omission of DOM in the elicitation task than in the narrative task. While L2 learners seem to perform better in explicit tasks that maximize metalinguistic knowledge, heritage speakers seem to perform better in implicit tasks that minimize metalinguistic knowledge.

Results: Oral tasks

Narrative oral task

Participants' answers were audio recorded and their answers transcribed and coded by a native speaker from Spain. All sentences containing object NPs were analyzed and the objects were coded for animacy and for DOM marking (present or absent). In situations where participants produced unexpected sentences, those sentences were coded as 'other' and were removed from the final statistical analyses. An example of a sentence coded as 'other' is when participants used the passive voice 'El alumno fue castigado' 'The student was punished' instead of the active sentence with DOM 'La profesora castigó al alumno' 'The teacher punished the student'. Results were analyzed with a bivariate logistic regression with the framework of *glm* (generalized linear model) using R (version 1.1.453 for Mac OS X, R Development Core Team, 2014), with participant and item as random effects and markedness ([+DOM] vs. [-DOM]), animacy of the object (animate vs. inanimate) and group (heritage speakers vs. L2 learners) as fixed effects. All fixed effects were coded as a binary variable using dummy coding (markedness: [+DOM] = 1, [-DOM] = 0; animacy of the object: animate object = 1, inanimate object = 0; group: heritage speakers = 1, L2 learners = 2). Each participant ended up with 4 percentage scores reflecting their use or omission of DOM with either animate or inanimate objects. Proficiency scores were included as covariates to assess the extent to which proficiency affected performance.

Table 3 shows that, as predicted, heritage speakers (8a) and L2 learners (8b) omitted DOM with animate objects; and that heritage speakers showed lower DOM omission rates than the L2 learners.

For heritage speakers, 80.40% of the animate objects were marked, while 19.60% were unmarked. However, for the L2 only 38.74% of the animate objects were marked and 61.26% were unmarked. Unlike native speakers of Mexican Spanish who have been shown to extend DOM to inanimate objects (Arechabaleta-Regulez and Montrul, 2021), these bilingual participants did not show much extension of DOM to inanimate objects. While heritage speakers did not produce any cases of inanimate objects with DOM extension, the L2 learners did so in 5 occasions, as in (7c).

(7)

- a. [Participant 302] ver su abuela
see his/her grandmother
'(She) see her grandmother'
- b. [Participant 254] comio la nina
ate the girl
'(the wolf) ate the girl'
- c. [Participant 237] mirando a las flores
staring DOM at the flowers
'(She/he) was staring at the flowers.'

The logistic regression revealed a significant effect of ANIMACY, as animate objects were marked with DOM significantly more than inanimate objects ($\beta = -4.24$, $SE = 0.0008$, $p < 0.0001$), and a significant GROUP effect ($\beta = -0.72$, $SE = 0.0008$, $p < 0.0001$), as heritage speakers used DOM significantly more often than L2 learners regardless of the animacy of the object. There was also a significant interaction between ANIMACY and GROUP ($\beta = 1.96$, $SE = 0.0008$, $p < 0.0001$). Tukey's multiple comparison test revealed that heritage speakers ($\beta = 7.43$, $SE = 1.11$, $p < 0.0001$) and L2 learners ($\beta = 3.05$, $SE = 0.5$, $p < 0.0001$) used DOM significantly more often with animate objects than with inanimate objects. However, when comparing the use of DOM with animate objects between the two groups of bilinguals, there was a significant effect ($\beta = 1.55$, $SE = 0.54$, $p = 0.02$) as heritage speakers used DOM significantly more often than the L2 learners. As for the use of DOM with inanimate objects, there was not a GROUP effect as the use of DOM was minimal for heritage speakers and L2 learners ($\beta = -2.82$, $SE = 1.22$, $p = 0.09$). Finally, there was a significant PROFICIENCY effect ($\beta = 0.093$, $SE = 0.0008$, $p < 0.0001$). As Figure 3 shows, participants with higher proficiency used DOM with animate objects more often than participants with lower proficiency.

However, proficiency seems to have a bigger effect on L2 learners than on heritage speakers. Interestingly, for the L2 learners, proficiency also had an effect on the extension of DOM to inanimate objects. It appears that L2 learners with a higher proficiency of Spanish used DOM more with both animate and inanimate objects. Participants may have acquired the rule that states that DOM is used with animate

TABLE 3 Use or omission of DOM with animate and inanimate objects (narrative).

	Heritage Speakers			L2 Learners		
	Total	Marked	Unmarked	Total	Marked	Unmarked
Animate	199 (100%)	160 (80.40%)	39 (19.60%)	222 (100%)	86 (38.74%)	136 (61.26%)
Inanimate	124 (100%)	0 (0%)	124 (100%)	102 (100%)	5 (4.90%)	97 (95.10%)

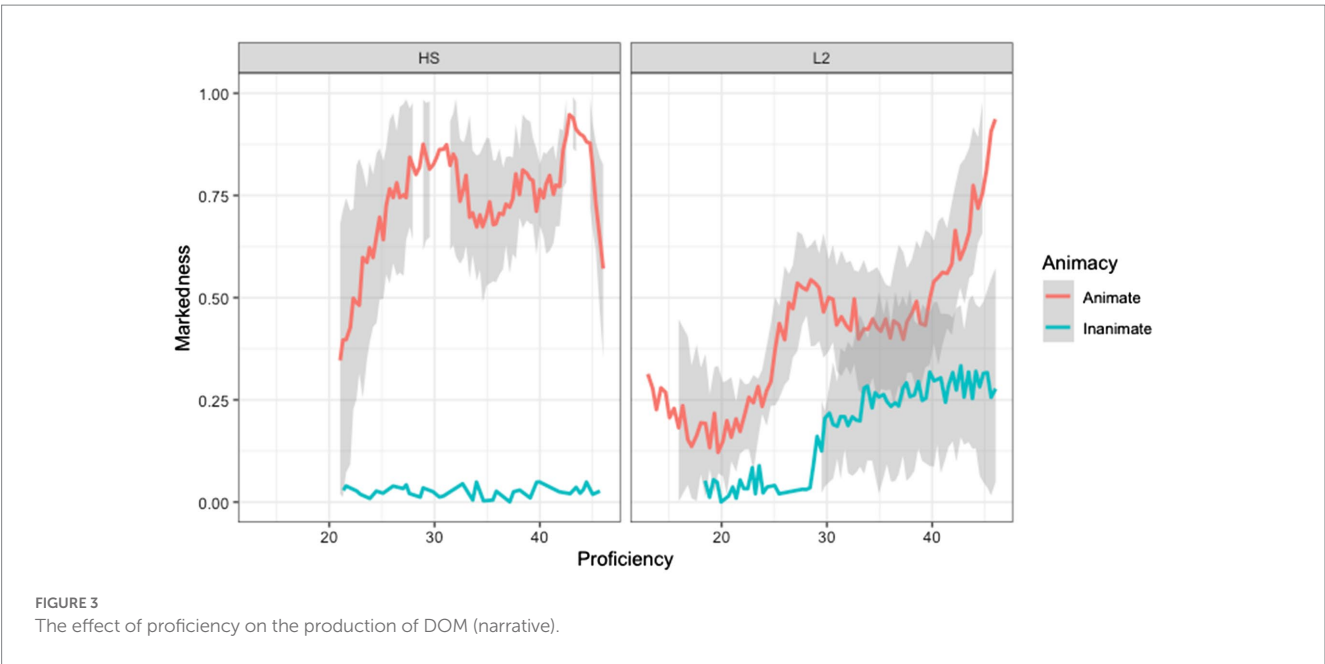


TABLE 4 Use or omission of DOM with animate and inanimate objects (elicitation task).

	Heritage Speakers			L2 Learners		
	Total	Marked	Unmarked	Total	Marked	Unmarked
Animate	412 (100%)	298 (72.33%)	114 (27.66%)	498 (100%)	205 (41.17%)	293 (58.83%)
Inanimate	411 (100%)	46 (11.19%)	365 (88.81%)	496 (100%)	65 (13.08%)	431 (86.92%)

objects, and they are now overextending this rule to inanimate objects. However, heritage speakers did not extend the use of DOM to inanimate objects.

Oral elicitation task

In total, 31 sentences were coded as ‘other’ and were removed from the statistical analyses. Heritage speakers (8a) and L2 learners (8b) omitted DOM with animate objects; however, heritage speakers again showed less DOM omission than expected: 27.45% of the animate objects were unmarked and 72.55% were marked. L2 learners produced 58.83% of the animate objects unmarked and 61.26% marked. Moreover, there were more cases of DOM extension to inanimate objects in this task by both heritage speakers, as in (8), and L2 learners, as in (5.2d) (see Table 4).

(8)

- a. [Participant 207] Cristina saludó los novios
Cristina said hi to the couple
‘Cristina said hi to the couple’
- b. [Participant 322] El ladrón atacó el presidente
the thief attacked the president
‘The thief attacked the president’
- c. [Participant 311] El viaje llevo al paraguas
the old mal brought DOM the umbrella
‘The old mal brought DOM the umbrella’
- d. [Participant 213] El hombre besó al trofeo

the man kissed the DOM trophy
‘The man kissed the trophy’.

The logistic regression revealed a significant effect of ANIMACY ($\beta = -2.48$, $SE = 1.06$, $p = 0.02$), because participants marked animate objects significantly more often than inanimate objects overall, and a significant ANIMACY and GROUP interaction ($\beta = 1.93$, $SE = 0.30$, $p < 0.0001$). Tukey’s multiple comparison test revealed that heritage speakers ($\beta = 4.20$, $SE = 0.42$, $p = 0.001$) and L2 learners ($\beta = 2.27$, $SE = 0.39$, $p < 0.001$) used DOM significantly more often with animate objects than with inanimate objects. However, heritage speakers and L2 learners did not significantly differ on either the use of DOM with animate objects ($\beta = 0.69$, $SE = 0.47$, $p = 0.45$) or on the use of DOM with inanimate objects ($\beta = -1.23$, $SE = 0.49$, $p = 0.06$). That is why in the logistic regression, GROUP did not turn out to be a significant effect ($\beta = -0.69$, $SE = 0.47$, $p = 0.14$). Finally, there was a PROFICIENCY effect ($\beta = 0.093$, $SE = 0.0008$, $p < 0.0001$), as participants with higher proficiency marked DOM with animate objects more often than participants with lower proficiency. Figure 4 shows that the production of DOM increase as participants’ proficiency increases. In the elicitation task, L2 learners also marked some inanimate objects, but it is not as correlated to proficiency as in the narrative task.

The reason for using two oral tasks was to analyze whether participants’ use of DOM would vary depending on whether they were completing a narrative task or an elicitation task. In order to analyze task effects, results for the animate objects and inanimate objects were analyzed individually with a bivariate logistic

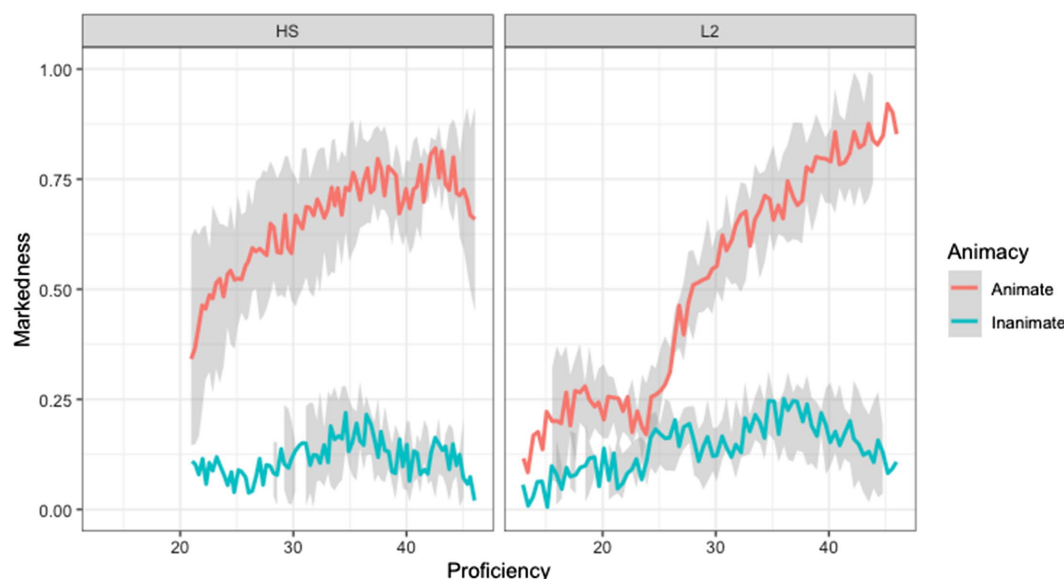


FIGURE 4
The effect of proficiency on the production of DOM (elicitation task).

TABLE 5 Sample sentences used in the AJT.

Direct object	[+DOM]	[-DOM]
Animate	El niño acusó al señor de las gafas azules.	*Diego acogió el estudiante de intercambio.
	'The kid accused the man with the blue glasses.'	'Diego welcomed the exchange student.'
Inanimate	El joven apreció al esfuerzo económico por parte de sus padres.	La actriz dibujó el carro de sus sueños
	'The young boy appreciated the economic effort that his parents made.'	'The actress drew her dream car.'

regression with the framework of glm in R with participant and item as random effects and markedness ([+DOM] vs. [-DOM]), task (narration vs. elicitation) and group (heritage speakers vs. L2 learners) as fixed effect. Results for the animate objects revealed a significant TASK effect ($\beta = 0.65$, $SE = 0.23$, $p = 0.006$) and a significant TASK*GROUP interaction ($\beta = -1.06$, $SE = 0.32$, $p = 0.0009$). Tukey's multiple comparison test revealed a significant difference between the use of DOM by the heritage speakers in the narrative task and in the elicitation task ($\beta = -0.65$, $SE = 0.23$, $p = 0.03$), as participants produced DOM with animate objects significantly more in the narrative task than in the elicitation task. However, for the L2 learners there was not a significant effect on the use of DOM between the two tasks. As for the inanimate objects, results revealed a significant TASK effect ($\beta = -1.68$, $SE = 0.60$, $p = 0.005$). Tukey's multiple comparison tests only revealed a significant effect when comparing the use of DOM with inanimate objects in the narrative task and in the elicitation task ($\beta = 1.68$, $SE = 0.60$, $p = 0.02$), as heritage speakers used DOM with inanimate objects significantly more often in the elicitation task than in the narrative task. For the L2 learners, there were not any significant comparisons.

Summary of results

As hypothesized, participants showed DOM omission in the narrative and in the elicitation task. Nevertheless, the L2 learners produced significantly more unmarked animate objects than the

heritage speakers. Moreover, the L2 learners also showed more extension of DOM to inanimate objects than the heritage speakers. Proficiency turned out to be a significant factor, especially for the L2 learners. Participants with a higher proficiency, used DOM significantly more than participants with a lower proficiency in Spanish. Proficiency also had an effect on the extension of DOM to inanimate objects for the L2 learners. L2 learners with a high proficiency in Spanish produced more marked inanimate objects.

Acceptability judgment task (AJT)

The aim of this task was to test participants' judgments of grammatical and ungrammatical sentences with DOM in both SVO and VOS sentences. Sentences varied by animacy of the object (animate vs. inanimate) and object marking ([+DOM] vs. [-DOM]) as shown in Table 5.

Based on previous studies (Guijarro-Fuentes and Marinis, 2007; Montrul and Sánchez-Walker, 2013), we predicted that the bilingual speakers would accept sentences with animate objects and DOM (*El niño acusó al señor de las gafas azules*) as well as sentences with unmarked inanimate objects (*La actriz dibujó el carro de sus sueños*), and would show more variability rejecting ungrammatical sentences with animate objects and DOM omission (**Diego acogió el estudiante de intercambio*).

As in previous studies (Montrul, 2010; Guijarro-Fuentes, 2012), proficiency was expected to play a role on L2 participants' rating as participants with a higher proficiency in Spanish were expected to show less acceptance of DOM omission. Finally, word order was also expected to play a role. Higher rejection of DOM omission with animate objects was expected in sentences with non-canonical word order, as DOM is more informative. With sentences with inanimate objects, participants were expected to reject ungrammatical sentences with DOM and to accept unmarked objects, which are grammatical (Jegerski, 2018). Finally, as this is a metalinguistic task, L2 learners were expected to reject ungrammatical DOM omission with animate objects and ungrammatical DOM extension to inanimate objects more than heritage speakers overall (Montrul, 2010).

Because the acceptability task used a rating scale, the results were analyzed using the *clmm* (cumulative link mixed model) function in the "ordinal" package (Christensen, 2015) using R (version 1.1.453 for Mac OS X, R Development Core Team, 2014). *Clmms* were performed on the ordinary-scaled data to model both participant- and item-variability (Agresti, 2002). The raw scores were entered as primary outcome measures (i.e., item ratings per participant and condition) into the statistical analyses. Markedness ([+DOM] vs. [-DOM]) and animacy of the object (animate vs. inanimate) were both fixed effects. Subject and item were included as random effects not standardized because *clmms* take inter-participant variation into consideration. *Clmms* were performed separately for each type of sentence (SVO vs. VSO), and the results obtained for each sentence type are discussed below. Proficiency scores were included as covariates to assess the extent to which proficiency of the participants affected their performance.

SVO sentences

Figure 5 shows that with animate objects, heritage speakers accepted more grammatical sentences with DOM ($M=4.57$, $SD=0.97$) than ungrammatical sentences with DOM omission ($M=3.19$, $SD=1.47$). However, heritage speakers seemed unsure about the rejection of sentences with unmarked animate objects. With inanimate objects, heritage speakers rejected more the use of DOM ($M=3.71$, $SD=1.41$) than the omission of DOM ($M=4.4$, $SD=1.01$). However, there was a lot of variation among heritage speakers' answers, especially with rejection of DOM omission with animate objects and the extension of DOM to inanimate objects. These patterns suggest that while some participants rejected unmarked animate objects and marked inanimate objects, others accepted them.

Figure 6 shows the results obtained by the L2 learners. Similar to the heritage speakers, the L2 learners accepted sentences with DOM and animate objects DOM ($M=4.07$, $SD=1.21$) more than the ungrammatical sentences with DOM omission and animate objects ($M=3.18$, $SD=1.42$). As for the sentences with inanimate objects, L2 learners rejected the sentences with DOM ($M=4.02$, $SD=1.11$) more often than the sentences with DOM omission ($M=3.67$, $SD=1.33$). Among the L2 learners there was also a lot of variation which suggests that participants had different judgments about the acceptance/rejection of these sentences. The cumulative link mixed model revealed a significant MARKEDNESS effect ($\beta=1.38$, $SE=0.18$, $t=7.65$, $p<0.0001$) and a significant ANIMACY effect ($\beta=1.19$, $SE=0.18$, $t=6.59$, $p<0.0001$). There was also a significant MARKEDNESS*ANIMACY interaction ($\beta=-1.92$, $SE=0.24$, $t=-7.73$, $p<0.0001$), a significant MARKEDNESS*GROUP interaction ($\beta=1.43$, $SE=0.26$, $p<0.0001$), a significant ANIMACY*GROUP interaction ($\beta=1.34$, $SE=0.28$, $t=4.70$, $p<0.0001$) and a

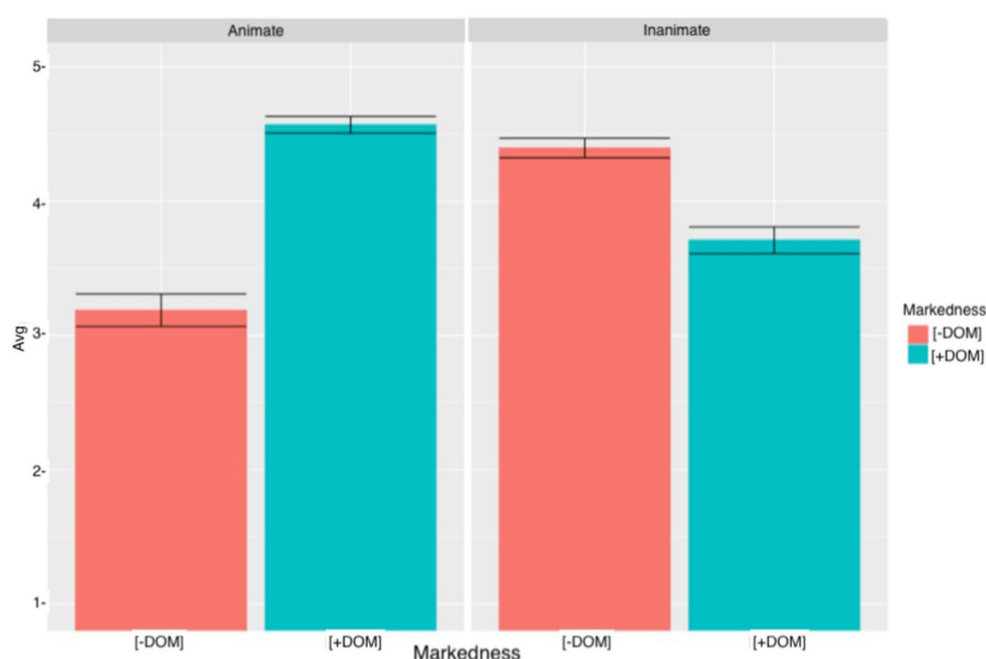


FIGURE 5
Heritage speakers' mean acceptability scores and errors bars (95% CI) for SVO sentences.

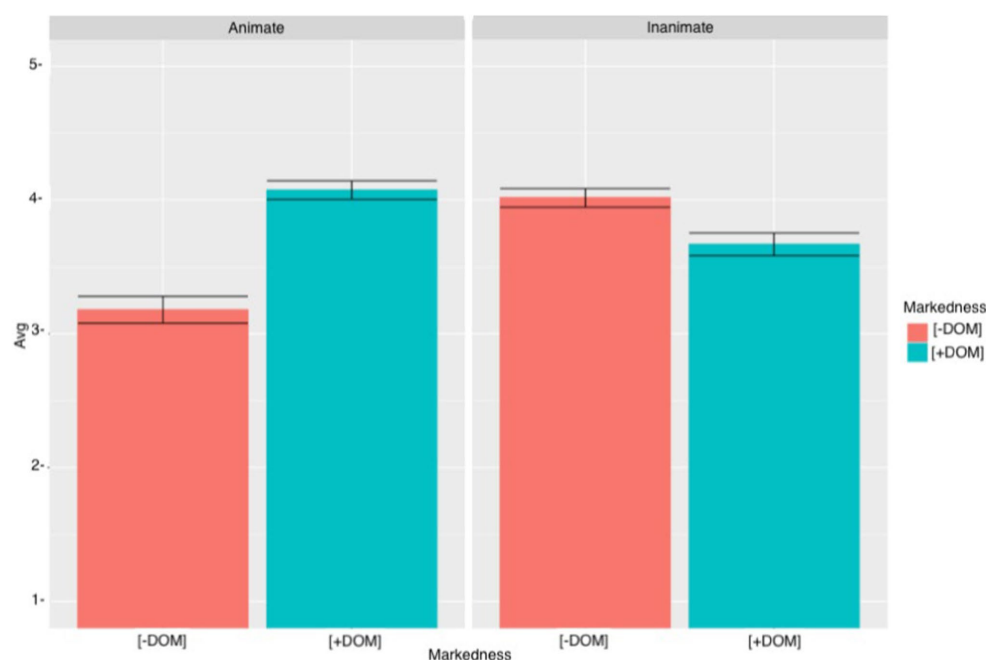


FIGURE 6
L2 learners' mean acceptability scores and errors bars (95% CI) for SVO sentences.

significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = -2.10$, $SE = 0.39$, $t = -5.35$, $p < 0.0001$). *Post hoc* analyses for the three-way interaction revealed that the heritage speakers ($\beta = -1.38$, $SE = 0.18$, $t = -7.65$, $p < 0.0001$) and the L2 learners ($\beta = -2.73$, $SE = 0.23$, $t = -11.76$, $p < 0.0001$) accepted sentences with DOM and animate objects significantly more than sentences with animate objects and DOM omission. Moreover, both groups rejected sentences with inanimate objects and DOM significantly more than sentences with animate objects and DOM ($\beta = 0.54$, $SE = 0.16$, $t = 3.19$, $p = 0.03$) ($\beta = 1.30$, $SE = 0.20$, $t = 6.23$, $p < 0.0001$). Interestingly, when comparing sentences with unmarked animate objects to sentences with marked inanimate objects, there was a significant effect for heritage speakers ($\beta = -0.73$, $SE = 0.20$, $t = -3.53$, $p = 0.009$) and for L2 learners ($\beta = -0.65$, $SE = 0.18$, $t = -3.62$, $p = 0.006$). These results suggest that, for heritage speakers and for L2 learners, there is more of a tendency to expand DOM to inanimate objects than to omit DOM with animate objects. Finally, when comparing sentences with marked animate objects to sentences with unmarked inanimate objects, there was not a significant effect for the heritage speakers ($\beta = 0.19$, $SE = 0.16$, $t = 0.16$, $p = 0.94$), but the difference was significant for the L2 learners ($\beta = 0.69$, $SE = 0.22$, $t = 3.07$, $p = 0.04$). The L2 learners accepted marked animate objects significantly more than unmarked objects. Proficiency was not significant, which suggests that participants' proficiency did not have an effect on their acceptability ratings.

Following previous studies, heritage speakers and L2 learners were expected to accept sentences with animate objects and DOM omission. Results showed that heritage speakers and L2 learners showed some acceptance of unmarked animate objects, but both groups still rated sentences with animate objects and DOM significantly higher. Nevertheless, as Figures 5, 6 show, there was a great deal of variation among heritage speakers' and L2 learners' responses, and while some participants appeared to reject sentences

with animate objects and no DOM, others accepted them. As for the sentences with inanimate objects, participants were expected to accept sentences with DOM omission and reject sentences with DOM. While results revealed a significant effect between sentences with DOM and sentences with DOM omission, participants did not always reject sentences with DOM, and there was notable variation among their answers. Moreover, heritage speakers and L2 learners preferred the extension of DOM to inanimate objects over the omission of DOM with animate objects. Because PROFICIENCY did not turn out to be significant ($\beta = 0.006$, $SE = 0.01$, $t = 0.03$, $p = 0.97$), it appears that participants' proficiency does not have an effect on their judgments.

VSO sentences

Figure 7 shows the results obtained for the heritage speakers. With animate objects, heritage speakers rated the sentences with DOM ($M = 3.46$, $SD = 1.39$) higher than the sentences with DOM omission ($M = 2.47$, $SD = 1.35$). However, when accepting sentences with inanimate objects, heritage speakers accepted ungrammatical DOM omission ($M = 3.85$, $SD = 1.42$) more than the use of DOM ($M = 3.45$, $SD = 1.49$).

Regardless of the type of the object or the use of DOM, there was variation on heritage speakers' answers regarding the acceptance of these sentences. Similar to the heritage speakers, when judging the sentences with animate objects, the L2 learners rated the sentences with DOM ($M = 3.57$, $SD = 1.36$) higher than the sentences with DOM omission ($M = 2.84$, $SD = 1.39$) (see Figure 8). As for the sentences with inanimate objects, contrary to what it was hypothesized, L2 learners preferred the sentences with DOM ($M = 3.63$, $SD = 1.28$) over the sentences with DOM omission ($M = 3.53$, $SD = 1.42$). Overall, there was a lot of variation in L2

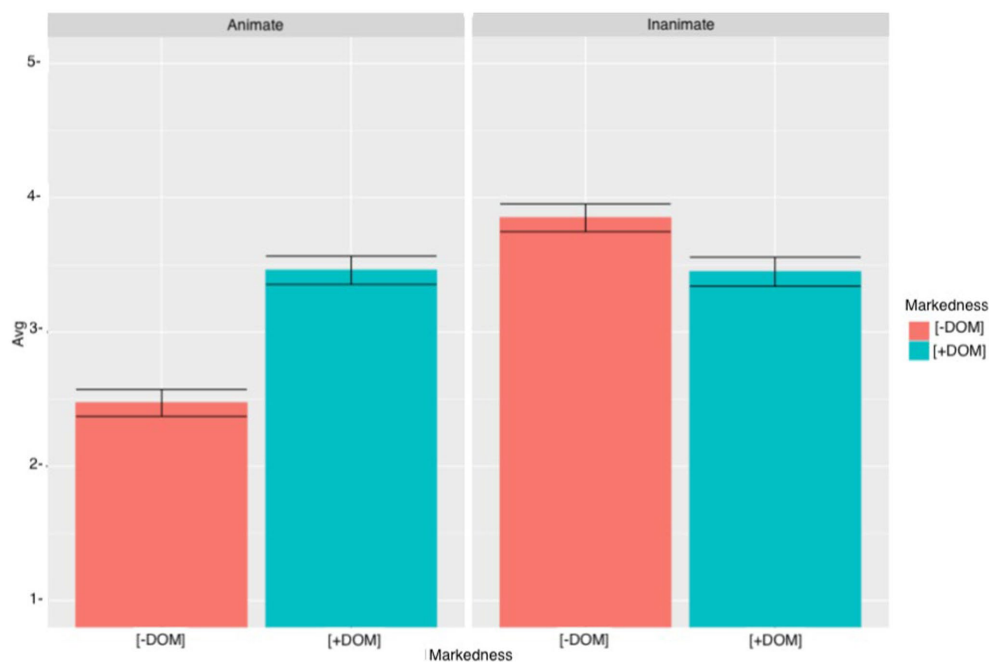


FIGURE 7
Heritage speakers means acceptability scores and errors bars (95% CI) for VSO sentences.

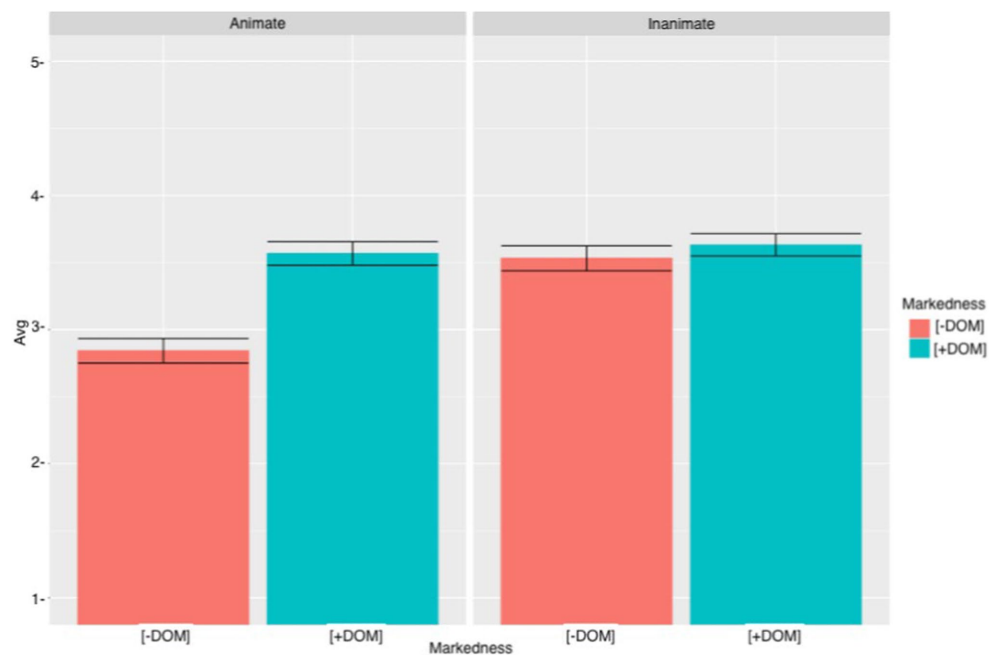


FIGURE 8
L2 learners' mean acceptability scores and errors bars (95% CI) for VSO sentences.

learners' answers. The cumulative link mixed model for VSO sentences revealed a significant MARKEDNESS effect ($\beta = 1.40$, $SE = 0.20$, $t = -6.80$, $p < 0.0001$) and a significant ANIMACY effect ($\beta = 2.24$, $SE = 0.21$, $t = -10.51$, $p < 0.0001$). There was also a significant MARKEDNESS * ANIMACY interaction ($\beta = -2.19$, $SE = 0.29$, $t = -7.45$, $p < 0.0001$), a significant ANIMACY * GROUP interaction ($\beta = -1.21$, $SE = 0.27$, $t = -4.40$, $p < 0.0001$) and a

significant MARKEDNESS*ANIMACY*GROUP ($\beta = 1.20$, $SE = 0.38$, $t = 3.13$, $p < 0.001$). Tukey's multiple comparison test for the three-way interaction revealed a significant effect when comparing sentences with animate objects with and without DOM for heritage speakers ($\beta = -1.40$, $SE = 0.20$, $t = -6.80$, $p < 0.0001$) and L2 learners ($\beta = -1.06$, $SE = 0.17$, $t = -5.99$, $p < 0.0001$). However, when comparing sentences with inanimate objects with DOM and without

DOM, there was only a significant effect for heritage speakers ($\beta = 0.79$, $SE = 0.20$, $t = -3.82$, $p < 0.0001$), but not for L2 learners ($\beta = -0.07$, $SE = 0.17$, $t = -0.43$, $p = 0.99$). Therefore, only the heritage speakers rejected the extension of DOM to inanimate objects. When comparing unmarked animate objects to marked inanimate objects, there was a significant effect for heritage speakers ($\beta = -1.45$, $SE = 0.20$, $t = -7.09$, $p < 0.0001$) and for L2 learners ($\beta = -0.07$, $SE = -1.11$, $t = -6.27$, $p < 0.0001$). These results suggest that heritage speakers and L2 learners prefer DOM with inanimate objects over the omission of DOM with animate objects. However, when comparing sentences with marked animate objects to sentences with unmarked inanimate objects there was a significant effect for heritage speakers ($\beta = -0.84$, $SE = 0.20$, $t = -4.03$, $p = 0.011$), but the difference was not significant for the L2 learners ($\beta = 0.03$, $SE = 0.17$, $t = 0.19$, $p = 1.00$). Heritage speakers, but not L2 learners, accepted unmarked inanimate objects significantly more than marked animate objects. Finally, proficiency did not turn out to be significant ($\beta = 0.01$, $SE = 0.02$, $t = 0.75$, $p = 0.45$), which suggests that participants' proficiency did not have an effect on their judgments.

Summary of results

As hypothesized, heritage speakers and L2 learners did not completely reject the omission of DOM with animate objects in any of the contexts. In most cases, heritage speakers and L2 learners appeared to be undecided when judging unmarked animate objects. However, heritage speakers and L2 learners with SVO and VSO rejected the omission of DOM with animate objects more than the use of DOM with inanimate objects. In fact, results obtained from the sentences containing inanimate objects were unexpected, as neither the heritage speakers nor the L2 learners showed a strong rejection of the use of DOM with inanimate objects. Moreover, proficiency did not appear to be significant in any of the analyses, and thus, contrary to what was predicted (Montrul, 2010), participants with higher proficiency did not behave differently than participants with lower proficiency.

It was also hypothesized that word order would have an effect on participants' judgments. Results partially support this hypothesis as word order had an effect only on sentences with inanimate objects and only with L2 learners: for sentences with a non-canonical word order, L2 learners' judgment did not differ between sentences with DOM and sentences with DOM omission. Therefore, results suggest that L2 learners sometimes accepted the use of DOM with inanimate objects. Participants may accept DOM with inanimate objects due to an overgeneralization error. However, because they did not accept the use of DOM with inanimate objects in sentences with a canonical word order, the fact that they accept DOM extension in VSO sentences may be more related to the word order of these sentences. When reading sentences with non-canonical word order, participants may find these sentences unnatural and pay less attention to the use of DOM. In fact, heritage speakers' and L2 learners' ratings were overall lower with sentences following a non-canonical word than with sentences following a canonical word order (SVO).

Heritage speakers and L2 learners showed some DOM retraction in both the oral tasks and the AJT. However, results suggest an opposite production-comprehension asymmetry: while heritage speakers showed more DOM omission in the AJT than in the oral tasks, L2 learners showed more DOM omission in the oral tasks than in the AJT. The next step is to analyze their online comprehension. It

TABLE 6 Sample sentences used in the eye-tracking task.

Direct object	[+DOM]	[-DOM]
Animate	El actor liberó al compañero con su llave.	*El actor liberó el compañero con su llave.
	'The actor freed DOM the companion with his key.'	'The actor freed the companion with his key.'
Inanimate	*El joven movió al sofá a la calle para dormir.	El joven movió el sofá a la calle para dormir.
	'The young man moved DOM the sofa to the street to sleep.'	'The young man moved the sofa to the street to sleep.'

seems that heritage speakers integrate DOM into their processing, and that is why they are able to almost always produce it. As for the L2 learners, following the MSIH, they may also integrate DOM into their online comprehension, but due to production specific problems brought on by communicative pressure, DOM is not part of their productive knowledge.

Reading comprehension task with eye-tracking

The aim of this task was to test heritage speakers and L2 learners' sensitivity to DOM while reading. This task measured participants' sensitivity to DOM during reading comprehension. The basic assumption in reading tasks with eye-tracking is that participants' eye movements are slower (fixed on the target longer) or produce more regressions (return to a specific region) when reading something unexpected. For example, when presented with sentences such as **Juan vio el policía* 'Juan saw the policeman' and *Juan vio al policía* 'Juan saw DOM-the policeman', participants are expected to take longer to read the first sentence or produce more regressions if they are aware that animate and specific objects must be marked with DOM.

Participants read sentences that varied by MARKING ([+DOM] vs. [-DOM]) and animacy of the object (animate vs. inanimate) and word order (SVO and VOS). Table 6 shows examples of the sentences used in this task.

Notice in (9) that all objects (e.g., *compañero*, *sofá*) were singular and masculine objects with the case marker merged with the article ($a + el = al$). In this way, it is possible to compare 'el' versus 'al' because they are segments of equal length. All sentences were between 8 and 9 words in length and were preceded by a prepositional phrase because it is recommended to avoid having the critical, or even the spillover, region at the beginning of a sentence in eye-tracking with text tasks. Fixations tend to be longer at the beginning of a sentence and people often make corrective saccades (Rayner, 1979; Heller, 1982). All experimental sentences and fillers were followed by comprehension questions about the content of the sentences. The fillers used in this task were very similar to the filler sentences used in the AJT. The comprehension questions had nothing to do with agent/patient relationships so as not to direct the participants' attention to the experimental manipulation, as in (9).

- (9) El actor liberó al compañero con su llave.
'The actor released his partner with his key.'
¿Qué usó el actor?

‘What did the actor use?’

A) Una llave
a key’

B) Unas tijeras.
‘a pair of scissors’

Based on previous studies, heritage speakers and L2 learners were expected to show no sensitivity to DOM with animate objects with canonical word order sentences (Arechabaleta-Regulez, 2016; Jegerski, 2018; Jegerski and Sekerina, 2019). Therefore, participants were expected to produce comparable reading times when reading sentences with marked animate objects than with sentences with DOM omission. As for sentences with inanimate objects, heritage speakers and L2 learners were expected to show sensitivity to DOM (Jegerski, 2018). Therefore, they were expected to produce longer reading times and more regressions with marked than with unmarked inanimate objects. Moreover, word order was hypothesized to play a role in participants’ sensitivity to DOM. If Heritage speakers and L2 learners showed some sensitivity to DOM, it would be more prominent with non-canonical word order sentences than sentences with canonical word order (Arechabaleta-Regulez, 2016; Jegerski and Sekerina, 2019). However, proficiency was expected to play a role and only those participants with a high proficiency in Spanish are expected to show DOM sensitivity, particularly with objects in sentences with non-canonical word order.

Reading task with eye-tracking

Results

Eye movement data was analyzed off-line to identify fixations and saccades using the DataViewer software package (SR Research Ltd., version 1.11.1). Data for the reading task with eye-tracking was analyzed with the *lmer* (linear mixed effect regression) function in the *lme4* package (Bates et al., 2015) using R (version 1.1.453 for Mac OS X, R Development Core Team, 2014) for every eye movement measurement. For all analyses, reading times were the dependent variable while markedness ([+DOM] vs. [-DOM]), animacy of the object (animate vs. inanimate) and group (heritage speakers vs. L2 learners) were all fixed effects. Subject and item were both included as random effects. Proficiency scores were included as covariates to assess the extent to which proficiency affected participants’ processing. When significant interactions were found, a Tukey’s multiple comparison *post hoc* test was performed with *lmeans* package to conduct multiple pairwise comparisons of the fixed variables and their interactions. To ensure that the descriptive and statistical analyses included only sentences that participants understood, sentences with incorrect responses to the post-stimulus comprehension questions were excluded from the analyses. Also, all fixations shorter than 80 ms and longer than 1,200 ms were excluded (Rayner, 1979). In total, this excluded 15.1% of the data (see Table 7).

TABLE 7 Mean accuracy scores for the comprehension questions.

	Heritage Speakers	L2 learners
Correct	91.20%	89.4%
Incorrect	8.8%	10.6%

Table 7 shows that, overall, heritage speakers were more accurate than the L2 learners with the post-stimulus comprehension questions; however, there was not a significant GROUP comparison ($\beta = 10.77$, $SE = 9.33$, $t = 1.01$, $p = 0.22$). Results for each type of sentence are discussed in the following subsections. Each discussion begins with a table displaying the mean reaction times in milliseconds as well as the standard errors for each of the 5 reading times and in each of the 4 regions: the Critical Region, Region 4, Region 5 and Region 6. Notice in (10) that all sentences were divided into 8 different regions (R) of interest. While the Critical Region (CR) was Region 3 (the region in which DOM is either used or omitted), processing effects could occur after the Critical Region (spillover effect) (Arechabaleta-Regulez, 2016). Therefore, not only the CR, but also Region 4(R4), Region 5(R5) and Region 6 (R6) were analyzed. Five reading times were analyzed: *second pass reading times*, *total reading times*, *number of regressions out* and *number of regression in*. Second pass reading times were analyzed to measure the time participants spend in each region when re-reading the sentence. Total reading times were run to measure the total time participants spent in each region of the sentence. Finally, number of regressions out and number of regressions in were calculated for each sentence. While number of regressions out refers to the number of times a region was exited (with an eye regression) to a previous region, number of regressions in refers to the number of times a region was entered (with an eye regression) from a later region. Only significant effects and significant interactions are discussed.

(10)

El	actor	libe	ró al	compañero	von	su	llave
R1	R2	CR	R4	R5	R6	R7	

SVO sentences

Supplementary Tables 8, 9 show the mean reaction times in milliseconds as well as the standard deviation for each of the 5 reading times and in each of the 4 regions analyzed for SVO sentences for heritage speakers and L2 learners, respectively.

Total reading times

In Region 4, there was a significant MARKEDNESS*ANIMACY interaction ($\beta = 89.05$, $SE = 42.50$, $t = 2.09$, $p = 0.03$). As Supplementary Table 10 shows, when reading sentences with animate objects, heritage speakers and L2 learners needed more time to read sentences with DOM omission than with DOM; however, when reading sentences with inanimate objects, heritage speakers and L2 learners needed more time to read sentences with DOM than with DOM omission. Because the animacy of the object caused opposite effects to the use or omission of DOM, the result is an interaction between the two factors without a main effect (known as a crossover interaction). Therefore, it appears that both groups were sensitive to the omission of DOM with animate objects and to the use of DOM with inanimate objects. In Region 6, there was a significant ANIMACY* GROUP interaction ($\beta = -89.23$, $SE = 39.4$, $t = -2.26$, $p = 0.02$). Tukey’s multiple comparison test revealed a significant difference between the reading times produced by L2 learners with animate and inanimate objects ($\beta = 62.99$, $SE = 23.07$, $t = 2.73$, $p = 0.03$). As Supplementary Table 11 shows, both groups produced longer reading times with animate than with inanimate objects.

First pass reading times

In Region 5, there was a significant MARKEDNESS* GROUP interaction ($\beta = 25.62$, $SE = 12.96$, $t = 1.97$, $p = 0.04$). Tukey's test did not reveal any significant comparisons. Heritage speakers and L2 learners behaved differently with regard to the use or omission of DOM. Heritage speakers took longer to read sentences without DOM, while L2 learners produced longer reading times with DOM-marked objects (crossover interaction) (see [Supplementary Table 12](#)).

Second pass reading times

There was a significant MARKEDNESS*ANIMACY interaction in Region 4 ($\beta = 96.02$, $SE = 41.89$, $t = 2.29$, $p = 0.02$). Tukey's multiple comparison tests did not reveal any significant comparisons (crossover interaction). As [Supplementary Table 13](#) shows, participants took longer to read unmarked objects than marked objects with sentences containing animate objects; however, for sentences containing inanimate objects, participants needed more time to read DOM-marked objects than unmarked objects. In Region 6, there was a significant GROUP effect ($\beta = 97.9$, $SE = 35.85$, $t = 2.73$, $p = 0.006$) and a significant ANIMACY* GROUP interaction ($\beta = -85.17$, $SE = 37.43$, $t = -2.27$, $p = 0.02$). Tukey's test revealed a significant comparison between the reading times produced by the heritage speakers and the L2 learners when reading sentences with animate objects ($\beta = -87.60$, $SE = 30.96$, $t = -2.82$, $p = 0.02$). There was also a significant comparison between the heritage speakers' reading times when reading sentences with inanimate objects and the L2 learners' reading times when reading sentences with animate objects ($\beta = -100.46$, $SE = 32.96$, $t = -3.04$, $p = 0.01$). Finally, the analysis found a significant difference for the L2 learners with sentences with animate and inanimate objects ($\beta = 65.94$, $SE = 21.00$, $t = -3.13$, $p = 0.01$). As [Supplementary Table 14](#) shows, heritage speakers and L2 learners took longer to read sentences with animate than with inanimate objects.

Number of regressions in

Number of regressions in revealed a significant ANIMACY effect ($\beta = -9.57$, $SE = 4.74$, $t = -2.01$, $p = 0.04$), a significant PROFICIENCY effect ($\beta = 4.88$, $SE = 2.48$, $t = 1.96$, $p = 0.05$) and a significant MARKEDNESS*ANIMACY interaction ($\beta = 1.24$, $SE = 6.72$, $t = 1.84$, $p = 0.05$) in Region 4. Tukey's multiple comparison test did not reveal any significant differences. With animate objects, heritage speakers and L2 learners both produced more regressions in with when animate objects were unmarked. However, they produced more regressions in with sentences that contained DOM-marked inanimate objects (see [Supplementary Table 15](#)). Proficiency turned out to be significant because participants with a lower proficiency produced more regressions overall than participants with a higher proficiency. In Region 5, there was a significant ANIMACY*GROUP interaction 2 ($\beta = 252.35$, $SE = 91.85$, $t = 2.75$, $p = 0.005$). As [Supplementary Table 16](#) shows, heritage speakers produced more regressions in with animate objects, while L2 learners did so with inanimate objects. In Region 6, there were not significant effects or significant interactions.

Number of regressions out

Number of regressions out revealed a significant ANIMACY*GROUP interaction ($\beta = 8.98$, $SE = 4.50$, $t = 1.99$, $p = 0.04$)

and a significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = -1.17$, $SE = 6.38$, $t = -1.83$, $p = 0.05$ in Region 4. Tukey's multiple comparison test did not reveal any significant comparisons. As [Supplementary Table 17](#) shows, both heritage speakers and L2 learners produced more regressions out with unmarked animate objects. As for inanimate objects, heritage speakers produced more regressions out with DOM-marked objects, while the L2 learners with unmarked objects. In Region 6, there was a significant ANIMACY*GROUP interaction ($\beta = 1.29$, $SE = 6.43$, $t = 2.01$, $p = 0.04$) and a significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = -1.17$, $SE = 6.38$, $t = -1.74$, $p = 0.05$). Tukey's test did not reveal any significant comparisons for either of the interactions. As [Supplementary Table 18](#) shows, when reading sentences with animate objects, heritage speakers, but not L2 learners, produced more regressions out with sentences with unmarked objects. With inanimate objects, heritage speakers and L2 learners both regressed out with DOM-marked objects.

Sum of skipped targets

In total, heritage speakers skipped DOM or the determiner 'el' 10% of the time and L2 learners 12%. There was not a significant GROUP effect ($\beta = -0.03$, $SE = 0.02$, $t = -1.68$, $p = 0.54$).

Reading times for SVO sentences suggest that heritage speakers and L2 learners were more sensitive to the omission of DOM with animate objects than previously expected. Heritage speakers and L2 learners were also sensitive to the extension of DOM to inanimate objects as previously suggested ([Jegerski, 2018](#)). With total reading times, with first pass reading times, and with regressions in, there was a significant MARKEDNESS*ANIMACY interaction in region 4, as heritage speakers and L2 learners produced longer reading times or more regressions with unmarked animate objects than with marked animate objects and with marked inanimate objects than with unmarked inanimate objects. However, in later regions, it is important to note that heritage speakers seemed to be affected by DOM omission regardless of the animacy of the object, while L2 learners were affected by DOM regardless of the animacy of the object. Finally, proficiency was only significant with regressions in, as participants with a lower proficiency produced more regressions than participants with a higher proficiency.

VSO sentences

The mean reaction times in milliseconds and the standard deviation for the 7 reading times and for all the 4 regions are represented in [Supplementary Table 19](#) for the heritage speakers and in [Supplementary Table 20](#) for the L2 learners.

Total reading times

Total Reading times did not show any significant effects or any significant interactions in any of the 4 regions that were analyzed. Therefore, with VSO sentences, the use or non-use of DOM did not cause any processing difficulties for the heritage speakers or for the L2 learners.

First pass reading times

First pass reading times revealed a significant GROUP effect ($\beta = -38.57$, $SE = 13.69$, $t = -2.81$, $p = 0.005$) in the Critical Region. In

Region 4, there was a significant MARKEDNESS*GROUP interaction ($\beta = -23.62$, $SE = 12.94$, $t = -1.82$, $p = 0.05$). Tukey's multiple comparison test did not reveal any significant comparisons. Heritage speakers and L2 learners reacted differently to the use or omission of DOM. Regardless of the animacy of the object, heritage speakers produced longer reading times when sentences with DOM, while L2 learners produced longer reading times with sentences without DOM (see [Supplementary Table 21](#)). In Region 5, there was also a significant ANIMACY*GROUP interaction ($\beta = -2.51$, $SE = 1.38$, $t = -1.82$, $p = 0.04$). As [Supplementary Table 22](#) shows, while heritage speakers needed more time to read sentences with inanimate objects, L2 learners needed more time to read sentences with animate objects.

Second pass reading times

In the Critical Region ($\beta = 105.65$, $SE = 43.04$, $t = 2.45$, $p = 0.01$) and in Region 4 ($\beta = 92.25$, $SE = 45.14$, $t = 2.04$, $p = 0.04$), there was a significant GROUP effect. Overall, heritage speakers were faster readers than L2 learners. However, there were not any significant effects or significant interactions.

Number of regressions in

In the Critical Region, there was a significant MARKEDNESS*ANIMACY interaction ($\beta = 1.15$, $SE = 6.51$, $t = 1.76$, $p = 0.04$) and a significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = -1.64$, $SE = 9.11$, $t = -1.80$, $p = 0.03$). Tukey's multiple comparison test did not reveal any significant comparisons in any of the interactions, as participants reacted to the use or omission of DOM differently (see [Supplementary Table 23](#)). When reading sentences with animate objects, heritage speakers produced more regressions in with sentences that omitted DOM than with sentences containing DOM. As for the L2 learners, they produced more regressions in with sentences with DOM than with sentences that omitted DOM. As for sentences with inanimate objects, heritage speakers produced more regressions in with sentences with DOM, while L2 learners produced more regression in with sentences without DOM. In Region 6, there was a significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = -1.82$, $SE = 1.01$, $t = -1.80$, $p = 0.04$). In this region, Tukey's test did not reveal any significant effects. However, estimated marginal means showed the same trend as in the critical region. When reading sentences with animate objects, heritage speakers produced more regressions in with sentences without DOM, while L2 learners produced more regressions in with sentences with DOM. As for sentences with inanimate objects, heritage speakers produced more regressions in with sentences with DOM, while L2 learners produced more regressions in with sentences without DOM (see [Supplementary Table 24](#)).

Number of regressions out

Number of regressions out revealed a significant MARKEDNESS*GROUP interaction ($\beta = -1.36$, $SE = 5.41$, $t = -2.52$, $p = 0.01$) and a significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = 1.46$, $SE = 7.47$, $t = 1.95$, $p = 0.05$) in the Critical Region. Tukey's multiple comparison test revealed an almost significant comparison between the regressions out produced by the L2 learners when reading sentences with animate objects and DOM omission and sentences with animate objects and DOM ($\beta = 0.11$, $SE = 0.04$, $t = 2.99$, $p = 0.06$). As [Supplementary Table 25](#) shows, when reading sentences with animate objects, heritage speakers and L2 learners produced more regressions out with unmarked objects than with marked objects. As

for sentences with inanimate objects, heritage speakers produced more regressions with marked objects, while L2 learners produced more regressions out with sentences that omitted DOM (see [Supplementary Table 25](#)). In Region 4, there was also a significant MARKEDNESS*ANIMACY*GROUP interaction ($\beta = -1.87$, $SE = 9.25$, $t = -2.02$, $p = 0.04$). For this region, Tukey's test did not reveal any significant comparisons. As [Supplementary Table 26](#) shows, heritage speakers produced more regressions out with unmarked animate objects and with marked inanimate objects than with marked animate objects and unmarked inanimate objects, respectively. L2 learners on the other hand, produced more regressions out with unmarked objects regardless of the animacy of the object. There was also a significant PROFICIENCY effect in Regions 5 ($\beta = -4.39$, $SE = 2.41$, $t = -1.82$, $p = 0.07$) and 6 ($\beta = -7.44$, $SE = 3.42$, $t = -2.17$, $p = 0.03$).

Sum of skipped targets

Overall, heritage speakers skipped the Critical Region 16% of the time and L2 learners 19% of the time. The pairwise comparison did not show a significant GROUP effect ($\beta = -0.04$, $SE = 0.02$, $t = -2.37$, $p = 0.16$).

Overall, reading times for sentences with VSO word order showed mixed results regarding the sensitivity to DOM with animate and inanimate objects by the heritage speakers and the L2 learners. The heritage speakers, as suggested by total reading times and second pass reading times, produced longer reading times with marked animate objects than with unmarked animate objects in early regions (R2, R3, R4 and R5); however, in later regions (R6, R7 and R8) they showed the opposite pattern: they produced longer reading times with sentences without DOM than with sentences with DOM. These results may suggest that their sensitivity to DOM omission with animate objects happened only in later regions as a spillover effect. Regressions in and regressions out supported this possibility, as the heritage speakers showed sensitivity to the omission of DOM with animate objects: heritage speakers produced more regressions (in and out) with unmarked than with marked animate objects. As for sentences with inanimate objects, sensitivity to the use or omission of DOM was only perceived by regressions in and regressions out: the heritage speakers produced more regressions (in and out) with marked inanimate objects than with unmarked inanimate objects. The L2 learners showed less sensitivity to the use or omission of DOM regardless of the animacy of the object. Some type of DOM sensitivity was only perceived by regressions in and regressions out and only for animate objects. The L2 learners tended to produce more regressions (in and out) with unmarked animate objects than with marked animate objects.

Summary of results

The aim of this task was to test participants' sensitivity to DOM while reading. With SVO sentences, heritage speakers and L2 learners were not expected to show DOM sensitivity. However, with non-canonical sentences participants were expected to show some sensitivity ([Arechabaleta-Regulez, 2016](#)) by producing longer reading times with unmarked animate objects than with marked animate objects and with marked inanimate objects than with unmarked inanimate objects. Heritage speakers and L2 learners were expected to rely on processing mechanisms (word order) in their stronger language (English) to comprehend these sentences instead of object

marking. However, with non-canonical sentences (VSO), participants were expected to rely more on DOM and thus show more sensitivity to it, as it was more relevant for comprehending these sentences. Finally, participants with a higher proficiency were expected to show more DOM sensitivity with both, animate and inanimate objects.

Contrary to what was hypothesized, the heritage speakers and the L2 learners showed more DOM sensitivity with SVO sentences than with VSO sentences. Sensitivity to DOM with SVO sentences happened with later measures, as both groups of bilinguals produced longer reading times and/or more regressions with sentences with unmarked animate objects than with marked animate objects and with marked inanimate objects than with unmarked inanimate objects. As for VSO sentences, heritage speakers appeared to show more sensitivity to DOM than the L2 learners. However, sensitivity to DOM by the heritage speakers was only observable with regressions (in and out) and in later regions of the sentences. Thus, sensitivity did not appear squarely in the Critical Region but as a post-critical effect. The L2 learners, on the other hand, did not show sensitivity to DOM with either animate or inanimate objects, contrary to what was hypothesized. Finally, while participants were expected to skip DOM more often with SVO than with VSO sentences, results did not support this hypothesis. In fact, heritage speakers and L2 learners skipped DOM most with VSO sentences.

Discussion and conclusion

The purpose of this study was to analyze the production, acceptability and online comprehension of Spanish DOM by two groups of bilingual speakers living in the U.S.: heritage speakers and L2 learners. While previous studies have reported these two groups of bilingual speakers tend to omit DOM with animate objects (Farley and McCollam, 2004; Guijarro-Fuentes and Marinis, 2007; Bowles and Montrul, 2008, 2009; Montrul, 2010, 2014; Montrul and Sánchez-Walker, 2013; Arechabaleta-Regulez, 2014; Jegerski, 2018; Jegerski and Sekerina, 2019), these studies have usually examined production, acceptability or online comprehension of DOM in isolation. Few, if any, have compared all three of these aspects with the same group of speakers. Therefore, this study employed tasks designed to elicit data related to all of participants' production, acceptability and online comprehension. The same group of heritage speakers and the same group of L2 learners completed all the tasks.

First, the oral tasks were used to analyze heritage speakers' and L2 learners' production of DOM. It was predicted that both groups would show significant DOM omission with animate objects, but that heritage speakers would show less DOM omission overall (Montrul, 2010). In addition, proficiency in Spanish was expected to play an important role. Lastly, heritage speakers' and L2 learners' production of DOM was hypothesized to depend on the type of task because the narrative task was seen as a more implicit task than the elicitation task. Results showed that: (1) as predicted, both groups omitted DOM with animate objects; (2) L2 learners showed more cases of DOM omission than heritage speakers; (3) proficiency played a significant role, as participants with a higher proficiency in Spanish marked animate objects more than participants with a lower proficiency; (4) proficiency was more significant for L2 learners than for heritage speakers; and (5) type of task indeed had an effect but not the effect that was expected. Results partially supported the hypotheses, as only the

heritage speakers were affected by the type of task. The heritage speakers showed more DOM retraction in the elicitation task than in the oral task and also produced significantly more DOM omission to inanimate objects in the elicitation task than in the narrative task. However, contrary to what was hypothesized, the L2 learners did not show less DOM omission in the elicitation task than in the narrative task. In fact, the L2 learners behaved very similarly in the two oral tasks.

Second, the aim of the AJT was to analyze participants' knowledge of DOM. Following previous studies, both heritage speakers and L2 learners were expected to accept DOM omission with animate objects and to reject the use of DOM with inanimate objects. Additionally, word order was manipulated in the AJT. It was hypothesized that participants would have to pay closer attention to the use or omission of DOM in sentences following a non-canonical word order and thus might show less DOM attrition with VSO sentences. Finally, proficiency was expected to play a role. Results showed that: (1) as predicted, heritage speakers and L2 learners had difficulty rejecting sentences with DOM omission and animate objects; (2) surprisingly, both heritage speakers and L2 learners also had a hard time rejecting sentences with DOM extension to inanimate objects, especially L2 learners. In fact, the acceptability ratings given by L2 learners to VSO sentences with DOM did not significantly differ from those given to sentences with DOM omission; (3) results did not support the hypothesis on the effects of word order. Participants did not integrate DOM more in sentences following a non-canonical word order. Overall, SVO sentences were rated higher than VSO sentences. The rejection of VSO sentences may be due to the fact that they follow a non-canonical word order. Thus, heritage speakers and L2 learners may not be as familiar with these sentences and could perceive them as less acceptable regardless of the use of DOM or the animacy of the objects. Finally, (4) proficiency did not turn out to be significant. Therefore, heritage speakers and L2 learners' performance did not depend on their proficiency in Spanish.

Considered together, the results obtained in the oral tasks and the results obtained in the AJT support the relevance of language experience and practice in language acquisition. Because heritage speakers have acquired DOM orally and implicitly, they relied on implicit knowledge and integrated DOM more in the oral tasks than in the AJT. L2 learners, on the other hand, have acquired DOM in the classroom and most likely *via* metalinguistic explanations. Thus, they applied that explicit knowledge in the AJT but not in the oral tasks (as suggested by the MSIH; Prévost and White, 1999, 2000). Nevertheless, as suggested by the results obtained in the reading task with eye-tracking, L2 learners can still integrate DOM into their online comprehension.

The reading task with eye-tracking aimed to analyze participants' processing mechanisms to test whether DOM omission is part of their competence. It was hypothesized that heritage speakers would show little sensitivity to unmarked animate objects, at least with sentences following a canonical word order (Jegerski, 2015, 2018; Arechabaleta-Regulez, 2016; Jegerski and Sekerina, 2019). Moreover, heritage speakers and L2 learners were expected to show sensitivity to DOM with inanimate objects regardless of the word order (Jegerski, 2018). Lastly, proficiency was also thought to play an important role, and participants with a higher proficiency were

expected to show more DOM sensitivity. Results showed that, contrary to what was predicted, both heritage speakers and L2 learners showed more DOM sensitivity with canonical word order sentences than with non-canonical word order sentences. With SVO sentences, the heritage speakers and the L2 learners showed sensitivity to unmarked animate objects and marked inanimate objects in late reading measures and immediately after the critical region. With VSO sentences, only the heritage speakers showed a degree of sensitivity to DOM omission with animate objects and to the use of DOM with inanimate objects. Sensitivity was only perceived with regressions in and out. Finally, proficiency did not play an important role in the reading mechanisms produced by the participants.

All in all, results showed that DOM variation exists among heritage speakers and L2 learners. Both heritage speakers and L2 learners can integrate DOM into their production, judgments and processing, but they do so inconsistently. Type of task and type of sentence each have an effect on speakers' use of DOM. These effects were not always the same for both heritage speakers and L2 learners, which corroborates the importance that language experience and language practice have on speakers' actual use of DOM.

Data availability statement

The original contributions presented in the study are included in the article/[Supplementary material](#), further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by University of Illinois at Urbana Champaign. The patients/participants provided their written informed consent to participate in this study.

References

- Aissen, J. (2003). Differential object marking: iconicity vs. economy. *Nat. Lang. Ling Theory* 21, 435–483. doi: 10.1023/A:1024109008573
- Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). Wiley.
- Alarcon, I. (2010). Advanced heritage learners of Spanish: a sociolinguistic profile for pedagogical purposes. *Foreign Lang. Ann.* 43, 269–288. doi: 10.1111/j.1944-9720.2010.01078.x
- Alfaraz, G. (2011). Accusative object marking: a change in progress in Cuban Spanish? *Span. Context* 8, 213–234. doi: 10.1075/sic.8.2.02alf
- Arechabaleta-Regulez, B. (2014). *Acquisition of Differential Object Making by L2 speakers of (unpublished master thesis)*. Champaign-Urbana, IL: University of Illinois.
- Arechabaleta-Regulez, B. (2016). *Online sensitivity to DOM violations by Spanish heritage speakers*. in "Paper presented at the international workshop of the Unity and Diversity in differential object marking, Institut national des Langues et Civilisations Orientales, Paris, France.
- Arechabaleta-Regulez, B., and Montrul, S. (2021). Psycholinguistic evidence for incipient language change in Mexican Spanish: the extension of differential object marking. *Languages* 6:131. doi: 10.3390/languages6030131
- Au, T., Knightly, L., Jun, S., and Oh, J. (2002). Overhearing a language during childhood. *Psychol. Sci.* 13, 238–243. doi: 10.1111/1467-9280.00444
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). lme4: linear mixed-effects models Using Eigen and S4. R package version 1.1.1-1.
- Beaudrie, S., and Ducar, C. (2005). Beginning level university heritage programmes: creating a space for all heritage language learners. *Heritage Lang. J.* 3, 1–26. doi: 10.46538/hlj.3.1.1
- Benmamoun, A., Montrul, S., and Polinsky, M. (2010). *Prolegomena to heritage linguistics*. Unpublished white paper. Champaign, IL: University of Illinois at Urbana-Champaign and Harvard University.
- Bowles, M. (2011). Measuring implicit and explicit knowledge. What can heritage language learners contribute? *Stud. Second. Lang. Acquis.* 33, 247–271. doi: 10.1017/S0272263110000756
- Bowles, M., and Montrul, S. (2008). "The role of explicit instruction in the L2 acquisition of the a-personal" in *Selected papers from the 10th Hispanic linguistics symposium*. eds. J. Bruhn de Garavito and E. Valenzuela (Somerville, MA: Cascadilla Press), 25–35.
- Bowles, M., and Montrul, S. (2009). "Instructed L2 acquisition of differential object marking in Spanish" in *Little words: their history, phonology, syntax, semantics, pragmatics and acquisition*. eds. R. Leow, H. Campos and D. Lardiere (Washington, DC: Georgetown University Press), 199–210.
- Bullock, B., and Toribio, A. J. (2004). Convergence as an emergent property in bilingual speech. *Biling. Lang. Cogn.* 7, 91–93. doi: 10.1017/S1366728904001506
- Carreira, M., and Kagan, O. (2011). The results of the National Heritage Language Survey: implications for teaching, curriculum design, and professional development. *Foreign Lang. Ann.* 44, 40–64. doi: 10.1111/j.1944-9720.2010.01118.x
- Carter, B. T., and Luke, S. G. (2020). Best practices in eye tracking research. *Int. J. Psychophysiol.* 155, 49–62. doi: 10.1016/j.ijpsycho.2020.05.010
- Christensen, R. H. B. (2015). Ordinal - Regression Models for Ordinal Data. R package version 2015, 6–28. Available at: <http://www.cran.r-project.org/package=ordinal/>

Author contributions

BA and SM: methodology and formal analysis. BA: data collection, writing—original draft preparation, and writing—review and editing. SM: supervision and editing. All authors contributed to the article and approved the submitted version.

Funding

This research was partially funded by the Love Fellowship (UIUC Center for Latin American and Caribbean Studies).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1106613/full#supplementary-material>

- Clahsen, H., and Felser, C. (2006). Grammatical processing in language learners. *Appl. Psycholinguist.* 27, 3–42. doi: 10.1017/S0142716406060024
- Farley, A., and McCollam, K. (2004). Learner readiness and L2 production in Spanish: Processability theory on trial. *Estudios de Lingüística Aplicada* 22, 47–69.
- Fedzechkina, M., Jaeger, T. F., and Trueswell, J. (2015). “Production is biased to provide informative cues early: evidence from miniature artificial languages” in *Proceedings of the 37th annual meeting of the cognitive science society*, eds. D. Noelle, A. Waelaumont, J. Yoshimi, T. Matlock, C. Jennings and P. Maglio (Austin, TX: Cognitive Science Society), 674–679.
- Field, J. (2004). An insight into listeners’ problems: too much bottom-up or too much top-down? *System* 32, 363–377. doi: 10.1016/j.system.2004.05.002
- Guijarro-Fuentes, P. (2012). The acquisition of interpretable features in L2 Spanish: personal a. *Biling. Lang. Cogn.* 15, 701–720. doi: 10.1017/S1366728912000144
- Guijarro-Fuentes, P., and Marinis, T. (2007). Acquiring phenomena at the syntax/semantics interface in L2 Spanish: the personal preposition a. *EUROSLA Yearb.* 7, 67–88. doi: 10.1075/eurosla.7.06gui
- Heller, D. (1982). Eye movements in reading. In R. Groner & P. Fraisse (Eds.), *Cognition and Eye movements* (pp. 487–498). Berlin: Deutscher Verlag der Wissenschaften.
- Ionin, T., Ko, H., Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition*, 12, 3–69.
- Jegerski, J. (2015). The processing of case in near-native Spanish. *Second. Lang. Res.* 31, 281–307. doi: 10.1177/0267658314563880
- Jegerski, J. (2018). The processing of the object marker a by heritage Spanish speakers. *International Journal of Bilingualism*, 22, 585–602.
- Jegerski, J., and Sekerina, I. A. (2019). The processing of input with differential object marking by heritage Spanish speakers. *Bilingualism* 22, 585–602. doi: 10.1017/S1366728919000087
- Jiang, N. (2004). Morphological insensitivity in second language processing. *Appl. Psycholinguist.* 25, 603–634. doi: 10.1017/S0142716404001298
- Krashen, Stephen D. (1982). *Principles and practice in second language acquisition*. New York: Pergamon Press.
- Kutlu, E. and Kircher, R. (2021). A corpus-assisted discourse study of attitudes toward Spanish as a heritage language in Florida. *Languages*, 6, 38.
- Lunn, P. (2002). “Tout se tient in Dominican Spanish” in *Structure, meaning and Acquisition in Spanish*, eds. J. Lee, K. Geelin and C. Clements (Sommerville, MA: Cascadia Press), 65–72.
- MacWhinney, B. (1987). “The competition model. (1987)” in *Mechanisms of language acquisition*, ed. B. MacWhinney (New Jersey: Erlbaum), 249–308.
- MacWhinney, B. (2005). “A unified model of language acquisition,” in *Handbook of bilingualism: Psycholinguistic approaches* eds J. Kroll and Groot A. M. B. de, editors. (New York, NY: Oxford University Press, 49–67
- McCollam Wiebe, K. (2004). The interface between instruction type and learner readiness in the acquisition of Spanish personal a and the subjunctive. Dissertation Abstracts International, A: The Humanities and Social Sciences, 4034-A.
- Mikulski, A. (2006). *Native intuitions, foreign struggles? Knowledge of the subjunctive in volitional constructions among heritage and traditional FL learners of Spanish*. Unpublished PhD dissertation. Iowa, IA: University of Iowa
- Montrul, S. (2010). Dominant language transfer in adult second language learners and heritage speakers. *Second. Lang. Res.* 26, 293–327. doi: 10.1177/0267658310365768
- Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: a case of morpho-syntactic convergenc. *Bilingual. Lang. Cogn.* 7, 125–142.
- Montrul, S. (2014). Searching for the roots of structural changes in the Spanish of the United States. *Lingua* 151, 177–196.
- Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge, UK: Cambridge University Press.
- Montrul, S. (2022). *Native speakers interrupted* Cambridge University Press.
- Montrul, S., and Bowles, M. (2009). Back to basics: Differential Object Marking under incomplete acquisition in Spanish heritage speakers. *Language and Cognition*, 12, 363–383.
- Montrul, S., and Bowles, M. (2010). Is grammar instruction beneficial for heritage language learners? Dative case marking in Spanish. *The Heritage Language Journal*, 7, 47–73.
- Montrul, S., Foote, R., and Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition. *Lang. Learn.* 58, 503–553. doi: 10.1111/j.1467-9922.2008.00449.x
- Montrul, S., and Ionin, T. (2010). Transfer effects in the interpretation of definite articles by Spanish heritage speakers. *Biling. Lang. Cogn.* 13, 449–473. doi: 10.1017/S1366728910000040
- Montrul, S., and Sánchez-Walker, N. (2013). Differential object marking in child and adult Spanish heritage speakers. *Lang. Acquis.* 20, 109–132. doi: 10.1080/10489223.2013.766741
- Prévost, P., and White, L. (1999). Accounting for morphological variability in second language acquisition: Truncation or Missing Inflection?. In *The Acquisition of syntax*, M.-A. Friedemann and L. Rizzi (eds.), 202–235. London: Longman.
- Prévost, P., and White, L. (2000). Missing surface inflection or impairment in second language? Evidence from Tense and Agreement. *Second Language Research*, 16 (2).
- R Development Core Team (2014). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014. Available at <http://www.R-project.org>
- Rayner, K. (1979). “Eye movements in reading: Eye guidance and integration” in *The Processing of Visible Language I* Eds Kolars P A, Wrolstad M, Bouma H (New York: Plenum Press).
- Reynolds, R. R., Howard, K. M., and Deak, J. (2009). Heritage language learners in first-year foreign language courses: a report of general data across learner subtypes. *Foreign Lang. Ann.* 42, 250–269. doi: 10.1111/j.1944-9720.2009.01020.x
- Sánchez, L. (2003). *Quechua-Spanish bilingualism: Interference and convergence in functional categories*. Amsterdam: Johns Benjamins.
- Sánchez, L., and Zdrojewski, P. (2013). Restricciones semánticas y pragmáticas al doblado de clíticos en el español de Buenos Aires y de Lima. *Lingüística* 29, 271–320.
- Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270. doi: 10.1016/j.cognition.2003.10.008
- Valdés, G. (2001). “Heritage language students: profiles and possibilities” in *Heritage languages in America: preserving a national resource*, eds. J. K. Peyton, D. Randard and S. McGinnis (Washington, DC: Center for Applied Linguistics), 37–80.
- von Heusinger, K., and Kaiser, G. (2005). The evolution of differential object marking in Spanish. In: HeusingerKlaus von, Georg Kaiser and Elisabeth Stark (eds) *Specificity and the evolution / emergence of nominal determination Systems in Romance*, 33–69. Konstanz: Universität Konstanz.



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY

Ethan Kutlu,
The University of Iowa, United States
Anamaria Bentea,
University of Konstanz, Germany

*CORRESPONDENCE

Brechje van Osch
✉ brechje.a.osch@uit.no

[†]These authors have contributed equally to this work and share first authorship

SPECIALTY SECTION

This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 02 January 2023

ACCEPTED 30 March 2023

PUBLISHED 02 May 2023

CITATION

van Osch B, Parafita Couto MC, Boers I and
Sterken B (2023) Adjective position in the
code-switched speech of Spanish and
Papiamentu heritage speakers in the
Netherlands: Individual differences and
methodological considerations.
Front. Psychol. 14:1136023.
doi: 10.3389/fpsyg.2023.1136023

COPYRIGHT

© 2023 van Osch, Parafita Couto, Boers I and
Sterken. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The
use, distribution or reproduction in other
forums is permitted, provided the original
author(s) and the copyright owner(s) are
credited and that the original publication in this
journal is cited, in accordance with accepted
academic practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Adjective position in the code-switched speech of Spanish and Papiamentu heritage speakers in the Netherlands: Individual differences and methodological considerations

Brechje van Osch^{1*†}, M. Carmen Parafita Couto^{2,3†}, Ivo Boers^{3,4}
and Bo Sterken²

¹Acqva Aurora Center, Department of Language and Culture, UiT, The Arctic University of Norway, Tromsø, Norway, ²Heritage Linguistics Lab, Leiden University, Center for Linguistics, Leiden, Netherlands, ³Language Variation and Textual Categorization, Faculty of Philology and Translation, University of Vigo, Vigo, Spain, ⁴Department of Netherlandic Studies, Károli Gáspár University, Budapest, Hungary

Introduction: This study examines adjective-noun order in code-switched constructions by heritage speakers of Spanish and Papiamentu in the Netherlands. Given that Dutch differs from Spanish and Papiamentu regarding the default position of the adjective, word order in the nominal domain creates a so-called “conflict site” in code-switching. Most accounts of word order patterns in code-switching focus on structural constraints, such as the matrix language or the strength of the EPP feature in Agr. Thus far, studies comparing the two models have not found compelling evidence for either of them.

Methods: The present study takes a more comprehensive approach and considers several linguistic (matrix language, adjective language, and type of insertion) as well as extra-linguistic variables (e.g., age, age of onset, and patterns of exposure and use). Moreover, we compare heritage speakers of two different heritage languages that are linguistically similar (both Spanish and Papiamentu exhibit postnominal adjectives), and share the same dominant societal language, but are likely to differ from each other in terms of certain sociolinguistic properties. 21 Spanish and 15 Papiamentu heritage speakers (aged 7–54) in the Netherlands carried out a Director-Matcher task, aimed at eliciting nominal constructions containing switches.

Results: The results show that either the ML or the language of the adjective, or both, are important predictors for word order, although the data cannot disentangle these two factors. Moreover, the type of insertion was found to play a role: word order patterns for noun insertions differed from other types of insertions. In addition, the two groups did not behave similarly: Papiamentu speakers were more categorical in their preference for noun-adjective order when inserting Dutch nouns into their heritage language than the Spanish speakers were. Finally, there was a great deal of individual variation, which seemed to be related mostly to the age of the participants: children and teen participants behaved differently from adults.

Discussion: These findings demonstrate that both linguistic and extra-linguistic play a role in determining how heritage speakers deal with conflict sites in the nominal domain. Particularly, the findings suggest that, at least for some communities and in some code-switching modes, children may need more time, or more input, too converge on adult-like code-switching norms.

KEYWORDS

code-switching, heritage bilingualism, adjective position, extra-linguistic variables, individual differences, Papiamentu, Dutch, Spanish

1. Introduction

Heritage speakers (HSs) are bi/multilingual speakers who, like most other multilinguals, commonly use elements from their languages in the same utterance (either within the same sentence or conversation). This phenomenon is known as code-switching (CS; Deuchar, 2012). In studies of heritage language (HL) acquisition, code-switching has often been overlooked, as the focus of most studies is on either the heritage language of the bilingual or their majority/dominant societal language. However, studying code-switching can make important contributions to our knowledge about heritage speakers' grammar, since it allows us to uncover patterns in a bilingual's grammar that remain hidden in the study of unilingual speech alone. In recent decades, a general consensus has emerged that code-switching is rule-governed (cf. Parafita Couto et al., 2021 for an overview). Nevertheless, "no clear evidence has emerged concerning the structural regularities that underlie mixed speech across language pairs, or even within the same language pair in different communities" (Parafita Couto et al., 2023). Recent studies suggest that different code-switching strategies may be used between members of the same community (e.g., Boers et al., 2020) and also that there are cross-community differences between communities that share the same language combinations, suggesting that sociolinguistic variables may in some cases override structural constraints. However, to date, we still do not have a clear picture of how the interaction of different linguistic and extralinguistic components shapes code-switching outcomes (Stell and Yakpo, 2015).

In this study, we look at two separate communities of heritage speakers who differ from each other in terms of age (comparing children, teens, and adults) as well as age of onset and patterns of use and exposure, in order to investigate which, if any, of these factors play a role in determining code-switching patterns. We focus on heritage speakers of Spanish and Papiamentu who live in the Netherlands, targeting switching where the structures of the two languages differ (conflict sites, cf. Vaughan-Evans et al., 2020 for a recent overview). In particular, we address word order in adjective-noun switches. Adjectives are pre-nominal in Dutch and (mostly) post-nominal in Papiamentu and Spanish (cf. section 2). Hence, Spanish-Dutch and Papiamentu-Dutch code-switching between the noun and the adjective could result in four potential noun-adjective combinations (Pap/Span N Dutch Adj, Pap/Span Adj Dutch N, Dutch N Pap/Span Adj, and Dutch Adj Pap/Span N), so the question that arises is whether they are all possible or whether some combinations are disallowed in the bilingual grammars of these speakers. Due to the generally low occurrence of attributive adjectives in production data (cf. Parafita Couto and Gullberg, 2019), several studies attempted to unveil the constraints that predict code-switching patterns at this conflict site in different bilingual populations (Spanish-English, Welsh-English, and Papiamentu-Dutch) using different methodologies (Parafita Couto et al., 2015, 2017a,b; Voss, 2018; Pablos et al., 2019; Stadthagen-González et al., 2019; Vaughan-Evans et al., 2020, i.a.). Most of these studies evaluated the predictions of two theoretical accounts: the Matrix Language Framework (MLF; Myers-Scotton, 1993) and the Minimalist Program approach (MP; Cantone and MacSwan, 2009), although no clear evidence to favor one model over the other was found. However, these studies provided valuable insight into a general preference for noun-insertions over adjective insertions (cf. Vaughan-Evans et al., 2020 for a detailed overview). In the next section,

we present a brief description of Papiamentu-Dutch and Spanish-Dutch bilingualism and word order.

2. Papiamentu-Dutch and Spanish-Dutch bilingualism

2.1. The Papiamentu and Spanish-speaking communities in the Netherlands

Papiamentu is a Portuguese-based creole (re)lexified by Spanish (Jacobs, 2012) spoken in Aruba, Bonaire, and Curaçao (known as the ABC islands, the Caribbean), where it is an official language together with Dutch and English. It is the first language of more than 80% of the population (Kester and Fun, 2012; Jacobs and Muysken, 2019). Papiamentu is also spoken by a large part of the 161,265 Antillean migrants who live in the (European) Netherlands [Central Bureau of Statistics (CBS), 2019],¹ a diverse community ranging from "well-established long-term residents of Antillean origin, students, and young people with little chance of employment and living in poor conditions" (Jacobs and Muysken, 2019). The ABC islands are part of the Kingdom of the Netherlands and have thus been in close contact with Dutch for over three and a half centuries. Because of the extensive historical contact with Dutch and because of wide-spread bilingualism in the country of origin, Papiamentu in the Netherlands has been described as post-colonial HL, in a similar situation as Hindi in the United Kingdom (Jacobs and Muysken, 2019). Several studies point to the fact that, despite the importance of Dutch in everyday life, Papiamentu dominance can still be found in bilingual populations residing in the Netherlands (Pablos et al., 2019; Suurmeijer et al., 2020), and their attitudes to their HL are positive (Kester and Hortencia 2010; Kester and Fun, 2012). Perhaps related to this, the most common code-switching pattern observed in the available data seems to be that Papiamentu is the matrix language and Dutch elements—often nouns—are inserted (Muysken et al., 1996; Parafita Couto and Gullberg, 2019).

According to data from the Central Bureau of Statistics (CBS) in 2019, a total of 130,160 people living in the Netherlands come from Spanish-speaking countries. About a third of the Spanish-speaking population comes from Spain, and many of these migrated to the Netherlands in the 1960 and 1970s as contracted workers. The rest came from a range of Spanish speaking countries in Latin America, where dictatorships and civil wars caused a wave of political refugees during the 1970 and 1980s. More Spanish-speaking people migrated to the Netherlands during the 1990s (mostly from Colombia and the Dominican Republic; van Suchtelen, 2016). In the Netherlands, we do not find tight-knit Spanish-speaking communities such as the ones that exist in certain areas in the United States. People tend to live dispersed across the country, and there is relatively little cohesion among its members (van Osch, 2019). Spanish speakers in the Netherlands are appreciated for their linguistic repertoire, as Spanish enjoys a relatively high prestige (van Osch, 2019). From personal communication with

¹ The Dutch Central Bureau of Statistics (CBS) reports this number for migrants from the Dutch Antilles, which also include the English-speaking islands of Sint-Maarten, Sint-Eustatius and Saba.

several Spanish heritage speakers who participated in the present as well as other studies, we know that many of them only speak Spanish with their direct family members. Therefore, we may even contend that there is no such thing as a “Spanish-speaking community” in the Netherlands, since the word community in and of itself implies membership of a group that has certain characteristics shared between all members, as well as close connections between those individual members.

We do not know of any studies that have investigated code-switching habits for this particular population. Therefore, we do not know whether there are any directionality asymmetries such as those that have been attested for the Papiamentu-speaking community.

2.2. Word order in Dutch, Spanish, and Papiamentu

Spanish and Papiamentu are different from Dutch when it comes to noun-adjective word order. While Dutch requires a pre-nominal position of the adjective (Broekhuis, 2013), as shown in (1), Spanish and Papiamentu use post-nominal adjectives, as shown in (2) and (3), even though pre-nominal adjectives are sometimes accepted in both languages (see Kouwenberg and Muysken, 1994 and Castillo, 2022 for Papiamentu and García-Bayonas, 2006 for Spanish).

Dutch

- (1) een zwarte hamer
a black hammer
“a black hammer”

Spanish

- (2) un martillo negro
a hammer black
“a black hammer”

Papiamentu

- (3) un martin pretu
a hammer black
“a black hammer”

In Spanish, the placement of a number of adjectives with respect to the noun varies depending on the semantic interpretation of the adjective, see examples in (4a) and (4b). Certain adjectives tend to be placed before the noun, such as *gran* (great) and *buen* (good), or can only appear before the noun, such as *mero* (mere). Most adjectives however tend to be placed after the noun, and some are strictly ungrammatical in prenominal position, such as adjectives which indicate nationalities or—important to this study—colors (2).

- (4) a. un hombre pobre
a man poor
“a poor (poverty-stricken) man”
b. un pobre hombre
a poor man
“a poor (piteous) man”

Papiamentu adjectives behave similarly to Spanish ones, and may appear prenominal, which then changes its meaning (Sledge, 2011),

“encoding a non-restrictive meaning that departs from the regular denotation” (Castillo, 2022, p. 53). Examples (9a) and (9b) demonstrate how the semantic interpretation of an NP differs with different noun-adjective word orders in Papiamentu (just as in 4a and 4b).

- (5) a. homber pober
man poor
“poor (poverty-stricken) man”
b. pober homber
poor man
“poor (piteous) man” (Parafita Couto et al., 2017a,b, p. 162)

The stimuli for the current study, however, were designed to elicit color adjectives, which leave no room for interpretation and are always postnominal in both Spanish and Papiamentu, and prenominal in Dutch.

3. Previous literature on word order in code-switching

3.1. Grammatical constraints

Poplack (1980) proposed the equivalence constraint, which states that “[c]ode-switches will tend to occur at points in discourse where juxtaposition of L1 and L2 elements does not violate a syntactic rule of either language, i.e., at points around which the surface structure of the two languages map onto each other” (p. 586). This implies that code-switching conflict sites should not happen, yet examples from spontaneous conversational data show that they do, as illustrated by Parafita Couto and Gullberg (2019) for Papiamentu-Dutch. In the example *un dushi verblijf* “a nice stay,” for example, the Papiamentu adjective “*dushi*” precedes the Dutch noun “*verblijf*,” contrary to what would be expected in unilingual Papiamentu constituent order (Parafita Couto and Gullberg, 2019). Below we provide a brief overview of the predictions of some theoretical models to account for such switches.

According to the Matrix Language Framework (MLF, Myers-Scotton, 1993, 2002), there is an asymmetry between the two languages in code-switched discourse, distinguishing between the ‘matrix language’ (ML), which provides the morphosyntactic frame for the clause, and the ‘embedded language’ (EL), which provides embedded elements. The MLF predicts that both finite verb morphology and word order within a clause will be sourced from the same language (the ML). As such, if the finite verb morphology is from language A, then the prediction would for the relative word order within the adjective-noun phrase to also be from language A.

Another approach, which is grounded in the Minimalist Program (MP), assumes that the features of the lexical items should account for CS/bilingual grammars (MacSwan, 1999). Thus, code-switching data should be explained in the same way we explain monolingual grammars. Regarding adjective-noun order, Cantone and MacSwan (2009) follow proposal of Cinque (1994, 1999, 2005) that adjectives universally precede nouns and that the postnominal position of the adjective in languages like Spanish and Papiamentu

follows from overt movement of the noun to a position to the left of the adjective, due to a strong EPP feature in Agr in those languages. Thus, they arrive at the descriptive generalization that “while the data remain slightly ambiguous, a relatively clear pattern has emerged in both the survey data and the naturalistic data confirming the general view of previous researchers, namely, that the word order requirements of the language of the adjective determine word order in code-switching in DP-internal contexts” (Cantone and MacSwan, 2009, pp. 266–267). Therefore, the language of the adjective, irrespective of the matrix language, is expected to determine the adjective’s position in code-switched phrases (Cinque, 2005; Cantone and MacSwan, 2009). However, Cantone and MacSwan (2009) did not control for the Matrix Language of the clause, so it is not clear whether these examples could also be explained by the MLF.

Several studies have tried to differentiate between these two models, but no clear conclusion can be drawn (cf. Parafita Couto et al., 2021 for an overview). For instance, for the specific case of Papiamentu-Dutch mixed nominal constructions, study of Pablos et al. (2019) used event-related brain potentials (ERPs) to measure online comprehension of adjective-noun switching, but leading to null results when trying to disentangle the predictions of the different theoretical models. Similarly, Voss (2018) used comparative judgments and showed that neither of the two theoretical models could fully account for the acceptability of Papiamentu-Dutch adjective-noun switches.

3.2. Extra-linguistic factors

Whereas previous studies on word order have mainly focused on comparing MLF and MP predictions, the current study takes a different approach, which leaves more room for extra-linguistic variables both at the individual level and at the level of the community (cf. Parafita Couto and Gullberg, 2019).

Variation at the individual level has been observed by Boers et al. (2020) and van Osch et al. (2022), who demonstrate that differences between speakers with respect to gender agreement strategies in code-switching are related to differences in dominance, in terms of proficiency, use and exposure. Similarly, Licerias et al. (2008) and Munarriz-Ibarrola et al. (2022) report differences in code-switching patterns between groups of bilinguals that seem to be related to the order of acquisition of the languages in the particular bilingual group.

There is also evidence from a usage-based perspective that suggests that code-switching patterns emerge through their increased use and subsequent entrenchment and such patterns can be community-specific (Backus, 2015; Valdés and Jorge, 2016; Blokzijl et al., 2017; Balam et al., 2020, 2022). It has been demonstrated that community-specific norms exist in certain parts of code-switching grammars, and that bilingual communities of the same language pair do not necessarily converge onto the same code-switching structures (e.g., Balam et al. (2020) for code-switched verbal constructions in Spanish-English bilingual communities or Królikowska et al. (2019) for gender assignment to English noun insertions in different Spanish-English communities). Such norms may depend on the frequency of code-switching within the community (Królikowska et al., 2019). It is hence

expected that cross-community variation may also affect environments about which the MP or MLF make predictions, such as adjective position, though these models do not account for this type of variation.

An interesting case of cross-community variation that may be relevant to the present topic of investigation concerns code-switching directionality or choice of matrix language. Several studies presenting natural production data show that, within specific communities, speakers tend to converge on one matrix or base language, inserting elements from the other language (e.g., Welsh for Welsh-English in northern Wales, Spanish for Spanish-English in Miami, English Creole for English Creole-Spanish in Nicaragua, Frisian for Frisian-Dutch in the Netherlands, cf. Breuker, 2001; Blokzijl et al., 2017; Bosma and Blom, 2019). As mentioned in section 2, a similar asymmetry has also been reported for Papiamentu, such that it is more common to insert Dutch elements (such as nouns) into Papiamentu, than vice versa (Muysken et al., 1996; Parafita Couto and Gullberg, 2019).

What determines the choice of matrix language is not clear, but previous research indicates that extralinguistic factors such as language prestige play a role (Blokzijl et al., 2017; Parafita Couto and Gullberg, 2019), suggesting that the language with the higher social status is the one that is inserted into the other (matrix) language. These findings highlight the extent to which code-switching practices are embedded in the sociocultural and sociohistorical experiences of the bilingual speakers (cf. Suurmeijer et al., 2020) and raise the question of whether exposure to asymmetries in the choice of matrix language or directionality of switching within the community would determine how speakers tackle code-switches at conflict sites such as the one reported on in the present paper. This issue is discussed by Vaughan-Evans et al. (2020), who looked at the relative order of adjectives and nouns in switched nominal constructions Welsh-English by means of an electrophysiological study. They observe stronger expectations about the placement of the code-switch when the ML is Welsh, than when the ML is English, which they attribute to the fact that in this particular community, English insertions into Welsh are considerably more common than vice versa. They argue that this finding could also explain some of the conflicting patterns observed in previous electrophysiological studies (Parafita Couto et al., 2017a,b on Welsh-English and Pablos et al., 2019 on Papiamentu-Dutch), which did not consider the frequency of the ML of the sentence as a confounding factor within their experimental design and analyses.

Finally, some studies have observed differences in code-switching patterns between child and adult bilinguals of the same language combination. For instance, Urbaneja (2020) showed that Spanish-English child bilinguals produced more English determiners than adult bilinguals, although not from the same community. Similarly, longitudinal study of Vihman (2018) two English-Estonian bilingual children (aged 2;10–7;2 and 6;6–11;0) shows the importance of considering age as a factor affecting code-switching patterns, as the grammar of the children in the study contains a lot of variation. They have not yet fully acquired adult grammar and therefore do not conform to the constraints of the MLF model, like adult bilinguals. This suggests that, as is the case for the development of unilingual grammars in language acquisition, children’s code-switching patterns and strategies may exhibit more flexibility and take time to converge onto adult-like

norms. On the other hand, [Balam et al. \(2021\)](#) and [Phillips and Deuchar \(2021\)](#), who compared children and adults from the same community with respect to gender and choice of the matrix language respectively, do not observe any differences between the different age groups in their studies. The children in the study of [Phillips and Deuchar \(2021\)](#) were aged between 1;9 and 2;6, leading them to conclude that the code-switching patterns in the linguistic input in the community begin to be reproduced in child productions from a very young age.

4. Research questions

In the present study, we focus on adjective-noun code-switched constructions in Spanish-Dutch and Papiamentu-Dutch bilinguals, and we aim to unveil the factors that determine which word order is preferred by heritage speakers from these languages. To this end, the following research questions were formulated:

1. Which linguistic factors (e.g., the ML and the language of the adjective) determine word order preferences?
2. What is the role of extralinguistic factors, both at the individual level and at the community level, in accounting for heritage speakers' preferences in code-switched speech?

5. Materials and methods

5.1. Participants

A total of 36 heritage speakers living in the Netherlands participated in this study. We would like to note that we use the term heritage speaker, even though not all participants are considered as such under all definitions, for example because they arrived in the Netherlands well after the onset of school. However, given that age of onset was one of our variables of interest, it was considered important that our sample included a wide range of ages of onset. Of the 21 Spanish heritage speakers, 11 participants were born in the Netherlands (two of whom spent a few years of their lives in another Spanish-speaking country later in childhood), four arrived in the Netherlands before starting their primary education, and the remaining six arrived in the Netherlands between the ages of 6 and 12. Of the 15 Papiamentu heritage speakers, three were born in the Netherlands, two arrived before going to primary school, and the remaining 10 arrived when they were between 6 and 21 years old. However, it must be noted that all Papiamentu-speaking participants were exposed to Dutch to a certain extent before arriving in the Netherlands, given that Dutch is an official language in Aruba and Curaçao, where all participants were from. As mentioned, our participants varied considerably regarding their ages at testing (8–54). The participants can be divided into three age groups: children (age 7–12, $n = 12$), teenagers (age 13–18, $n = 7$), and adults ($n = 17$). The Spanish-speaking participants had backgrounds from a range of Spanish speaking countries, such as Argentina, Colombia, Ecuador, Mexico, Paraguay, Peru, and Spain among others. The Papiamentu heritage speakers all came from or have a family background in Curaçao and Aruba. The background questionnaire contained several questions about the participants' patterns of use and exposure to both languages. They were asked to report their usage of Dutch and of the HL, both with immediate family and non-immediate

TABLE 1 Socio-linguistic information about the participants.

	Spanish (N=21)	Papiamentu (N=15)
Age at testing	M: 17,19	M: 27,27
	Range: 8–52	Range: 9–54
Age of arrival	M: 3,23	M: 8,26
	Range: 0–12	Range: 0–21
Length of residence	M: 13,62	M: 18,87
	Range: 4–37	Range: 3–42
Self-reported skill across domains in the HL (0–3)	M: 2,42	M: 2,13
	Range: 1–3	Range: 0.5–3
Heritage language usage immediate family	M: 47,61%	M: 49,35%
	Range: 10–100	Range: 9–100
Heritage language usage non-immediate family	M: 27,43%	M: 23,31%
	Range: 0–91	Range: 0–90
Usage of Dutch immediate family	M: 48,82%	M: 44,98%
	Range: 10–100	Range: 9–100
Usage of Dutch non-immediate family	M: 68,66%	M: 71,73%
	Range: 5–100	Range: 40–100
Other input in HL (hours a week)	M: 12,1	M: 8,18
	Range: 0–33	Range: 0–67
Heritage language classes (yes/no)	May-21	0/15
(Children) Current input HL	M: 54,63%	M: 45,6%
	Range: 40–100	Range: 24–85
(Children) Current input Dutch	M: 38,44%	M: 43,10%
	Range: 0–55	Range: 15–75
(Children) Previous input HL (0–4 years old)	M: 60,34%	M: 71,9%
	Range: 47,5–100	Range: 46,5–80
(Children) Previous input Dutch (0–4 years old)	M: 33,13%	M: 19,6%
	Range: 0–52,5	Range: 11–25

family, the number of hours per week they received other input (which refers to media such as music, books, television, and social media) in their HL, the frequency with which they visited their country of origin [on a scale from 1 (never) to 4 (once or multiple times a year)], their self-reported skill in their HL (on a scale from 0 to 3 for reading, writing, speaking, and listening separately), and whether they had received any official classes/courses in their heritage language while living in the Netherlands. The questionnaire for the children also contained questions about current input and input in their heritage language at age 0–4. This information is summarized in [Table 1](#) below.

5.2. Materials

The participants completed a Director-Matcher task (*cf.* [Gullberg et al., 2009](#)), a method used to elicit nominal constructions consisting of a determiner, noun and adjective (e.g., “above the green painting is a blue lamp”). This task, which has been used before by [Bellamy et al. \(2018\)](#) and [Munarriz Ibarrola et al. \(2022\)](#), consists of a board game involving two people; the director and the matcher. The participants sit

across from each other with a cardboard box dividing them, so that they cannot see each other's board. Both participants have a set of cards laid out depicting different objects in different colors. The goal is for the director to communicate to the matcher where to put the cards, describing the images on each card. If the game is played correctly, both the director and the matcher end up having their cards in the same order on their boards. Both the director and the matcher were given the same set of 30 cards depicting 15 different highly frequent objects (a house, a hat, a bed, etc.) in four different colors: red, white, black and green.

As described in the participants' section, the background questionnaire was mainly aimed at participants' current use and exposure to both languages, their education in the HL, and their self-rated proficiency in the HL. The background questionnaire for participants under the age of 12 was filled out by the parents, and also contained a part on the age, education and language use of the father and mother, as well as questions about previous input.

Both the materials and the language background questionnaire can be found on: https://osf.io/3srzv/?view_only=a38aceb650a04dbd8eeff1c84ea867c0

5.3. Procedure

The participants completed the task four times in total. Examples 6–9 show samples in the four modes of a Spanish HS. The order of administration was as follows: the first two modes elicited nominal constructions in the two languages in unilingual mode [first the HL, then Dutch—examples (6) and (7)], in order to check whether the participants were able to use the target word order in each of their languages. Immediately after the unilingual modes, they carried out the same task in two different code-switching modes. First, they were instructed to complete the task in their HL again, but this time they were asked to name just the object in Dutch (8). Finally, the participants were instructed to use Dutch, and name the object in the heritage language (9). This order was chosen for two reasons. Based on the assumption that our participants were not likely to highly frequent code-switchers, we considered that it would be easier for them to understand the task if they started with the unilingual mode. Moreover, maintaining the same order for every participant allowed us to analyze observed differences between groups without having to take into account any potential effect of order.

- (6) Arriba de la casa roja está el libro blanco
Above de house red is the book white
“Above the red house is the white book”
- (7) Naast de zwarte kam ligt de groene hoed
Next to the black comb is the green hat
“Next to the black comb is the green hat”
- (8) A la derecha de la **bloem** blanca está el **boek** verde²
To the right of the flower white is the book green
“To the right of the white flower is the green book”

² In this and all other examples containing code-switching, the matrix language is in italics and the inserted element in bold font.

- (9) *Onder het zwarte **casa** is de rode **flor***
Underneath the black house is the red flower
“Underneath the black house is the red flower”

At the beginning of the procedure, the participants were asked in which language they would like to receive instructions, the questionnaire, and consent forms, in Dutch or in their heritage language. The participants (or their parents in the case of child participants) first signed a consent form. After this, they completed the task while being given precise instructions. Only after they had completed the first round of the task in the heritage language were they told to do the next round in Dutch, and so on. After having completed all four rounds of the task, the participants (and/or parents) were asked to fill out the background questionnaire.

6. Analysis and results

In the analysis presented below, we only included those instances where an adjective was produced either directly preceding or following the noun. Those cases that lacked an adjective ($n = 39$) or where the adjective was part of a relative clause construction ($n = 29$; *een hoed que es verde*—“a hat that is green”) were excluded.

6.1. Unilingual mode

Table 2 shows the frequencies of the produced word orders by both groups combined in the unilingual modes. In the Dutch mode, participants produced almost exclusively adjective-noun word order, except for five instances of noun-adjective order, four of which were produced by the same participant, a Spanish heritage speaker. In the unilingual HL mode, there were 27 occurrences of adjective noun orders, 23 of which were produced by the same participant, a Papiamentu HS.

6.2. Code-switching mode

In code-switching mode, HSs tended to adhere to the word order from the experimental mode they were in, that is: they used prenominal adjectives more when they were instructed to speak Dutch with nouns inserted from the HL and they produced postnominal adjectives more when they had to insert Dutch nouns into their respective heritage languages (see Table 3). However, there is variation: in the Dutch mode with HL insertions, 278 (24,11%) of all inserted nouns have a postnominal adjective, and in the HL mode with Dutch insertions, adjective-noun order was used 130 (11,31%) times.

6.2.1. Linguistic variables

In this section, we ask to what extent this variability can be explained by linguistic factors. In this part of the analysis,

TABLE 2 Produced word order in the unilingual experimental modes.

	Adjective-Noun	Noun-Adjective
Dutch mode	1,186	5
HL mode	27	1,108

TABLE 3 Produced word order in the code-switching experimental modes.

	Adjective-Noun	Noun-Adjective
Dutch mode with HL insertions	875	278
HL mode with Dutch insertions	130	1,019

TABLE 4 Production of word orders by matrix language and adjective language.

Matrix language	Language adjective	Adjective-Noun	Noun-Adjective
Dutch	Dutch	607	152
	Spanish/Papiamentu	0	22
Spanish/Papiamentu	Spanish/Papiamentu	38	712
	Dutch	7	36

we collapse the data for the two heritage groups, given that Spanish and Papiamentu behave similarly when it comes to word order in the nominal domain. In the introduction, two linguistic variables were mentioned that have been proposed to account for word order constraints in code-switching: the matrix language and the language of the adjective. While identifying the language of the adjective is straight-forward, the same is not true when it comes to determining the matrix language. Even though the participants were instructed to speak one language and embed nouns from the other language, it is not guaranteed that they in fact consistently follow these instructions. A potential solution to this problem is to determine the matrix language for each clause based on the language of the verb (cf. Herring et al., 2010; Blokzijl et al., 2017; Urbaneja, 2020). However, in our dataset, only 51.6% of utterances included a verb. Of the sentences that lacked a verb, there were sometimes other elements, such as adverbs and/or conjunctions (*en daarnaast weer een zwarte casa*—“and next to that again a black house”). In 99% of these cases, the languages of the verb or these other elements coincided with the language of the experimental mode. Based on this information, it was considered safe to assume that the language of the verb and/or other elements in the sentence could be used as an indicator for the matrix language.

A total of 711 instances that consisted of *only* noun phrases were excluded, leaving us with 1,574 instances. Table 4 presents the word orders produced for these 1,574 cases, by matrix language and adjective language.

What immediately becomes clear from this table, is that the matrix language almost always coincides with the language of the adjective (1,509 out of 1,574–95.9%). This could be due to the nature of our task: participants were explicitly instructed to name *only* the object in the other language. This led to a high number of noun insertions [example (10); $n = 1,441$].

When the matrix language and the adjective were Dutch, adjective-noun (the Dutch word order) was used more often (607 out of 759 cases), whereas when the matrix language and the adjective were Spanish/Papiamentu, noun-adjective (the Spanish/Papiamentu word order) was preferred (712 out of 750 cases). For the few cases where the language of the adjective did not coincide with the matrix language (65 in total), we see a general preference for noun-adjective word order, which sometimes aligned with the matrix language

($n = 36$) and other times language of the adjective ($n = 22$). These data thus suggest that either the matrix language or the language of the adjective, or both, seem to play a role in determining word order in code-switched productions. However, the data cannot help us disentangle between these two factors. Moreover, even when both the matrix language and the language of the adjective align, there is still variation, which suggests there may be other factors playing a role.

Taking a closer look at our data, we noticed that the type of insertion mattered. In addition to the 1,441 noun insertions (example 10), there were also 66 determiner-noun insertions (example 11), 18 adjective insertions (example 12), 30 noun + adjective insertions (example 13), and 18 det + noun + adjective insertions (example 14).³

(10) “*El kam negro está arriba*” (Spanish ML, Dutch insertion)

The comb black is above
“The black comb is above”

(11) “*Después es de bloem negro*” (Spanish ML, Dutch insertion)

Next is the flower black
“Next is the black flower”

(12) “*...en een bloem blanku*” (Dutch ML, Papiamentu insertion)

and a flower white
“... and a white flower”

(13) “*...met daaronder een kama pretu*”

(Dutch ML, Papiamentu insertion)

With underneath a bed black
“...with underneath a black bed”

(14) “*Daarna un llave rojo*” (Dutch ML, Spanish insertion)

after that a key red
“After that, a red key”

Table 5 below shows the word order preference for each type of insertion that was observed in the dataset.

What becomes clear from Table 5 is that, apart from noun insertions, all other types of insertions seem to favor noun-adjective order, regardless of the ML.

To see whether any of these effects was statistically significant, we ran a series of linear mixed effects regression models, using the lme4 package in R (R Core Team, 2021). The dependent variable was word order (adjective-noun vs. noun-adjective). Our three predictor variables of interest were matrix language (Dutch vs. HL), adjective language (Dutch vs. HL), and insertion type (noun insertion vs. other insertion), which were all sum-coded. It was problematic to include all three independent variables in a single analysis, for two reasons. First, as explained above, there was a considerable overlap between the matrix language and the language of the adjective: these two factors overlapped for 96% of the data. In addition, the third variable, insertion type, is partially derived from the other two variables, because if the ML and the language of the adjective do not coincide, this automatically implies that the insertion

³ Some participants used non-standard gender agreement in unilingual Spanish or Dutch utterances, which is discussed in both Boers et al., 2020 and van Osch et al., 2022. Gender assignment in code-switched nominal constructions is discussed in these publications as well.

TABLE 5 Production of word order by matrix language and type of insertion.

Matrix Language	Type of insertion	Adjective-Noun	Noun-Adjective
Dutch	Noun	605 (<i>het rode casa</i>)	115 (<i>het casa rood</i>)
	Determiner + noun	0 (<i>la rode casa</i>)	37 (<i>la casa rood</i>)
	Determiner + noun + adjective	0 (<i>la roja casa</i>)	4 (<i>la casa roja</i>)
	Noun + adjective	0 (<i>het roja casa</i>)	7 (<i>het casa roja</i>)
	Adjective	5 (<i>het roja huis</i>)	2 (<i>het huis roja</i>)
Spanish/Papiamentu	Noun	38 (<i>la roja huis</i>)	683 (<i>la huis roja</i>)
	Determiner + noun	0 (<i>het roja huis</i>)	29 (<i>het huis roja</i>)
	Determiner + noun + adjective	1 (<i>het rode huis</i>)	13 (<i>het huis rood</i>)
	Noun + adjective	2 (<i>la rode huis</i>)	21 (<i>la huis rood</i>)
	Adjective	0 (<i>la rode casa</i>)	11 (<i>la casa rood</i>)

Bold type is used to indicate the inserted elements, while italics are used to indicate the matrix language.

TABLE 6 Output for the final model including linguistic variables.

Predictors	Word_order				
	Estimate	Std. error	CI	Statistic	p
(Intercept)	4.35	1.57	1.28–7.42	2.77	0.006
ML based verb or other elements SPAPAP merged 1	9.40	3.17	3.20–15.61	2.97	0.003
Noun vs. other based on verb other elements 1	9.66	2.74	4.28–15.04	3.52	<0.001
ML based verb or other elements SPAPAP merged 1 × noun vs. other based on verb other elements 1	–22.60	5.48	–33.34––11.86	–4.12	<0.001
Random effects					
σ^2	3.29				
τ_{00} Subject	74.33				
τ_{11} Subject.ML_based_verb_or_otherelements_SPAPAPmerged1	259.00				
ρ_{01} Subject	–0.43				
ICC	0.98				
N Subject	31				
Observations	1,574				
Marginal R^2 /Conditional R^2	0.426/0.987				

Statistically significant p -values are indicated in bold type.

contains at least the adjective, whereas if they do coincide, the insertion can only contain the determiner and/or the noun, but not the adjective. To avoid issues with multicollinearity, we therefore decided to first run three models for each of the three variables separately, and check which of the variables explained the most variance. Each of these models was compared to a null model, i.e., a model only containing the intercept and the random intercept for subject. All three variables improved the model fit significantly, but the model including matrix language showed the most improvement, in terms of both the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In the next step we added first the main effect of insertion type and then the interaction between the two variables, and both improved the model significantly.

The final model (Table 6), which also included the random slope for matrix language (the model did not reach convergence when we added the slope for the interaction), showed strong and significant effects for matrix language ($\beta = 9.4$, $SE = 3.17$, $z = 2.97$, $p = 0.003$), insertion type ($\beta = 9.66$, $SE = 2.74$, $z = 3.52$, $p < 0.001$), and the interaction between these two variables ($\beta = -22.60$, $SE = 5.48$, $z = -4.12$, $p < 0.001$), which confirmed the observation that the word

order is determined by the Matrix language in the case of noun insertions, but not for all other types of insertions, in which case noun-adjective is the preferred word order overall (Figure 1).⁴

6.2.2. Extra-linguistic variables

In addition to the linguistic variables discussed in the previous section, we were also interested to what extent extra-linguistic variables played a role in determining word order variation. This is

⁴ Finally, there was an interesting relation between word order and gender in Dutch, which is discussed in van Osch et al. (2022). In Dutch, attributive adjectives are inflected for common nouns and uninflected for neuter nouns, but when used predicatively, the adjective is uninflected for both common and neuter gender. van Osch et al. (2022), which is based on the same dataset as the present paper, but focusing on gender, show that whenever the Dutch adjective is placed after the noun, it is uninflected, which seems to suggest that it may be used as a predicative adjective rather than an attributive one.

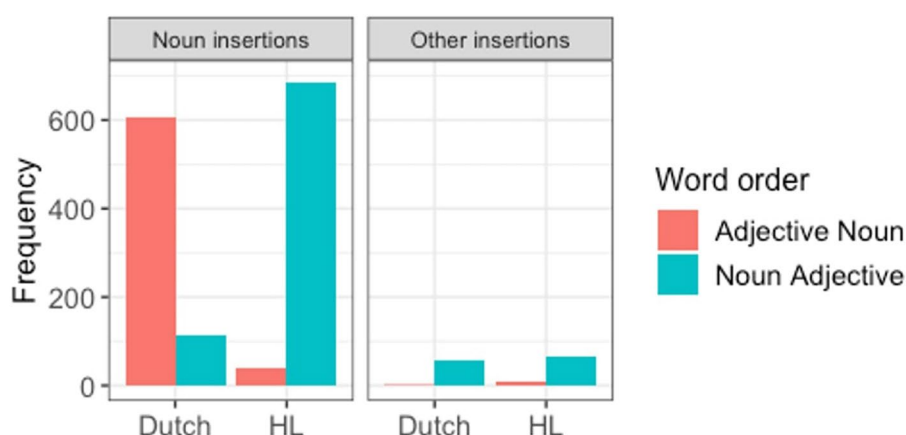


FIGURE 1
Production of word order in code-switching mode by ML and insertion type.

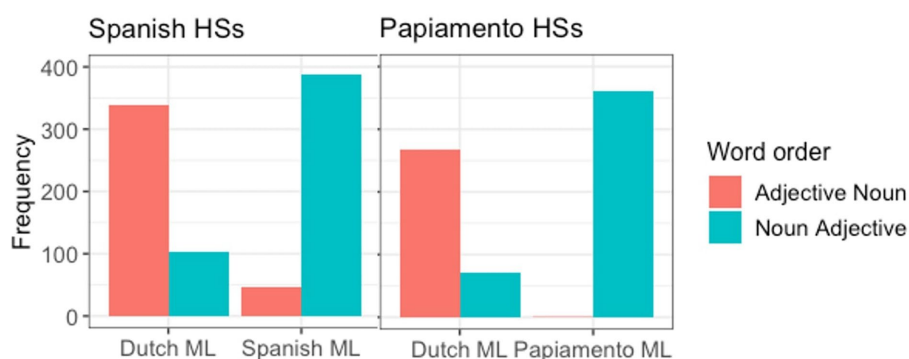


FIGURE 2
Production of word order by ML, separated between groups.

why we collected data from two different communities of heritage speakers in the Netherlands, and we also included a wide range of speakers of different ages, different lengths of residence, etc. In this section, we focus on the code-switching data, because both groups were very categorical in the unilingual modes.

First, we compare the two communities to each other (Figure 2). While the Spanish HSs show variation in terms of their word order preferences both when the matrix language is Dutch and when it is Spanish, the Papiamento speakers very categorically choose noun-adjective when Papiamento is the matrix language.

In addition to the difference between these two communities, a large part of the observed variation was found to derive from individual variation between subjects. This is illustrated in Figures 3, 4 for the Spanish group mode and the Papiamento group, respectively.

In a second analysis, we explored which socio-linguistic variables, if any, could account for the observed variation between participants. From the background questionnaire, we had gathered information about the participants concerning their age at testing, age of onset of the societal language, the length of residence in the Netherlands, the amount of use of both languages with their immediate family and in other contexts, the amount of “other” exposure to their HL through

TV, music, reading and social media, and their self-rated proficiency in their HL (averaged across four domains; reading, writing, listening, and speaking).

We performed two analyses, one on the Dutch mode for both groups, and one on the HL mode for the Spanish group only, given that there was close to zero variation in the Papiamento group in this mode. For the analysis on the Dutch ML experimental mode, the dependent variable was word order (adjective-noun vs. noun-adjective). We considered the following predictor variables: heritage community, age at testing, age of onset of the societal language, length of residence in the Netherlands, use of Dutch with immediate family, use of Dutch with non-immediate family, total use of Dutch, average “other” exposure to the HL (i.e., through books, music, TV, and social media), whether or not they had had any instruction in their HL, and self-rated proficiency in their HL. Heritage community was a binary variable with two levels: Spanish and Papiamento. Similarly, instruction in the HL was a binary variable with two levels: yes and no. For these two binary variables, sum-coding was used. Age at testing was a categorical variable with three levels (children, teens, and adults), for which orthogonal sum-to-zero coding was used such that contrast 1 compared teens and children (+1/3 for both) to adults (−2/3) and contrast 2

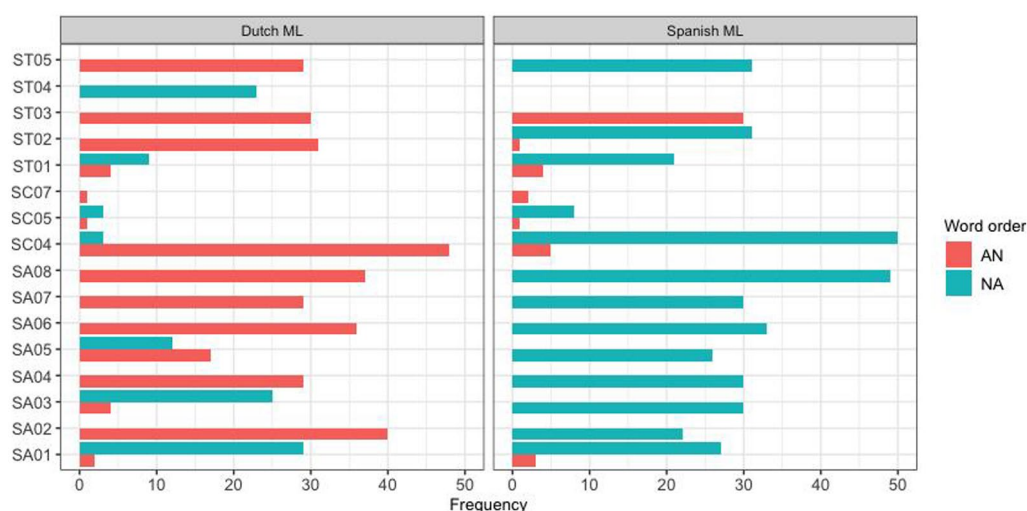


FIGURE 3
Word order production pattern for individual Spanish HSs, separated by the ML.

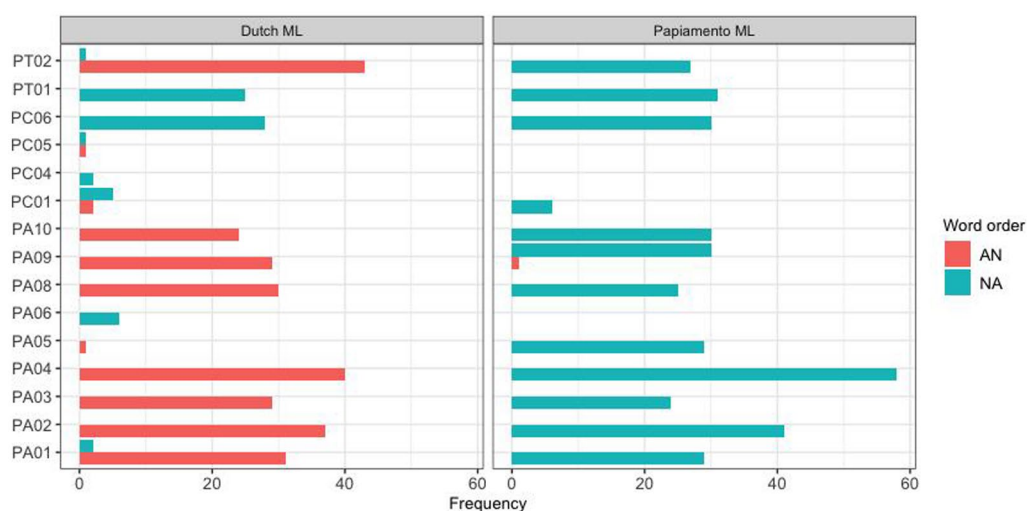


FIGURE 4
Word order production pattern for individual Papiamentu HSs, separated by the ML.

compared teens (+0.5) to children (−0.5). The remaining predictors of interest were continuous variables which were centered and standardized.

Some of these variables are inherently related. For instance, age at testing, age of onset of the societal language, and length of residence are all derived from one another. To avoid multicollinearity issues, we first checked for each of them to what extent they improved the model fit compared to a null model which just included a random intercept for subject. The variable that explained most variability was age at testing. In a similar way, it was decided to include both usage of Dutch with immediate family and with non-immediate family, but not total usage of Dutch.

We used the package *buildmer* (Voeten, 2021) for automatic model selection. The advantage of this package is that it first identifies the maximal model that converges with the variables of interest, and

subsequently uses this as a baseline for backward stepwise elimination. However, it does not check multicollinearity for each of the possible models. Therefore, to determine the degree of the correlation between predictor variables, we checked the variance inflation factors (VIF) for the final model, and eliminated several variables based on this information. The final model (Table 7) contained significant effects for age group, for the contrast between children and teens vs. adults ($\beta = 10.21$, $SE = 3.13$, $z = 3.26$, $p = 0.001$), heritage community ($\beta = -9.66$, $SE = 4.01$, $z = -2.41$, $p = 0.016$), as well as a significant interaction between these two ($\beta = -15.51$, $SE = 6.30$, $z = -2.46$, $p = 0.014$), which indicated for the Papiamentu HSs, younger participants use noun-adjective order relatively more when they insert HL nouns into Dutch, whereas for the Spanish HSs, this is not the case (Figure 5).

For the analysis on the Spanish experimental mode, the dependent variable was again word order (adjective-noun vs. noun-adjective). For

TABLE 7 Output for the final model containing extra-linguistic variables in Dutch mode with HL insertions.

Predictors	Word_order				
	Estimate	Std. error	CI	Statistic	p
(Intercept)	−3.07	1.51	−6.03 to −0.11	−2.03	0.042
HL1	−9.66	4.01	−17.51 to −1.81	−2.41	0.016
Age group1	10.21	3.13	4.08–16.34	3.26	0.001
Age group2	−1.86	4.24	−10.17 to 6.44	−0.44	0.660
HL1 × age group1	−15.51	6.30	−27.86 to −3.16	−2.46	0.014
HL1 × age group2	−3.24	8.52	−19.95 to 13.46	−0.38	0.704
Random effects					
σ^2	3.29				
τ_{00} Subject	94.76				
ICC	0.97				
N_{Subject}	31				
Observations	779				
Marginal R^2 /Conditional R^2	0.269/0.975				

Statistically significant *p*-values are indicated in bold type.

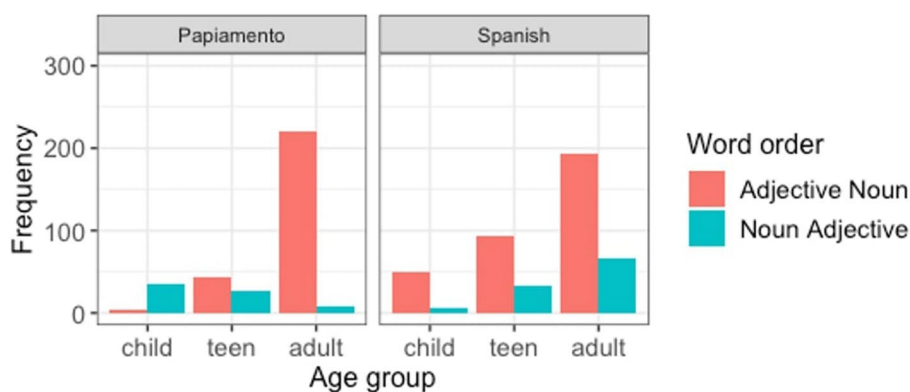


FIGURE 5

Word order preference for Dutch as a matrix language, by age group by heritage community.

the independent variables, the following were considered: age at testing, age of onset of the societal language, length of residence in the Netherlands, usage of Spanish with the immediate family, usage of Spanish with non-immediate family, total usage of Spanish, exposure to “other” exposure to Spanish (i.e., through books, music, TV, and social media), self-rated proficiency in Spanish, and whether or not they had received instruction in Spanish. Similar to the model for the Dutch experimental mode, instruction in the HL was a binary variable which was sum-coded, age at testing was a ternary variable for which orthogonal sum-to-zero coding was applied as described above, and all other variables were continuous and were centered and standardized.

Through a similar procedure as described above, age at testing was selected over age of onset and length of residence, and usage of Spanish with both immediate and non-immediate family were selected over total usage of Dutch. The final model (Table 8) contained one significant effect of age ($\beta = 8.77$, $SE = 4.34$, $z = 2.02$, $p = 0.04$), as well as a significant intercept for subject. The effect of age indicates that younger participants use the adjective-noun orders relatively

more when they insert Dutch nouns into Spanish (Figure 6, right panel). As mentioned earlier, Papiamentu speakers of all age groups categorically produced noun-adjective order while inserting Dutch nouns into their HL (Figure 6, left panel).

7. Discussion

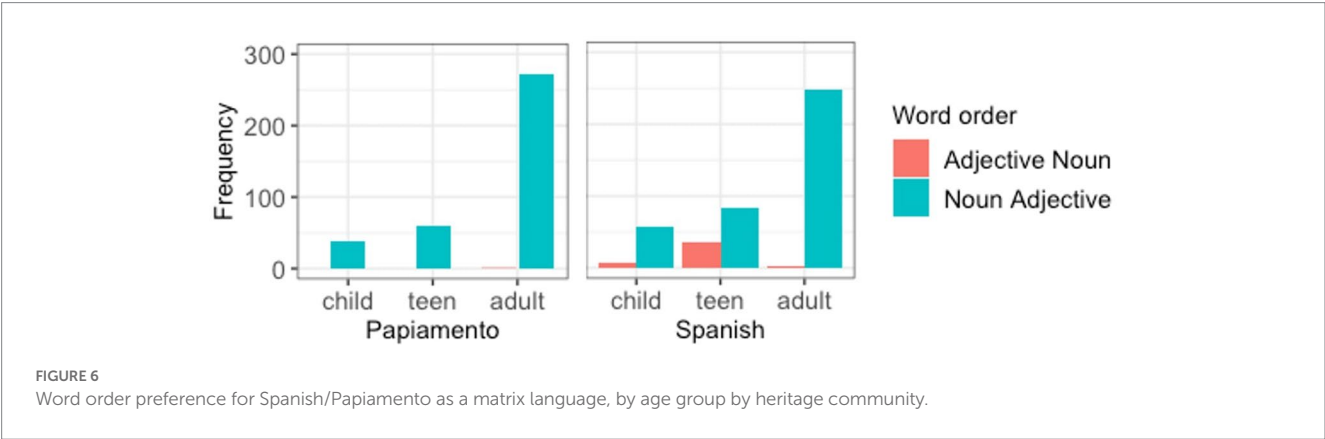
The study presented in this paper was concerned with the investigation of word order in the nominal domain in both unilingual and code-switched speech of bilingual speakers of Dutch (a language that has prenominal adjectives) and Spanish or Papiamentu (in which adjective are typically placed in the postnominal position). We observed that word order in these cases is constrained both by linguistic factors and by non-linguistic factors.

Concerning linguistic factors, similar to previous studies (Voss, 2018; Stadthagen-González et al., 2019; Vaughan-Evans et al., 2020)

TABLE 8 Output for the final model containing extra-linguistic variables in HL mode with Dutch insertions.

Predictors	Word_order				
	Estimate	Std. error	CI	Statistic	p
(Intercept)	7.98	2.79	2.50–13.45	2.86	0.004
age	8.77	4.34	0.25–17.28	2.02	0.044
Random effects					
σ^2	3.29				
τ_{00} Subject	16.47				
ICC	0.83				
N Subject	15				
Observations	434				
Marginal R^2 /Conditional R^2	0.795/0.966				

Statistically significant *p*-values are indicated in bold type.



we found effects of the matrix language and the language of the adjective. When both the matrix language and the adjective were in Dutch, the preferred order was adjective-noun, and when the matrix language and the adjective were in Spanish or Papiamentu, noun-adjective was the preferred order. These findings may indicate support for the role of the Matrix Language Framework (cf. Myers-Scotton, 2002). However, it may also be the language of the adjective (or the strength of the EPP feature in AGR, cf. Cantone and MacSwan, 2009) that is responsible for the patterns we observe. It is worth noting, however, that almost all switches that adhered to the predictions of both the Matrix Language Frame and the MP included a noun insertion (which are frequent in naturalistic production, Muysken et al., 1996; Parafita Couto and Gullberg, 2019). Like previous studies (Parafita Couto et al., 2015; Voss, 2018; Pablos et al., 2019; Parafita Couto and Gullberg, 2019; Stadthagen-González et al., 2019 among others), our data do not allow us to differentiate between the matrix language and the language of the adjective. Most of the data either are in line with the predictions of both these theories, or they contradict both theories, or they support either of the two. Nonetheless, a novel finding in the present study is the relation between the type of insertion and word order. We noted that noun insertions behaved differently from all other types of insertions, that is: for noun insertions, the above mentioned effects of the matrix language and/or the language of the adjective apply, but for all other types of

insertions, noun-adjective was the preferred option across the board. How can we explain this?

Let us start with the second most frequently produced type of insertion after noun insertions: determiner-noun insertions, illustrated in example 15 for Spanish with a Dutch insertion (repeated here) and 16 for Dutch with a Papiamentu insertion.

- (15) “Después es de **bloem** negro” (Spanish ML, Dutch insertion)

Next is the flower black

“Next is the black flower”
- (16) “Onder die rooie **kas**, un **kurason** wit” (Dutch ML, Papiamentu insertion)

Below that red house a heart white

“Below that red house, a white heart”

This type of insertion occurred 66 out of 1,574 times in our data (37 times for Dutch as the ML and 29 times for Spanish/Papiamentu as the ML) and in all cases, the adjective followed the noun. Note that the second example contradicts both the predictions from the MLF and the MP. We would like to suggest the preference for the postnominal adjective in these cases may be explained from the perspective of processing economy. If the adjective would precede the noun, the speaker would have to switch back and forth between languages several times: the verb

in the ML, the determiner in the inserted language, then the adjective in the inserted language and the noun in the ML again, which may not be the most economic strategy.

In addition to these determiner-noun insertions, there are some insertion types that seem to be used as specific strategies by individual speakers. For instance, one Spanish heritage speaker uses almost exclusively Dutch (det-)noun-adjective insertions, always with a postnominal adjective, as in example 17:

(17) “*Arriba del hartje wit hay un sleutel groen*”

Above the heart white there is a key green

“Above the white heart there is green key”

The same individual variation was found in other insertion types as well. For instance, postnominal adjectives with noun insertions into Dutch were dispreferred by most participants, but for some speakers this was actually the preferred option. This indicates that different participants seem to adhere to different strategies.

In part, these different strategies were related to the specific linguistic communities. For instance, Papiamento speakers of all age groups categorically produced postnominal adjectives when the ML was Papiamento, whereas the Spanish speaking participants showed variation in the same context. However, this variation mostly pertained to the younger participants; the adult Spanish speakers almost categorically preferred noun-adjective order, similarly to the Papiamento speakers. Interestingly, the reversed pattern was observed when Dutch was the matrix language: here, an age effect was observed for the Papiamento speakers, but not the Spanish speakers. While Papiamento speaking children preferred noun-adjective word order, the adults almost categorically produced prenominal adjectives. This difference between children and adults is in line with studies by Vihman (2018) and Urbaneja (2020), although the former was a case study of two children and the second did not compare children and adults from the same community. Two studies that have compared children and adults from the same community (Balam et al., 2021; Phillips and Deuchar, 2021) did not find any differences between the two age groups. This topic needs to be investigated further in future studies.

The difference between the Papiamento and the Spanish speakers in our study is most likely not related to linguistic differences between Papiamento and Spanish, given that the two languages overlap in terms of word order in the nominal domain. However, there are important sociolinguistic and sociohistorical differences between these communities that may explain their differential behavior. First, the Papiamento community in the Netherlands is bigger and more established, compared to the migrant Spanish community, in part because it has a longer history of post-colonial relationship. This may mean there is more contact between the members of the Papiamento community than between Spanish-speaking immigrants and their descendents. Second, all Papiamento HSS, even those who were born in Aruba or Curaçao had knowledge of Dutch before migration given the official stats of Dutch, contrary to Spanish HSS who were born in Spanish-speaking countries. Therefore, it is possible that language mixing is more common in the Papiamento community, and that for this reason there are clearer community norms than for our Spanish-speaking participants. In fact, we know from previous research (Muysken et al., 1996; Parafita Couto and Gullberg, 2019) that Papiamento speakers in the Netherlands have clear norms when it comes to the directionality of code-switching: they tend to use Papiamento as the matrix language and insert Dutch elements.

This may explain why, in this direction of code-switching, Papiamento-speaking children converge on the adult pattern from an early age, as they are exposed to this type of switches relatively more often and from an early age onward.⁵ The opposite direction—inserting nouns from the HL into Dutch—is less common in the Papiamento community, which may explain why children take more time to converge on the adult-like adjective-noun word order. In fact, Papiamento-Dutch bilingual children start out preferring the opposite word order—noun adjective—during childhood and, to some extent, still produce it during the teenage years. It is not until adulthood that they converge on what seems to be the target pattern in their community.

The Spanish-speaking differs from the Papiamento-speaking group in several ways. First of all, while the adult participants categorically prefer noun-adjective order when Spanish is the ML, similar to the Papiamento speakers, Spanish-speaking children and teens show more variability in this direction than their Papiamento-speaking counterparts. It may be the case that these speakers are less accustomed to code-switching in general, and as a result of this, children need more time and exposure to code-switching in the input to converge on the adult norm. The two groups also differ in the other code-switching direction: Dutch as the ML with HL words inserted. While the Papiamento speakers categorically choose adjective-noun order in this direction, all Spanish-speaking age groups, including the adults, show a considerable degree of variation. The increased variability in this code-switching direction may indicate that they are less accustomed to this direction, and therefore no clear-cut norms have been established. Given that we do not have information on the code-switching habits for our Spanish-speaking participants, these explanations remain rather speculative and need to be substantiated by further research.

In sum, our data suggest that word order variation in code-switched constructions in the nominal domain is determined by various factors, both linguistic ones (the matrix language and/or the language of the adjective, the type of insertion) and extra-linguistic ones (community and age group). Therefore, the field needs to broaden its focus and take into account all the different variables that may play a role, either by careful controlling of the materials and/or the participants, or by including many variables as potential predictors, which is the approach taken in this study. We contend that, while theories such as the MLF or the MP have been essential in our understanding of code-switching, we also need to acknowledge that any theory that focuses on purely grammatical factors probably cannot be considered an accurate reflection of what happens in reality. As our study, as well as other recent studies (cf. Parafita Couto et al., 2021) demonstrate, the reality of code-switching is too complex to reduce it to a single variable. We would like to propose that, rather than talking in terms of pure grammatical “constraints” on code-switching, we may need to talk about a set of predictors that can have different weights, and it is our challenge as researchers to identify

⁵ In addition, there are many Dutch borrowings in Papiamento (e.g., *stòfzuiger* “vacuum cleaner” <Du. *stofzuiger*). In nominal constructions with an adjective, the adjective follows the noun, i.e., the default word order in Papiamento (e.g., *e stòfzuiger pretu* “the black vacuum cleaner” <Du. *stofzuiger*), which is similar to the code-switching pattern found in our data (e.g., *e huis pretu* “the black house”). The complete convergence toward noun-adjective word order when inserting Dutch words into Papiamento may be reinforced by the large amount of Dutch borrowings in Papiamento (cf. Muysken et al., 1996).

which predictors should be included in this set and to estimate their relative weights. This aligns with the proposal of Muysken (2013) for modeling and interpreting language contact phenomena, with speakers' bilingual strategies in specific scenarios of language contact as the starting point. Muysken claims that bilingual strategies are conditioned by social factors, processing constraints of speakers' bilingual competence, and perceived language distance. As such, the different outcomes should correspond to different interactions of these strategies in bilingual speakers and their communities and more attention should be paid to the links between these strategies and factors.

Finally, we need to acknowledge that our study has some limitations that may have affected our results. Given that adjectives do not occur often in spontaneous speech (Parafita Couto and Gullberg, 2019), and even in semi-spontaneous elicited production (Parafita Couto et al., 2015 found similar patterns using a toy task), we applied a method to specifically elicit them. While this method was successful in eliciting adjectives, it may have made the task less natural. From literature on spontaneous oral production, we know that speakers usually do not frequently switch between the noun and the adjective (Parafita Couto and Gullberg, 2019). When adjectives are produced, they usually form an island with the noun, that is, the noun and adjective are inserted together. Conversely, in our data, the vast majority consisted of noun insertions or determiner-noun insertions. This is a clear consequence of the nature of our task: people were explicitly instructed to only name the object in the other language. The effect of the task on the type of insertion is important given that the type of insertion, in turn, was related to word order as well. A challenge for future studies could lie in finding the right balance between leaving the participants free to switch when they choose to, and at the same time make sure they use adjectives. Another recommendation for future work is that it is crucial to collect information about our participants' code-switching habits and their general proficiency in both their languages. This information would have been very useful to support some of the claims we make based on our data.

Another issue to consider includes priming whereby one speaker's code-switching facilitates another speaker's similar switching (Kootstra et al., 2010; Fricke and Kootstra, 2016). A recent study by Berghoff et al. (2023) focused on code-switching at points of non-shared word order across a bilingual's two languages. Their study delved into the scope of code-switching priming by investigating whether lexical repetition across target and prime, a factor known to boost structural priming, can increase code-switching at points of word order divergence. They tested Afrikaans–English bilinguals and showed that lexical repetition boosts the priming of code-switching in a non-shared word order. Their findings demonstrate that code-switching in production is therefore affected by a dynamic interplay between factors both language-internal (i.e., word order) and language-external (i.e., priming, and specifically lexical repetition).

The research outlined in the present study constitutes an attempt to keep widening the research perimeter on code-switching. Our (so far preliminary) findings call for further research to be able to establish the theoretical and empirical implications of our findings. Only after studying different and similar language combinations in different contact situations will we be able to arrive at a description of the different dimensions that characterize code-switching and unveil the factors that modulate bilingual grammars.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/supplementary material.

Ethics statement

The studies involving human participants were reviewed and approved by Complying with the Ethics Code for linguistic research at the Faculty of Humanities at Leiden University, for this study written informed consent from all participants (including the parents/guardians of the child participants) was obtained prior to their participation. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

BO: data curation, statistical analysis, and writing. BS and IB: methodology, data collection, data transcription and coding, and writing. MP: conceptualization, methodology, writing, and funding acquisition. All authors contributed to the article and approved the submitted version.

Funding

BO acknowledges that this work has received support by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101024053. MP acknowledges support from the Maria Zembrano program (funded by the European Union, #NextGenerationEU) and the Traineeship in the Humanities Program at Leiden University (<https://www.student.universiteitleiden.nl/en/vr/humanities/research-traineeship?cf=humanities&cd=latin-american-studies-ma>).

Acknowledgments

We would like to thank these students for their help at different stages of the project: Serenay Kaykaç (experiment design), Samantha Angela (data collection, transcription, and coding), Eva Šipulová (data collection and transcription), Machteld van Kooten (coding), and Tessa van der Meijden (coding). We would also like to thank Janet Grijzenhout and Deniz Tat from the Herling Lab for their feedback in early stages of this project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations,

or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Backus, A. (2015). "A usage-based approach to code-switching: the need for reconciling structure and function" in *Code-Switching Between Structural and Sociolinguistic Perspectives*. eds. G. Stell and K. Yakpo, vol. 43 (Göttingen: De Gruyter), 19–37.
- Balam, O., Parafita Couto, M. C., and Chen, M. (2021). Being in bilingual speech: an analysis of Estar 'be' constructions in Spanish/English code-switching. *J. Monoling. Biling. Speech* 3, 238–264. doi: 10.1558/jmsb.19374
- Balam, O., Parafita Couto, M. C., and Stadthagen-González, H. (2020). Bilingual verbs in three Spanish/English code-switching communities. *Int. J. Biling.* 24, 952–967. doi: 10.1177/1367006920911449
- Balam, O., Stadthagen-González, H., Rodríguez-González, E., and Parafita Couto, M. C. (2022). On the grammaticality of passivization in bilingual compound verbs. *Int. J. Biling.* doi: 10.1177/1367006922109772
- Bellamy, K., Parafita Couto, M. C., and Stadthagen-Gonzalez, H. (2018). Investigating gender assignment strategies in mixed Purepecha–Spanish nominal constructions. *Language* 3:28. doi: 10.3390/languages3030028
- Berghoff, R., Gullberg, M., and Kootstra, G. J. (2023). Structural priming of code-switches in non-shared-word-order utterances: the effect of lexical repetition. *Biling. Lang. Cogn.*, 1–14. doi: 10.1017/S1366728923000044
- Blokzijl, J., Deuchar, M., and Parafita Couto, M. C. (2017). Determiner asymmetry in mixed nominal constructions: the role of grammatical factors in data from Miami and Nicaragua. *Language* 2:20. doi: 10.3390/languages2040020
- Boers, I., Sterken, B., van Osch, B., Parafita Couto, M. C., Grijzenhout, J., and Tat, D. (2020). Gender in unilingual and mixed speech of Spanish heritage speakers in the Netherlands. *Language* 5:68. doi: 10.3390/languages5040068
- Bosma, E., and Blom, E. (2019). A code-switching asymmetry in bilingual children: code-switching from Dutch to Frisian requires more cognitive control than code-switching from Frisian to Dutch. *Int. J. Biling.* 23, 1431–1447. doi: 10.1177/1367006918798972
- Breaker, P. (2001). "West Frisian in language contact" in *Handbuch des Friesischen/ Handbook of Frisian Studies*. eds. O. Vries, V. F. Faltings, A. G. H. Walker, O. Wilts, N. Århammar and J. F. Hoekstraet al. (Tübingen: Niemeyer), 121–129.
- Broekhuis, H. (2013). *Syntax of Dutch*. Amsterdam: Amsterdam University Press.
- Cantone, K. F., and MacSwan, J. (2009). "Adjectives and word order" in *Multidisciplinary Approaches to Code Switching*. eds. L. Isurin, K. de Bot and D. Winford (Amsterdam: John Benjamins), 243–278.
- Castillo, Y. R. (2022) in *A Description of Papiamentu: A Creole Language of the Caribbean Area*. ed. P. Bakker (Leiden, Boston: Brill).
- Central Bureau of Statistics (CBS) (2019). Population; sex, age, migration background and generation. Available at: <https://opendata.cbs.nl/statline/#/CBS/en/dataset/37325eng/table?ts=1614200262031>
- Cinque, G. (1994). "On the evidence for partial N-movement in the romance DP" in *Paths Towards Universal Grammar: Studies in Honour of Richard Kayne*. eds. G. Cinque, J. Koster, J.-Y. Pollock, L. Rizzi and R. Zanuttini (Washington, DC: Georgetown University Press), 85–110.
- Cinque, G. (1999). *Adverbs and Functional Heads: A Cross-Linguistic Perspective*. Oxford: Oxford University Press.
- Cinque, G. (2005). Deriving Greenberg's universal 20 and its exceptions. *Linguist. Inq.* 36, 315–332. doi: 10.1162/0024389054396917
- Deuchar, M. (2012). "Code switching" in *The Encyclopedia of Applied Linguistics*. ed. C. A. Chapelle (Hoboken: Wiley-Blackwell).
- Fricke, M., and Kootstra, G. J. (2016). Primed codeswitching in spontaneous bilingual dialogue. *J. Mem. Lang.* 91, 181–201. doi: 10.1016/j.jml.2016.04.003
- García-Bayonas, M. (2006). "Adjective placement and noun semantics in Spanish" in *Functional Approaches to Spanish Syntax*. eds. J. Clancy Clements and J. Yoon (London: Palgrave Macmillan), 219–235.
- Gullberg, M., Indefrey, P., and Muysken, P. (2009). "Research techniques for the study of code-switching" in *The Cambridge Handbook on Linguistic Code-Switching*. eds. B. E. Bullock and J. A. Toribio (Cambridge: Cambridge University Press), 21–39.
- Herring, J. R., Margaret Deuchar, M., Parafita Couto, M. C., and Quintanilla, M. M. (2010). I saw the madre: evaluating predictions about codeswitched determiner-noun sequences using Spanish–English and Welsh–English data. *Int. J. Biling. Educ. Biling.* 13, 553–573. doi: 10.1080/13670050.2010.488286
- Jacobs, B. (2012). *Origins of a Creole: The History of Papiamentu and Its African Ties*. Boston, MA; Berlin: Walter de Gruyter.
- Jacobs, B., and Muysken, P. (2019). "Heritage languages in a post-colonial setting: focus on Papiamentu" in *Heritage Languages: A Language Contact Approach*. eds. S. Aalberse, A. Backus and P. Muysken (Amsterdam: John Benjamins Publishing Company), 204–223.
- Kester, E. P., and Hortencia, T. (2010). "Language use, language attitudes and identity among Curaçaoan high school students" in *Continuity, Divergence and Convergence in Language, Culture and Society on the ABC-Islands*. eds. N. Faracías, R. Severing, C. Weijer and E. Ehteld (Willemstad: Fundashon di Planifikashon di Idioma), 25–38.
- Kester, E. P., and Fun, J. (2012). "Language use, language attitudes and identity among Aruban students in the Netherlands" in *Multiplex Cultures and Citizenships*. eds. N. Faracías, R. Severing, C. Weijer and L. Ehteld, vol. 2012 (Willemstad: University of Curaçao/Fundashon pa Planifikashon di Idioma), 231–248.
- Kootstra, G. J., Van Hell, J. G., and Dijkstra, T. (2010). Syntactic alignment and shared word order in code-switched sentence production: evidence from bilingual monologue and dialogue. *J. Mem. Lang.* 63, 210–231. doi: 10.1016/j.jml.2010.03.006
- Kouwenberg, S., and Muysken, P. (1994). "Papiamentu" in *Pidgins and Creoles*. eds. J. Arends, P. Muysken and N. Smith (Amsterdam: John Benjamins Publishing Company), 205–218.
- Królikowska, M. A., Bierings, E., Beatty-Martínez, A. L., Navarro-Torres, C., Dussias, P. E., and Parafita Couto, M. C. (2019). Gender assignment strategies within the bilingual determiner phrase: Four Spanish-English communities examined. *Paper presented at 3rd Conference on Bilingualism in the Hispanic and Lusophone World (BHL), January 9–11. The Netherlands*. Leiden: Leiden University.
- Liceras, J. M., Fernández Fuertes, R., Perales, S., Pérez-Tattam, R., and Spradlin, K. T. (2008). Gender and gender agreement in bilingual native and non-native grammars: A view from child and adult functional-lexical mixings. *Lingua* 118, 827–851. doi: 10.1016/j.lingua.2007.05.006
- MacSwan, J. (1999) in *A Minimalist Approach to Intrasentential Code Switching (Outstanding Dissertations in Linguistics)*. ed. L. Horn (New York, NY: Garland Publishing Inc).
- Munarriz-Ibarrola, A., Ezeizabarrena, M. J., Arrazola, D. C., Arrazola, V., and Parafita Couto, M. C. (2022). Gender assignment strategies and L1 effects in the elicited production of mixed Spanish-Basque DPs. *Linguist. Appro Bilingua*. 12, 778–851. doi: 10.1075/lab.20016.mun
- Muysken, P. (2013). Language contact outcomes as the result of bilingual optimization strategies. *Biling. Lang. Cogn.* 16, 709–730. doi: 10.1017/S1366728912000727
- Muysken, P., Kook, H., and Vedder, P. (1996). Papiamentu/Dutch code-switching in bilingual parent-child reading. *Appl. Psycholinguist.* 17, 485–505. doi: 10.1017/S0142176400008213
- Myers-Scotton, C. (1993). Common and uncommon ground: social and structural factors in codeswitching. *Lang. Soc.* 22, 475–503. doi: 10.1017/S0047404500017449
- Myers-Scotton, C. (2002). *Contact Linguistics: Bilingual Encounters and Grammatical Outcomes*. Oxford: Oxford University Press.
- Pablos, L., Parafita Couto, M. C., Boutonnet, B., De Jong, A., Perquin, M., De Haan, A., et al. (2019). Adjective-noun order in Papiamentu-Dutch code-switching. *Linguist. Approach. Bilingual.* 9, 710–735. doi: 10.1075/lab.17036.pab
- Parafita Couto, M. C., Bellamy, K., and Ameika, F. (2023). "Theoretical approaches to multilingual switching" in *Cambridge Handbook of Third Language Acquisition and Processing*. eds. J. Cabrelli, A. Chaouch-Orozco, J. G. Alonso, S. M. P. Soares, E. Puig-Mayenco and J. Rothman (Cambridge: Cambridge University Press).
- Parafita Couto, M. C., Boutonnet, B., Hoshino, N., Davies, P., Deuchar, M., and Thierry, G. (2017a). "Testing alternative theoretical accounts of code-switching using event-related brain potentials: a pilot study on Welsh-English" in *Bilingualism and Minority Languages in Europe: Current Trends and Developments*. eds. F. Lauchlan and M. C. Parafita Couto (Cambridge: Cambridge Scholars Publishing), 240–254.
- Parafita Couto, M. C., and Gullberg, M. (2019). Code-switching within the noun phrase: evidence from three corpora. *Int. J. Biling.* 23, 695–714. doi: 10.1177/1367006917729543
- Parafita Couto, M. C., Greidanus Romanelli, M., and Bellamy, K. (2021). Code-switching at the interface between language, culture, and cognition. *Lapurdum* [in press].
- Parafita Couto, M. C., Deuchar, M., and Fusser, M. (2015). "How do Welsh-English bilinguals deal with conflict? Adjective-noun order resolution" in *Code-Switching*

Between Structural and Sociolinguistic Perspectives, eds. G. Stell and K. Yakpo, vol. 43 (Göttingen: De Gruyter), 65–84.

Parafita Couto, M. C., Pérez-Tattam, R., and Guijarro-Fuentes, P. (2017b) "The Nominal Domain in Dutch-Papiamentu-Spanish Multilinguals" in *Bilingualism and Minority Languages in Europe: Current Trends and Developments*, eds. F. Lauchlan and M. C. Parafita Couto (Cambridge: Cambridge Scholars Publishing), 160–182.

Phillips, S., and Deuchar, M. (2021) in *The Role of the Input in the Acquisition of Code-Switching. Multilingualism Across the Lifespan*, eds. U. Røyneland and R. Blackwood (New York, NY: Routledge), 56–80.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: toward a typology of code-switching. *Linguistics* 18, 581–618.

R Core Team (2021). "R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2012." Available at: <https://www.R-project.org/>

Sledge, F. (2011). Let's learn Papiamentu (a wiki). Available at: <http://papiamentu.pbworks.com/w/page/8963491/FrontPage> (Accessed December 15, 2022).

Stadthagen-González, H., Parafita Couto, M. C., Alejandro Párraga, C., and Damian, M. F. (2019). Testing alternative theoretical accounts of code-switching: insights from comparative judgments of adjective–noun order. *Int. J. Biling.* 23, 200–220. doi: 10.1177/1367006917728390

Stell, G., and Yakpo, K. *Code-Switching Between Structural and Sociolinguistic Perspectives*. Göttingen: De Gruyter (2015).

Suurmeijer, L., Parafita Couto, M. C., and Gullberg, M. (2020). Structural and extralinguistic aspects of code-switching: evidence from Papiamentu-Dutch auditory sentence matching. *Front. Psychol.* 11:592966. doi: 10.3389/fpsyg.2020.592266

Urbaneja, D. R. (2020). ¿Tú tienes una little pumpkin? Mixed noun phrases in Spanish-English bilingual children and adults. *Int. J. Biling.* 24, 824–839. doi: 10.1177/1367006919888580

Valdés, K., and Jorge, R. (2016). "Mixed NPs in Spanish-English bilingual speech" in *Spanish-English Codeswitching in the Caribbean and the US* 11, eds. R. E. Guzzardo Tamargo, M. C. Parafita Couto and C. M. Mazak (Amsterdam: John Benjamins Publishing Company), 281.

van Osch, B. (2019). Vulnerability in heritage speakers of Spanish in the Netherlands: An interplay between language-internal and language-external factors. Doctoral dissertation.

van Osch, B., Boers, I., Janet Grijzenhout, M., Parafita Couto, M. C., Sterken, B., and Tat, D. (2022). "Cross-linguistic influence in bilingual grammars" in *The Acquisition of Gender*, ed. D. Ayoun (Amsterdam: John Benjamins). doi: 10.1075/sibil.63.09van

van Suchtelen, P.I. *Spanish as a Heritage Language in the Netherlands. A Cognitive Linguistic Exploration*. Utrecht: LOT (2016).

Vaughan-Evans, A., Parafita Couto, M. C., Boutonnet, B., Hoshino, N., Webb-Davies, P., Deuchar, M., et al. (2020). Switchmate! An electrophysiological attempt to adjudicate between competing accounts of adjective-noun code-switching. *Front. Psychol.* 11:549762. doi: 10.3389/fpsyg.2020.549762

Vihman, V. A. (2018). Language interaction in emergent grammars: morphology and word order in bilingual children's code-switching. *Language* 3:40. doi: 10.3390/languages3040040

Voeten, C.C. (2021). Maintainer Cesko C. Voeten. "Package builder."

Voss, D. (2018). "Insights from comparative judgments of adjective-noun order in Papiamentu-Dutch code-switching." PhD diss., BA thesis, Leiden University, Leiden.



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY

Nuria Sagarra,
Rutgers,
The State University of New Jersey,
United States
Pablo E. Requena,
University of Texas at San Antonio,
United States

*CORRESPONDENCE

Jill Jegerski
✉ jegerski@illinois.edu

RECEIVED 29 September 2022

ACCEPTED 04 April 2023

PUBLISHED 03 May 2023

CITATION

Jegerski J and Keating GD (2023) Using self-paced reading in research with heritage speakers: a role for reading skill in the online processing of Spanish verb argument specifications.
Front. Psychol. 14:1056561.
doi: 10.3389/fpsyg.2023.1056561

COPYRIGHT

© 2023 Jegerski and Keating. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Using self-paced reading in research with heritage speakers: a role for reading skill in the online processing of Spanish verb argument specifications

Jill Jegerski^{1*} and Gregory D. Keating²

¹Department of Spanish and Portuguese, University of Illinois, Urbana, IL, United States, ²Department of Linguistics and Asian/Middle Eastern Languages, San Diego State University, San Diego, CA, United States

Relatively little is known about how heritage speakers process language in real time, despite recent calls for the use of online methods such as self-paced reading, eyetracking, and ERPs (event-related potentials) in research on this early bilingual population. The present study addressed this gap with an empirical study of the online processing of heritage speakers of Spanish in the U.S. using self-paced reading, which is the online method that is most accessible to a wide body of researchers because it does not require specialized equipment. The processing target was related to the online integration of verb argument specifications, which was chosen because it does not involve ungrammatical sentences and therefore may be less likely to involve metalinguistic knowledge and less likely to put heritage speakers at a disadvantage than measures that rely on the recognition of grammatical errors. More specifically, this study examined an effect that occurs when a noun phrase appears after an intransitive verb, which can cause processing difficulty relative to a comparison condition in which the verb is transitive. The participants were 58 heritage speakers of Spanish and a comparison group of 16 first-generation immigrants raised in Spanish-speaking countries. Both groups showed the expected transitivity effect on the post-verbal noun phrase during self-paced reading, but the heritage speaker group also showed a spillover effect on the post-critical region. Among the heritage speakers, these effects were associated with lower self-ratings for reading skill in Spanish and with slower average reading speed during the experiment. Three theoretical accounts of the apparent susceptibility to spillover effects among heritage speakers are proposed: that it is a characteristic of shallow processing, that it is due to underdeveloped reading skill, and that it is an artifact of the self-paced reading method. The latter two possibilities are especially consistent with a role for reading skill in these results.

KEYWORDS

heritage speakers, Spanish (in the U.S.), self-paced reading (SPR), online methods, verb transitivity, reading skill

1. Introduction

Heritage speakers are bilingual users of a minority or community language that they have been exposed to at home from birth, but which they have typically had limited opportunities to develop, especially with language skills that are associated with formal education. For instance, heritage speakers often have underdeveloped literacy, metalinguistic skills, and formal register, as compared to their own language skills in the majority language and relative to their counterparts raised in other countries with the same L1 as a majority language (Carreira and Kagan, 2011). This difference between populations of language users raises the question of whether heritage speakers may respond differently to research methods that rely on those skills than other participant groups to which they are frequently compared in empirical studies, including more prototypical L1 users with formal education in the language and classroom-instructed adult second language (L2) learners. A number of scholars have therefore advocated for a move away from experimental tasks that rely on metalinguistic knowledge (Benmamoun et al., 2010) and toward online (i.e., real-time) methods like self-paced reading, eyetracking, and ERPs (event-related potentials; Bolger and Zapata, 2011; Jegerski, 2018a; Jegerski and Sekerina, 2021). One key advantage of online methods is that they record data in real time, while the participant is engaged in a language-related experimental task such as reading or listening, in which words are often processed in as little as 250 milliseconds. These time constraints are thought to reduce the application of metalinguistic knowledge (Montrul et al., 2008; Carreira and Kagan, 2011), so there may be less room for distortion in the data.

The present study employed the self-paced reading method because it is more accessible to a wider body of researchers than other online methods in terms of cost and the level of technical knowledge required. The focus was on a previously-documented processing phenomenon related to the online integration of verb argument specifications and for which the stimulus sentences are all grammatical, so the effect may be less likely to involve metalinguistic knowledge than processing effects that occur with ungrammatical sentences. The use of a reading-based method could be a potential concern with heritage speakers, given that they tend to have underdeveloped literacy, but over 90% of heritage speakers of Spanish in the U.S. have at least an intermediate reading level (Carreira and Kagan, 2011), which would likely be enough for self-paced reading. Indeed, previous research had successfully used self-paced reading with this population on several occasions prior to the present study (Foote, 2011; Jegerski et al., 2016; Keating et al., 2016; Jegerski, 2018b,c).

Like most previous work on heritage speakers using self-paced reading and other online methods (to be discussed in greater detail in the following sections), the present study examined the processing of different sources of linguistic information in real time. The fundamental question driving such research is whether grammatical details such as gender agreement or relative clause attachment are quickly accessed and integrated into an underlying representation of the segment of language that is being comprehended. Some theories propose that such grammatical details are sometimes overlooked during real-time processing, especially among less proficient language users such as adult second language learners (e.g., Clahsen and Felser, 2006; Christianson, 2016). The limited research on online processing

among heritage speakers has mostly taken a similar approach, often with the inclusion of some type of comparison group of more prototypical L1 users as a point of reference for sensitivity to the linguistic target during processing (e.g., Sekerina and Trueswell, 2011; Jegerski, 2018c). More recently, some researchers have begun to examine within-group variability in processing among heritage speakers in relation to individual difference variables such as proficiency (Bice and Kroll, 2021) and age of acquisition of the L2 majority language (Keating, 2022). The present study took a combined approach, including both a comparison group of L1 users of Spanish who were first-generation immigrants (to approximate the language that heritage speakers received as input while growing up in the U.S.; Polinsky and Scontras, 2020) and an analysis of within-group variability among heritage speakers.

2. Literature review

2.1. The processing of verb transitivity in Spanish

The present study examined a processing effect that is known to occur in both English and Spanish and which has been observed with both self-paced reading (Berghoff, 2020) and eyetracking (Staub, 2007). In sentences like (1) below, processing difficulty typically occurs on the first post-verbal noun phrase *the veterinarian* when the verb it follows is intransitive as in (1b), compared to when the verb is transitive as in (1a).

- (1)
 - a. After the dog scratched the veterinarian took off the muzzle.
TRANSITIVE.
 - b. After the dog struggled the veterinarian took off the muzzle.
INTRANSITIVE.

The observed processing difficulty is thought to arise from a conflict between the processing principle of Late Closure, which is a preference to incorporate each word into the current phrase whenever possible rather than to initiate a new clause (Frazier and Fodor, 1978), and the argument specifications of the intransitive verb, which do not allow such a structure. In other words, there is a tendency to process the post-verbal noun phrase as a direct object, but this is not possible when the verb in question is intransitive. A similar processing effect has been observed when the verb is transitive but the post-verbal noun phrase is semantically implausible as a direct object, as in *As the men drank the song...* versus *As the men drank the beer...* (Roberts and Felser, 2011). There is evidence these effects can be more robust among slower readers (Roberts and Felser, 2011; Jegerski, 2012), possibly because slower reading may lead to more incremental processing than faster reading (Roberts and Felser, 2011, p. 323), and also that L1 users do not always show an online plausibility effect during reading (Roberts and Felser, 2011).

From the perspective of acquisition, language users who exhibit the aforementioned effect must have acquired the relevant verb argument specifications and have the ability to apply them efficiently during online processing, along with the processing principle of Late Closure. Otherwise, the two would not be in conflict and there would not be an increase in processing difficulty.

2.2. The processing of verb transitivity in bilingual populations

To our knowledge, no prior investigation has examined the verb transitivity effect targeted in the present study among heritage speakers, so it is not known whether they are sensitive to this type of information during online processing. There have been two related studies that included early bilinguals (Berghoff, 2020; McCormick, 2020), but the target languages were not minoritized, so the participants were not heritage speakers. The first of these two prior investigations was Berghoff's (2020) self-paced reading study, in which the bilingual participants were L1 Afrikaans speakers tested in their L2, English, in South Africa. Another difference between that investigation and the present one was that the linguistic stimuli for that study manipulated the semantic plausibility of the post-verbal noun phrase as an object (e.g., *As the men drank the song...* versus *As the men drank the beer...*; Roberts and Felser, 2011) to create conflict with the processing principle of Late Closure, whereas the present study manipulated the transitivity of the verb. Berghoff (2020) compared the early Afrikaans-English bilinguals to late English-Afrikaans bilinguals and both groups showed the expected reading time increases on a post-verbal noun phrase that was implausible as a direct object of the verb versus a noun phrase that was a plausible object. In other words, they appeared to rapidly integrate verb argument specifications during online processing.

The second related investigation was McCormick (2020), which included simultaneous (2L1) Catalan-Spanish bilinguals in Spain. This self-paced reading study used Spanish stimuli that were nearly identical to those for the present study (both taken from Jegerski, 2012), so verb transitivity was manipulated. The expected reading time increase was observed when a post-verbal noun phrase followed an intransitive verb rather than a transitive one. Additionally, McCormick (2020) also employed a cognitive control engagement paradigm, in which the self-paced reading stimuli alternated with trials in a flanker task,¹ and found that the verb transitivity effect was diminished when the stimulus was read after an incongruent flanker trial as opposed to a congruent trial. This appeared to be a Gratton effect (Gratton et al., 1992), in which there is prolonged cognitive engagement following stimuli with conflict, so the researcher concluded that the same cognitive control mechanisms were used to resolve conflict in the intransitive verb stimuli as in the flanker task.

Other relevant previous research comes from adult L2 acquisition, where the focus is on the L2 of late bilinguals. In this area, at least five studies have reported apparent processing difficulty on a post-verbal noun in similar stimuli, although the effect was not always due to verb

transitivity and in some cases it was an incidental finding rather than the focus of the investigation. Jegerski (2012) observed the same effect targeted in the present study in the self-paced reading times of adult L2 learners of Spanish and a monolingual L1 comparison group, which suggests that L2 learners can acquire verb subcategorization information and apply it efficiently during online processing. On the other hand, the L2 participants in that study were of very high proficiency and two subsequent studies with L2 learners at a lower proficiency level found that verb transitivity did not affect their processing of the post-verbal noun phrase (Brothers et al., 2021: eyetracking study of L2 English; McCormick, 2020: self-paced reading study of L2 Spanish), so it appears that a certain level of language proficiency is necessary to successfully acquire and integrate verb subcategorization specifications during the processing of post-verbal nouns. Frenck-Mestre and Pynte (1997) observed a similar lack of transitivity effect among a group of L2 users of English that were of higher proficiency than in Brothers et al. (2021) and probably lower than in Jegerski (2012), but the L1 group in that study also showed only a marginal effect ($p = 0.09$) and the L2 group did show a nonsignificant numerical difference in the expected direction, so it seems possible that this eyetracking study may have been underpowered with only 16 participants in each group, L1 and L2. Finally, Roberts and Felser (2011) observed that advanced proficiency L2 learners showed even more robust processing effects than L1 users, but their stimuli manipulated the semantic plausibility of the post-verbal noun phrase as a direct object of the verb rather than verb transitivity. Roberts and Felser (2011) therefore concluded that L2 learners are overly sensitive to semantics during processing and argued that this is a compensatory strategy to make up for a lack of sensitivity to syntax and morphosyntax, in line with the theory of Clahsen and Felser (2006).

Thus, early bilinguals appear to integrate verb argument specifications during online processing in their L1 and L2, although the participants in previous work were speakers of two mainstream languages rather than heritage speakers of a minoritized language. There is also evidence that such effects may be related to cognitive control mechanisms, in addition to the linguistic knowledge and processing strategy (Frazier and Fodor, 1978) involved. Late bilinguals have exhibited similar verb transitivity effects while processing their L2, but this appears to require a relatively high level of proficiency. Based on this existing body of evidence, a reasonable expectation is that heritage speakers would also show online verb transitivity effects in their home language, assuming they have acquired a sufficiently high level of proficiency. Hence, our prediction for the present study was that at least some heritage speakers of Spanish would show a verb transitivity effect during self-paced reading and that the effect might vary according to Spanish proficiency level as an individual difference variable. Such an observation would contribute to our knowledge of areas of resilience among heritage speakers, in line with a recent call to broaden the prevailing research focus beyond areas of vulnerability and divergence (Polinsky and Scontras, 2020).

2.3. Online methods in research with heritage speakers

As outlined in the introduction to this article, a number of scholars have called for research on heritage speakers using online

¹ The flanker task (Eriksen and Eriksen, 1974) is widely used as a measure of inhibitory control. In this technique, a target stimulus is flanked by non-target stimuli that can be congruent, non-congruent, or neutral. For instance, participants may be asked to indicate whether the center arrow in a row of seven is pointing to the left or right and the three arrows appearing on either side of the target might be pointing in the same direction as the target or in the opposite direction. The presence of non-congruent stimuli is typically associated with slower and less accurate responses, so a left-pointing arrow is harder to identify when it is flanked by right-pointing arrows than by left-pointing ones.

(i.e., real-time) psycholinguistic methods (Bolger and Zapata, 2011; Jegerski, 2018a; Jegerski and Sekerina, 2021). To date, only a limited number of studies have been conducted using these techniques, but there are enough to suggest that the methods can be useful in work on heritage speakers.

Of the three most common online research methods, self-paced reading, eyetracking, and ERPs, self-paced reading is the most accessible to a wide body of researchers because it does not require specialized equipment, it is much less expensive than the other two methods, and it does not require as much technical training. Nevertheless, it does require dedicated software, extensive knowledge of materials design (see Keating and Jegerski, 2015, for more information), and knowledge of the current approaches to statistical analysis, which have become more complex over time. At least five research studies using self-paced reading with heritage speakers have been published to date (i.e., prior to the publication of this research topic in *Frontiers in Psychology*). One early example is Foote's (2011) investigation of the processing of agreement by heritage speakers of Spanish, which revealed online sensitivity to both gender and number agreement that was similar to that of a comparison group of more prototypical L1 users who were raised abroad with Spanish as a majority language. Additional previous research with self-paced reading has examined relative clause attachment (Jegerski et al., 2016; Jegerski, 2018b), pronominal reference (Keating et al., 2016), and differential object marking (Jegerski, 2018c), all in heritage Spanish.

Eyetracking has also been used in several studies of heritage speakers, both with text and with the visual world paradigm (in which the linguistic stimuli are auditory and the visual stimuli are images or physical objects). For instance, Sekerina and Trueswell's (2011) visual world experiment showed that heritage speakers of Russian were slower than monolinguals to integrate word order and visual context in the processing of contrastive focus during auditory processing. In another example using the visual world eyetracking paradigm, Jegerski and Sekerina (2020) observed that heritage speakers of Spanish showed similar online sensitivity to the object marker "a" in auditory questions as more prototypical L1 users raised abroad, even though the heritage speakers were less accurate in their offline responses to the questions. Finally, the results of Fuchs (2021) visual world eyetracking study suggested that heritage speakers of Spanish were able to use grammatical gender for predictive processing of auditory stimuli, similar to a monolingual comparison group.

In a study that employed eyetracking with written stimulus sentences rather than auditory stimuli, Keating (2022) compared the processing of Spanish gender agreement among heritage speakers who had acquired their two languages simultaneously to those who had acquired them sequentially and found that online sensitivity occurred earlier in the eye movement record for the sequential bilinguals, who had longer exposure to just Spanish before beginning to acquire the majority language, English. Parshina et al. (2022) also used eyetracking with text to show that heritage speakers of Russian could predict lexical and morphosyntactic information for upcoming words while reading and that this ability appeared to correlate with literacy experience in Russian. Lastly, Parshina et al. (2021) used eyetracking to document some general tendencies in the reading behavior of heritage speakers of Russian as compared to monolingual readers, more specifically, that they read more slowly, that they were less likely to skip words (which is a normal part of fluent reading), and that they were more likely to reread than the comparison group.

Finally, we are aware of three studies that have employed the ERP method, all with heritage speakers of Spanish. First, Martohardjono et al. (2017) observed that heritage speakers, like a comparison group of more prototypical L1 users raised in Spanish-speaking countries, exhibited the expected P600 and N400 waveforms in response to different types of syntactic anomalies. Second, Rossi (2021) found individual variation in the ERP responses of heritage speakers to gender and number violations. Specifically, the group as a whole did not show sensitivity to gender and number, but a subset of participants did show the expected P600 waveform response, while others showed an N400, which is typically observed in response to semantic anomalies rather than morphosyntactic ones. Finally, Bice and Kroll (2021) observed smaller P600 and N400 responses among heritage speakers than with monolinguals and the researchers also found that variability in the heritage speakers was related to proficiency, whereas with the monolinguals the main factor was working memory.

To summarize, a number of prior empirical investigations have employed online methods in research with heritage speakers. Although a number of scholars working with heritage languages have proposed that online methods can and should be used with this population because they tend to be less metalinguistic than more traditional techniques (e.g., Bolger and Zapata, 2011; Jegerski, 2018a; Jegerski and Sekerina, 2021), empirical experimentation is a critical piece that can establish support for such claims (e.g., Martohardjono et al., 2017). Hence, it is important to note that the outcomes of these studies have been generally positive with regard to methodology, meaning that heritage speakers have often shown the effects that would be expected in research on other populations of language users. Some of the cited researchers have even concluded that online methods are especially appropriate because they can reveal a higher level of heritage language ability than would be evident in other measures (e.g., Martohardjono et al., 2017). On the other hand, an additional goal of this line of work is to determine to what extent there may be special considerations that should guide work using online methods with heritage speakers. This is particularly true with reading-based methods, as pointed out by some of the researchers cited above (Jegerski, 2018b; Parshina et al., 2021, 2022), because literacy tends to be underdeveloped among heritage speakers (Carreira and Kagan, 2011).

Given this background and the goals of this research topic in *Frontiers in Psychology*, the objective of the present study was to contribute to the very limited but growing body of work on heritage speakers using online methods, with particular attention to the research methodology and its effectiveness with this participant population. More specifically, the current investigation was a self-paced reading study of the processing of verb transitivity (as outlined above) among heritage speakers of Spanish and a comparison group of more prototypical L1 users raised with Spanish as a majority language and who, as first-generation immigrants to the U.S., were also bilingual. We also examined the role of several background variables that tend to vary between heritage speakers and more prototypical L1 users, that pertain to reading specifically, or that are of general interest with heritage speakers: self-rated reading ability in Spanish, average reading speed during the experimental self-paced reading task, age of acquisition of English, Spanish proficiency test score, and self-reported current exposure to Spanish.

3. Method

3.1. Participants

The two participant groups for this study were selected to (1) represent U.S. heritage speakers of Spanish with a range of key individual difference variables such as proficiency, age of onset of bilingualism, and measures of reading skill and (2) to compare the sentence processing of heritage speakers with that of L1 users who would have provided them with input while they were growing up in the U.S., meaning first generation immigrants (the “appropriate baseline” for heritage speakers, as per Polinsky and Scontras, 2020, p. 5). The primary group of interest was comprised of 58 heritage speakers of Spanish who were all early Spanish-English bilinguals that were exposed to Mexican Spanish from birth. The comparison group was comprised of 16 immigrants who were also native speakers of Mexican Spanish and who had acquired the language as children in Mexico, where they received formal education. The participants in the comparison group were also bilingual because they knew English, but they had not begun to acquire the language until at least age 12. All participants were recruited on the campus of a large public university in non-borderland Texas and tested in person. More detailed participant background information is provided in Table 1, where it can be seen that the two groups differed with regard to the individual difference variables of Spanish proficiency test score, self-ratings for speaking, understanding, and reading skills in Spanish and English, and estimated relative exposure to both languages. In addition, as seen in the standard deviations in Table 1, there was variability within each group with regard to these measures and greater variability among the

heritage speakers, which is common with this population and was intentional in this study because of the analysis of individual differences.

3.2. Materials

An example of the self-paced reading stimuli can be seen in (2) below, where the slashes indicate how the sentence was segmented into phrases. The 20 sentences for the present experiment were based on those employed in two prior self-paced reading studies (Jegerski, 2012; McCormick, 2020), so they were known to elicit the desired processing effects in monolingual native speakers, 2L1 Catalan-Spanish bilinguals, and late L2 learners of Spanish with high proficiency. As described in the literature review above, sentences such as those in (2) below are typically associated with longer reading times on the first post-verbal noun phrase *el violín* “the violin” when the verb it follows is intransitive as in (2b), compared to when the verb is transitive as in (2a).

(2) *Stimulus for Self-Paced Reading.*

a. Mientras el maestro/tocaba /el violín/resonaba/por todo el salón. TRANSITIVE.

“While the maestro/played /the violin/resonated/throughout the hall.”

b. Mientras el maestro/descansaba /el violín/resonaba/por todo el salón. INTRANSITIVE.

“While the maestro/took a break/the violin/resonated/throughout the hall.”

The observed processing difficulty is thought to arise when the comprehender integrates both the processing principle of Late Closure, which is a preference to incorporate each word into the current phrase whenever possible rather than to initiate a new clause (Frazier and Fodor, 1978), and the argument specifications of the intransitive verb, which do not allow such a structure. In other words, there is a tendency to process the post-verbal noun phrase as a direct object, but this is not possible when the verb in question is intransitive, so there is a conflict that needs to be resolved during processing.

Each of the 20 stimuli appeared in two conditions, transitive and intransitive. The transitive and intransitive verbs were similar to each other in terms of frequency (Davies, 2005), according to an independent-samples *t*-test: $t(38) = 0.048$, $p = 0.962$. This was only to ensure ease of lexical access; reading times for the different verbs were not compared to each other in any of the statistical analyses. The critical region of interest for which data were analyzed was the post-verbal noun phrase, which was identical in both conditions, so all relevant linguistic variables were controlled. The post-critical region was also identical in both the transitive and intransitive conditions. Each sentence began with a subordinating conjunction such as *mientras* “while,” *antes de que* “before,” or *cuando* “when.”

The stimulus materials design and counterbalancing were as recommended by Jegerski (2014) and Keating and Jegerski (2015). Two counterbalanced presentation lists were created with 10 critical sentences in each condition and each sentence appearing only once in any condition per list. The 20 target stimuli were combined with 140 total distractors and fillers. The distractors were 40 stimuli for another experiment that focused on relative clause attachment (Jegerski, 2018b), as exemplified in (3) below, and the fillers were

TABLE 1 Language background information.

	Heritage speakers (<i>n</i> =58)			Comparison group (<i>n</i> =16)		
	<i>M</i>	<i>SD</i>	Range	<i>M</i>	<i>SD</i>	range
Age of acquisition						
English	3.12	2.63	0–8	14.00	1.55	12–16
Spanish	0.17	0.60	0–3	0.00	0.00	0
DELE score	35.81	6.80	24–47	45.56	3.42	38–49
Self-rating of skills: English						
Understanding	9.43	0.80	7–10	8.31	1.08	7–10
Speaking	9.36	0.87	6–10	7.75	1.13	5–10
Reading	9.34	0.85	7–10	8.19	0.75	7–9
Self-rating of skills: Spanish						
Understanding	8.49	1.66	1–10	9.63	0.50	9–10
Speaking	7.89	1.60	1–10	9.44	0.63	8–10
Reading	7.51	1.67	3–10	9.44	0.63	8–10
Current exposure						
English	64.12%	16.87	25–100	43.75%	23.92	3–80
Spanish	35.21%	17.04	0–75	55.31%	23.92	20–97
Average reading speed	995 ms	249	675–1836	886 ms	200	550–1246
Age	22.91	9.31	18–60	25.75	5.73	18–38

The maximum DELE score was 50 and the maximum for self-rated proficiency was 10.

non-experimental sentences that did not target or manipulate any particular linguistic form. The filler and distractor sentences varied in terms of length, but most were complex with two clauses. The stimuli were presented in pseudo-random order such that no two sentences of the same type appeared in succession.

(3) Distractor for Self-Paced Reading.

El jurado / consultó/con la abogada/del acusado/que estaba parada.

“The jury/consulted/ with the lawyer/of the defendant/who was standing.”

Beyond the self-paced reading task, the materials included a Spanish proficiency test and a background questionnaire. The proficiency test was one that has been used for at least 20 years in research on the acquisition of Spanish, starting with Montrul and Slabakova (2003), and which has more recently been shown to correlate with other measures of proficiency such as elicited imitation among heritage speakers (Solon et al., 2022). It is a modified version of the written DELE (*Diploma de español como lengua extranjera* “Diploma of Spanish as a Foreign Language”) with 50 items targeting grammar and vocabulary and for which the maximum score is 50.

The questionnaire was the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian et al., 2007), which included the key individual difference variables of age of acquisition of English, self-reported current exposure to Spanish (“What percentage of the time are you currently and on average exposed to each language?”), and self-rated reading ability in Spanish, plus the additional participant background information that is reported in Table 1.

3.3. Procedure

The self-paced reading stimuli were presented in a left-to-right, non-cumulative format using SuperLab (Cedrus Corporation, 2007). Each trial started with a “+” cue symbol that appeared at the leftmost edge of where the first segment of the stimulus sentence would appear; this was to encourage participants to look at the stimulus right away, beginning with the first word, rather than at other parts of the display. Words were masked with dashes but spaces and punctuation remained visible. Participants used a button on a response pad to proceed through each segment of a stimulus sentence at their own pace. Each segment contained one or more words, as illustrated above in (2).

After all segments of a stimulus sentence had been read, a subsequent display screen showed a binary choice comprehension question. As seen in Example (4) below, which followed the example stimulus in (2) above, the post-stimulus questions targeted the meaning of the sentence rather than the participant’s interpretation of a specific linguistic form (and this is why we refer to them as *comprehension* questions rather than *interpretation* questions). Participants responded to the questions using two keys on a Cedrus RB-730 response pad marked with the letters “A” and “B,” which were on the left and right sides of the response pad, respectively. The target responses were counterbalanced such that half were “A” and half were “B” and they were also randomized, to avoid the effects of handedness or other biases.

(4) Post-Stimulus Comprehension Question.

¿Dónde puede estar este músico?

a. En un parque.

b. En un teatro.

“Where might this musician be?”

“a. In a park.”

“b. In a theater.”

Detailed instructions and five practice items were presented prior to the experimental block. Participants were told that the test targeted reading comprehension in Spanish and no feedback was provided during the experiment. An optional 10-min break was offered when the participant had read half of the 160 total sentences included in the self-paced reading. Each participant completed all the experimental tasks in a single session lasting 60 to 90 min, including the background questionnaire, the self-paced reading, and the proficiency test, in that order. Participants were paid for their time.

3.4. Statistical analysis

All data were analyzed via mixed effects linear and logistic regression using R (R Development Core Team, 2019) with the *lme4* package (Bates et al., 2015). The models included verb transitivity, group, and the transitivity \times group interaction as fixed effects, plus subject and item as crossed random effects. Because the two participant groups were not exactly matched for age (heritage speakers $m = 22.9$, comparison group $m = 25.8$; see Table 1) and age can affect reading times, it was included as a covariate in all of the statistical models. Deviation coding was used to obtain main effects. Logit models were used with binary choice comprehension accuracy data (Jaeger, 2008). Following current procedure in psycholinguistics for identifying the maximal random effect structure appropriate for the sample (Barr, 2013), each model was first run with the maximal random effect structure, then in cases where that model did not converge, it was incrementally simplified to identify the maximal effect structure that still converged. In the case of interactions in the primary models, pairwise comparisons were examined using the *emmeans* package with the Bonferroni correction (Lenth et al., 2018). R code with the final random effects structure for each of the main analyses can be found under the corresponding output tables. *p* values were obtained using Satterthwaite’s approximation for degrees of freedom with the *lmerTest* package for R (Kuznetsova et al., 2014). Prior to statistical analysis, outlying reading times of less than 100 milliseconds were eliminated because they are more likely to represent errors (e.g., premature button presses in this study) than true linguistic processing (Rayner, 1998) and those beyond 5000 milliseconds were trimmed to the cutoff value, which affected 0.39 and 0.64% of the data, respectively. Response times were also log transformed to reduce the positive skew. Alpha was set at 0.05 for all analyses and *p* values of 0.05 to 0.10 would have been considered to be marginally significant in order to reduce the chance of a Type II error (i.e., a false negative; Larson-Hall, 2010), although none of the analyses for this study yielded any such marginal *p* values.

4. Results

Mean accuracy proportions and response times for the post-stimulus comprehension questions, by group and transitivity, can be found in Table 2. The statistical analyses for these data are reported

in Table 3. Accuracy was high overall (heritage speakers: $M = 0.892$, $SD = 0.311$; comparison group: $M = 0.934$, $SD = 0.248$), which indicates that participants generally paid attention while reading, although the heritage speakers were less accurate overall than the comparison group. There was no effect of transitivity or interaction between the two factors in the analysis of the accuracy data. There was also no effect of age. In addition, the analysis of the response times for the post-stimulus comprehension questions also revealed a main effect of group, in which the heritage group was slower to respond than was the comparison group. There was also an effect of age, which reflected longer response times among older participants. There was no effect of transitivity and no interaction of transitivity with group.

Mean self-paced reading times by group and transitivity condition from the critical NP and the post-critical word (i.e., the main clause verb) can be found in Table 4. The main statistical analysis of the reading time data is reported in Table 5. At the critical region with the post-verbal NP, there was a main effect of transitivity, in which reading times were longer when the NP followed an intransitive verb versus a transitive one, and a main effect of group, with the reading times of heritage speakers being generally longer than those of the comparison group. There was no effect of age. There was no interaction of transitivity with group, which indicates that the transitivity effect was similar across both groups.

At the post-critical word, the main clause verb that followed the critical NP, there was no main effect of transitivity, but there was a main effect of group, in which the reading times of heritage speakers were generally longer than those of the comparison group and there was a main effect of age, in which the reading times of older participants were also generally longer. Most importantly, transitivity interacted with group. Pairwise comparisons conducted to probe the interaction revealed that the effect of transitivity was significant for the heritage group, estimate = 0.033, $SE = 0.012$, $t = 2.841$, $p = 0.009$, but not for the comparison group, estimate = 0.009, $SE = 0.019$, $t = 0.483$, $p = 0.630$.

The transitivity effect can be taken as a sign of efficient online processing across both groups, but the spillover effect that was evident only among the heritage speakers might be related to any of several language background variables that differed both between the two groups and especially among the heritage speakers. For this reason, we conducted a secondary set of statistical analyses to explore what language background and reading-based variables might play a role in this aspect of sentence processing among heritage speakers. Each model examined the effect of transitivity, one centered background variable (run separately to avoid issues with multicollinearity), and their interaction. Age was again included as a covariate. As with the main analyses above, each model had random intercepts for subject

and item and random slopes for transitivity by subject and by item wherever possible (i.e., as with the main analyses above, the slopes were simplified if the model did not converge). A total of five background variables from Table 1 were analyzed for both the critical noun phrase and the post-critical region: age of acquisition of English, DELE proficiency test score, self-reported current exposure to Spanish, self-rated reading ability in Spanish, and average reading speed for the self-paced reading task (calculated as the mean reading time across all sentence regions and across all sentences in the self-paced reading task, including experimental items, distractors, and fillers).

All ten models showed a main effect of transitivity ($ts > 2.3$ and $ps < 0.03$) and most also showed a significant or marginally significant effect of age, consistent with the main analyses above. There were also main effects at both stimulus regions for the DELE proficiency test score (R3: estimate = 0.006, $SE = 0.002$, $t = 3.436$, $p = 0.001$; R4: estimate = 0.006, $SE = 0.002$, $t = 3.498$, $p = 0.001$), for self-rated reading ability in Spanish (R3: estimate = 0.018, $SE = 0.006$, $t = 2.944$, $p = 0.005$; R4: estimate = 0.024, $SE = 0.006$, $t = 4.031$, $p < 0.001$), and average reading speed (R3: estimate = 0.000, $SE = 0.000$, $t = 16.752$, $p < 0.001$; R4: estimate = 0.000, $SE = 0.000$, $t = 13.014$, $p < 0.001$), but not for age of acquisition of English or self-reported current exposure to Spanish (all $ts < 0.90$ and $ps > 0.40$). The main effects reflected generally longer reading times with a lower DELE score, with a lower self-rating for reading ability, and with slower average reading speed. The interaction with transitivity was significant only at the critical NP and only in the models with self-rated reading ability in Spanish (R3: estimate = 0.008, $SE = 0.003$, $t = 3.041$, $p = 0.004$; R4: estimate = 0.003, $SE = 0.002$, $t = 1.382$, $p = 0.167$) and average reading speed (R3: estimate = 0.000, $SE = 0.000$, $t = 3.962$, $p < 0.001$; R4: estimate = 0.000, $SE = 0.000$, $t = 0.744$, $p = 0.457$); other $ts < 1.4$ and $ps > 0.15$. These interactions reflected a more pronounced transitivity effect at the critical NP with lower self-ratings for reading and with slower average reading speed.

Thus, the main results of this investigation can be summarized as follows:

- The expected main effect of transitivity was evident on the critical NP across both groups: reading times were longer when the NP followed an intransitive verb than when it followed a transitive verb.
- The same effect spilled over to the post-critical region, the main clause verb, but only among the heritage speakers.
- Additional analysis of the heritage speaker data with language background variables revealed that greater transitivity effects were associated with lower self-ratings for reading and with slower average reading speed, but this was only on the critical NP and not on the spillover region.
- The heritage speakers also showed generally longer reading times for the stimulus sentences and longer response times and lower accuracy for the post-stimulus comprehension questions than the comparison group.

5. Discussion

The present study examined the processing of verb transitivity among heritage speakers of Spanish and a comparison group of more prototypical L1 users who had acquired Spanish in a majority language context before immigrating to the U.S. as adults. Both groups showed

TABLE 2 Comprehension question accuracy and response times (SDs in parenthesis).

	Heritage speakers	Comparison group
<i>Accuracy</i>		
Transitive	0.89 (0.31)	0.94 (0.24)
Intransitive	0.89 (0.31)	0.93 (0.25)
<i>Response times</i>		
Transitive	3839 (1026)	3547 (1064)
Intransitive	3895 (1022)	3664 (1020)

TABLE 3 Analysis of responses to comprehension questions: output from logistic and linear mixed-effects models.

	Estimate	SE	z/t	p
<i>Comprehension accuracy</i>				
Intercept	2.941	0.283	10.383	0.000*
Transitivity	0.061	0.126	0.486	0.627
Group	0.336	0.166	2.025	0.043*
Age	0.019	0.014	1.394	0.163
Transitivity × group	0.005	0.126	0.040	0.968
<i>Response times</i>				
Intercept	3.553	0.018	197.099	0.000*
Transitivity	0.006	0.005	1.260	0.219
Group	0.019	0.009	2.091	0.040*
Age	0.002	0.001	2.299	0.024*
Transitivity × group	0.002	0.003	0.550	0.582

*Effect significant at $\alpha = 0.05$.

The R code for these models was as follows: ACC = glmer (Accuracy ~ 1 + Transitivity * Group + Age + (1|ITEM) + (1|SUBJECT), data = R99, family = binomial). RT = lmer (log(RT) ~ 1 + Transitivity * Group + Age + (1 + Transitivity + Group|ITEM) + (1|SUBJECT), data = R99).

TABLE 4 Trimmed response times (SDs in parenthesis).

	Heritage speakers	Comparison group
<i>Critical NP</i>		
Transitive	866 (482)	748 (386)
Intransitive	987 (583)	868 (498)
<i>Critical NP + 1</i>		
Transitive	773 (366)	725 (350)
Intransitive	875 (599)	713 (396)

the expected effect during self-paced reading, which suggests that they successfully integrated verb transitivity specifications and the structural principle of Late Closure during online processing, as it is the conflict between the two that is thought to underlie the processing effect in question. This outcome was tentatively predicted based on previous research that had observed the same processing effect with other populations of early bilinguals (Berghoff, 2020; McCormick, 2020) and with adult L2 learners with advanced proficiency (Roberts and Felser, 2011; Jegerski, 2012). Thus, there is now a growing body of evidence that shows that a range of bilingual language users are sensitive to verb transitivity during processing, although it should be noted that some groups of L2 participants have failed to show the online effects in question (Frenck-Mestre and Pynte, 1997; McCormick, 2020; Brothers et al., 2021), probably due to having a lower level of proficiency. This observation also suggests that the processing of verb argument specifications may be an area of so-called “resilience” among heritage speakers, which is an important gap in the knowledge base noted by Polinsky and Scontras (2020).

In addition to the basic effect that occurred at the critical region of the stimulus sentences (i.e., the post-verbal noun phrase), the

TABLE 5 Analysis of self-paced reading times: output from linear mixed-effects models.

	Estimate	SE	t	p
<i>Critical NP</i>				
Intercept	2.883	0.025	117.564	0.000*
Transitivity	0.025	0.007	3.422	0.002*
Group	0.032	0.015	2.234	0.029*
Age	0.002	0.001	1.665	0.100
Transitivity × group	0.002	0.006	0.257	0.798
<i>Critical NP + 1</i>				
Intercept	2.836	0.021	134.915	0.000*
Transitivity	0.006	0.006	0.995	0.330
Group	0.032	0.014	2.331	0.023*
Age	0.004	0.001	2.929	0.005*
Transitivity × group	0.011	0.005	2.218	0.027*

*Effect significant at $\alpha = 0.05$.

The R code for these models was as follows: NP = lmer (log(RT) ~ 1 + Transitivity * Group + Age + (1 + Transitivity|ITEM) + (1 + Transitivity|SUBJECT), data = R3). NP + 1 = lmer (log(RT) ~ 1 + Transitivity * Group + Age + (1 + Transitivity + Group|ITEM) + (1|SUBJECT), data = R4).

heritage speakers displayed a continued effect that carried over into the following region. This apparent spillover effect was not evident among the comparison group in the present study, nor was it observed among any of the participant groups in two previous studies with very similar stimuli that also manipulated verb transitivity (Jegerski, 2012; McCormick, 2020).

On the other hand, two prior investigations of similar processing effects with stimuli that manipulated noun plausibility (as a direct object of the verb that preceded it) rather than verb transitivity had observed some type of continuation of processing effects among other participant populations. Specifically, the L2 participants in Roberts and Felser (2011) displayed longer reading times on both the critical noun and the verb that followed it, similar to the heritage speakers in the present study. An L1 comparison group in Roberts and Felser (2011) failed to show the effect on either stimulus region. In addition, Berghoff (2020) observed a prolonged processing effect among a childhood L2 group, with longer reading times on two post-critical words (and no effect on the critical noun itself). The L1 group in that study also showed a prolonged effect over the same two stimulus regions as did the L2 group, although the numerically longer reading times were only marginally significant on the second post-critical word with the L1 group. In both studies, the results were taken as evidence of greater sensitivity to semantic information such as plausibility in L2 processing as compared to L1 processing, in line with the Shallow Structure Hypothesis (Clahsen and Felser, 2006), which claims that L2 processing is more sensitive to semantic information because it can help to compensate for purported deficiencies in syntactic processing.

Thus, one explanation for the extended effect observed among the heritage speakers in the present study is that they were more sensitive to verb transitivity than the comparison group, perhaps because of a need to compensate for a lack of grammatical detail in processing, in line with the claims of the Shallow Structure Hypothesis for L2

processing (Clahsen and Felser, 2006). However, one potential problem with this account is that it is not clear to what extent the semantic plausibility effect from these previous studies is comparable to the verb transitivity effect in the present and two previous studies (Jegerski, 2012; McCormick, 2020). There is evidence that verb subcategorization specifications may be of higher priority than the semantic plausibility of nouns as objects, at least in monolingual L1 processing (Garnsey et al., 1997), which is in line with the observation from previous research that online effects appear to have been more consistent and localized with verb transitivity (Jegerski, 2012; McCormick, 2020) than with plausibility (Roberts and Felser, 2011; Berghoff, 2020). Still, verb transitivity is similar to semantic plausibility in terms of where it fits in the Shallow Structure Hypothesis, meaning it would be intact or even over-emphasized during so-called “shallow” processing (Clahsen and Felser, 2006, p. 18).

A second explanation for the extended effect observed among the heritage speakers in the present study is that it is related to reading skill. Literacy skills are typically underdeveloped in heritage speakers (Carreira and Kagan, 2011) and one prior study of heritage speakers using self-paced reading found that sentence processing was related to reading (Keating et al., 2016). Specifically, the participants who read in Spanish more often were more similar to a monolingual comparison group in their processing of pronominal reference. Along these same lines, the analysis of individual difference variables in the present study showed that greater transitivity effects were associated with slower average reading speed during the self-paced reading task and also with lower self-ratings for reading ability in Spanish. In other words, slower and less skilled readers had a larger reading time effect (i.e., greater processing difficulty) upon encountering a noun phrase that followed an intransitive verb versus a transitive one. This outcome is broadly consistent with the results of two prior investigations, albeit with different participant populations. As mentioned above, the two previous studies employed slightly different types of stimulus sentences, with those of Jegerski (2012) very closely resembling the verb transitivity stimuli from the present study and those of Roberts and Felser (2011) instead manipulating the semantic plausibility of a post-verbal noun phrase. Jegerski (2012) subdivided monolingual L1 and very advanced L2 participant groups based on a median split for average reading speed (during the experimental self-paced reading task, as in the present study) and found that only the slower L1 readers showed the processing effect in question. Reading speed did not appear to matter for the L2 group in that study, which showed the effect regardless of sub-group. Roberts and Felser (2011) performed a similar analysis and observed that slower L2 readers exhibited an extended processing effect over two stimulus regions, whereas the faster L2 readers showed the effect only on the second stimulus region. In that study, reading speed did not seem to matter for the L1 group.

It is interesting to note that both L1 and L2 processing can vary according to reading speed, but do not seem to do so consistently (i.e., across both studies). Most relevant to the present study is that greater processing difficulty on a post-verbal noun that cannot be integrated as an object of the verb immediately before it does not appear to be unique to heritage speakers. Nevertheless, to the extent that they are generally slower and less skilled readers in the heritage language, heritage speakers could potentially be more susceptible to such effects than other participant populations such as L2 and monolingual L1 users.

A third consideration in the interpretation of the extended reading time effect observed among the heritage speakers in the present study is the self-paced reading method that was employed to measure language processing. Self-paced reading appears to be particularly conducive to delayed or *spillover* effects (Just et al., 1982; Frank et al., 2013), in which a reading time difference caused by a critical word or phrase in the stimuli carries over to the following word or phrase. One reason why spillover effects might be especially common is that self-paced reading does not allow participants to reread prior text, which is very much a part of normal reading. Moreover, heritage speakers can show generally higher rates of rereading, or regressive eye movements, than monolingual L1 users (Parshina et al., 2021), so they may be more affected by self-paced reading during language processing experiments.

As outlined in the literature review sections of this paper, several previous studies have used the self-paced reading method with heritage speakers. Some of these segmented the stimulus sentences in a way that did not yield detailed enough data to observe spillover (Jegerski et al., 2016; Keating et al., 2016) or did not observe any online effects with the potential for spillover (Jegerski, 2018b), but both of the prior investigations that were able to gauge spillover reported extended reading time effects that occurred on both the critical stimulus region and the following region. In one case, this was with phrase-by-phrase self-paced reading (Jegerski, 2018c), as in the present study, and in the other it was with word-by-word self-paced reading (Foote, 2011). In both cases, the same prolonged effect was displayed by a comparison group of monolinguals (Jegerski, 2018c) or of more prototypical L1 users raised in a majority language context (Foote, 2011). Thus, previous research offers no particular evidence either for or against the supposition that heritage speakers are especially likely to show spillover effects during self-paced reading, although it does serve as a reminder that such effects are common in general, not just with heritage speakers. The present study appears to be the first with heritage speakers in which the comparison group has not shown spillover, which is the best scenario for testing whether heritage speakers are more likely to show such effects.

Looking to the future, it is clear that there is a need for more research using self-paced reading with heritage speakers, in line with the broader motivation for the use of online methods laid out in the introduction and literature review sections of this article. In addition, the present study has suggested that heritage speakers may be especially likely to show spillover effects with self-paced reading, but further research using the method is needed to determine to what extent the findings of this single study may generalize to other samples of heritage speakers and other aspects of sentence processing. In addition, a follow-up study using eyetracking, which is already in progress, could help clarify to what extent shallow processing (Clahsen and Felser, 2006) may underlie the observations of the present study. Specifically, evidence from eyetracking could help tease apart shallow processing from the self-paced reading method, as an effect caused by shallow processing should hold even if the experimental method is changed to eyetracking, whereas an artifact of the self-paced reading method should not.

In conclusion, a primary finding of the present study was that heritage speakers of Spanish exhibited prolonged effects for verb transitivity across two stimulus regions during self-paced reading, whereas a comparison group of more prototypical L1 users raised with Spanish as a majority language displayed the effect only on the

immediate region, with no spillover. Analysis of individual background variables revealed that reading-related metrics predicted the degree of sensitivity to verb transitivity. Three explanations for the apparent susceptibility to spillover effects among heritage speakers were proposed: that it is a characteristic of shallow processing (Clahsen and Felser, 2006), that it is due to underdeveloped reading skill (i.e., reading speed, more frequent rereading, and other skills that form the basis for self-ratings), and that it is an artifact of the self-paced reading method. The latter two possibilities are especially consistent with a role for reading skill in these results, although the three explanations are not mutually exclusive, so they might all apply to varying degrees or in different contexts.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Texas Tech University Protection of Human Subjects

Committee. The participants provided their written informed consent to participate in this study.

Author contributions

Both authors made an equal contribution to this work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Front. Psychol.* 4:328. doi: 10.3389/fpsyg.2013.00328
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Benmamoun, E., Montrul, S., and Polinsky, M. (2010). *White paper: Prolegomena to heritage linguistics*. Los Angeles: National Heritage Language Resource Center.
- Berghoff, R. (2020). Syntactic processing in English–Afrikaans bilinguals [doctoral dissertation, Stellenbosch University]. SUNScholar research repository. Available at: <http://hdl.handle.net/10019.1/108292> (accessed April 19, 2023).
- Bice, K., and Kroll, J. F. (2021). Grammatical processing in two languages: how individual differences in language experience and cognitive abilities shape comprehension in heritage bilinguals. *J. Neurolinguistics* 58:100963. doi: 10.1016/j.jneuroling.2020.100963
- Bolger, P. A., and Zapata, G. C. (2011). Psycholinguistic approaches to language processing in heritage speakers. *Heritage Lang. J.* 8, 1–29. doi: 10.46538/hlj.8.1.2
- Brothers, T., Hoversten, L. J., and Traxler, M. J. (2021). Bilinguals on the garden-path: individual differences in syntactic ambiguity resolution. *Biling. Lang. Cogn.* 24, 612–627. doi: 10.1017/S1366728920000711
- Carreira, M., and Kagan, O. (2011). The results of the National Heritage Language Survey: implications for teaching, curriculum design, and professional development. *Foreign Lang. Ann.* 44, 40–64. doi: 10.1111/j.1944-9720.2010.01118.x
- Cedrus Corporation. (2007). *SuperLab (Version 4.5)*. San Pedro, CA: Cedrus Corporation.
- Christianson, K. (2016). When language comprehension goes wrong for the right reasons: good-enough, underspecified, or shallow language processing. *Q. J. Exp. Psychol.* 69, 817–828. doi: 10.1080/17470218.2015.1134603
- Clahsen, H., and Felser, C. (2006). Grammatical processing in language learners. *Appl. Psycholinguist.* 27, 3–42. doi: 10.1017/S0142176406060024
- Davies, M. (2005). *Frequency dictionary of Spanish: Core vocabulary for learners*. New York: Routledge.
- Eriksen, B. A., and Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Percept. Psychophys.* 16, 143–149. doi: 10.3758/BF03203267
- Foot, R. (2011). Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Appl. Psycholinguist.* 32, 187–220. doi: 10.1017/S0142176410000342
- Frank, S. L., Fernandez Monsalve, I., Thompson, R. L., and Vigliocco, G. (2013). Reading time data for evaluating broad-coverage models of English sentence processing. *Behav. Res. Methods* 45, 1182–1190. doi: 10.3758/s13428-012-0313-y
- Frazier, L., and Fodor, J. D. (1978). The sausage machine: a new two-stage parsing model. *Cognition* 6, 291–325. doi: 10.1016/0010-0277(78)90002-1
- Freneck-Mestre, C., and Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *Quar. J. Exp. Psychol.* 50, 119–148. doi: 10.1080/027249897392251
- Fuchs, Z. (2021). Facilitative use of grammatical gender in heritage Spanish. *Ling. Approach. Biling.* 12, 845–871. doi: 10.1075/lab.20024.fuc
- Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. (1997). The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *J. Mem. Lang.* 37, 58–93. doi: 10.1006/jmla.1997.2512
- Gratton, G., Coles, M., and Donchin, E. (1992). Optimizing the use of information: strategic control of activation of responses. *J. Exp. Psychol. Gen.* 121, 480–506. doi: 10.1037/0096-3445.121.4.480
- Jaeger, T. F. (2008). Categorical data analysis: away from ANOVAs (transformation or not) and towards logit mixed models. *J. Mem. Lang.* 59, 434–446. doi: 10.1016/j.jml.2007.11.007
- Jegerski, J. (2012). The processing of temporary subject-object ambiguities in native and near-native Mexican Spanish. *Biling. Lang. Cogn.* 15, 721–735. doi: 10.1017/S1366728911000654
- Jegerski, J. (2014). “Self-paced reading,” in *Research methods in second language psycholinguistics*. eds. J. Jegerski and B. VanPatten (New York: Routledge), 20–49.
- Jegerski, J. (2018a). “Psycholinguistic perspectives on Spanish as a heritage language,” in *Routledge handbook of Spanish as a heritage/minority language*. ed. K. Potowski (New York, NY: Routledge).
- Jegerski, J. (2018b). Sentence processing in Spanish as a heritage language: relative clause attachment in early bilinguals. *Lang. Learn.* 68, 598–634. doi: 10.1111/lang.12289
- Jegerski, J. (2018c). The processing of the object marker a by heritage Spanish speakers. *Int. J. Biling.* 22, 585–602. doi: 10.1177/1367006916681083
- Jegerski, J., Keating, G. D., and VanPatten, B. (2016). On-line relative clause attachment strategy in heritage speakers of Spanish. *Int. J. Biling.* 20, 254–268. doi: 10.1177/1367006914552288
- Jegerski, J., and Sekerina, I. A. (2020). The processing of input with differential object marking by heritage Spanish speakers. *Biling. Lang. Cogn.* 23, 274–282. doi: 10.1017/S1366728919000087
- Jegerski, J., and Sekerina, I. (2021). “The psycholinguistics of heritage languages,” in *The Cambridge handbook of heritage languages and linguistics*. eds. S. Montrul and M. Polinsky (Cambridge, UK: Cambridge University Press), 449–470.
- Just, M. A., Carpenter, P. A., and Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol. Gen.* 111, 228–238. doi: 10.1037/0096-3445.111.2.228

- Keating, G. D. (2022). The effect of age of onset of bilingualism on gender agreement processing in Spanish as a heritage language. *Lang. Learn.* 72, 1170–1208. doi: 10.1111/lang.12510
- Keating, G. D., and Jegerski, J. (2015). Experimental designs in sentence processing research: a methodological review and user's guide. *Stud. Second. Lang. Acquis.* 37, 1–32. doi: 10.1017/S0272263114000187
- Keating, G. D., Jegerski, J., and VanPatten, B. (2016). Online processing of subject pronouns in monolingual and heritage bilingual speakers of Mexican Spanish. *Biling. Lang. Cogn.* 19, 36–49. doi: 10.1017/S1366728914000418
- Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (2014). lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). Available at: <https://cran.r-project.org/web/packages/lmerTest/index.html> (accessed April 19, 2023).
- Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York, NY: Routledge.
- Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2018). Emmeans: Estimated marginal means, aka least-squares means. Available at: <https://cran.r-project.org/web/packages/emmeans/index.html> (accessed April 19, 2023).
- Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)
- Martohardjono, G., Phillips, I., Madsen, C. N., and Schwartz, R. (2017). “Cross-linguistic influence in bilingual processing: an ERP study,” in *Proceedings of the 41st Annual BOSTON University Conference on Language Development*. eds. M. LaMendola and J. Scott (Somerville, MA: Cascadia Press), 452–465.
- McCormick, T. (2020). Early and emergent bilinguals: the role of cognitive control in the processing of structural ambiguity [doctoral dissertation, Georgetown University]. Georgetown University Institutional Repository. Available at: https://repository.library.georgetown.edu/bitstream/handle/10822/1059656/McCormick_georgetown_0076D_14469.pdf?sequence=1&isAllowed=y (accessed April 19, 2023).
- Montrul, S., and Slabakova, R. (2003). Competence similarities between native and near-native speakers: an investigation of the preterite/imperfect contrast in Spanish. *Stud. Second. Lang. Acquis.* 25, 351–398. doi: 10.1017/S0272263103000159
- Montrul, S., Foote, R., and Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition. *Lang. Learn.* 58, 503–553. doi: 10.1111/j.1467-9922.2008.00449.x
- Parshina, O., Laurinavichyute, A. K., and Sekerina, I. A. (2021). Eye-movement benchmarks in heritage language reading. *Biling. Lang. Cogn.* 24, 69–82. doi: 10.1017/S136672892000019X
- Parshina, O., Lopukhina, A., and Sekerina, I. A. (2022). Can heritage speakers predict lexical and morphosyntactic information in reading? *Languages* 7:60. doi: 10.3390/languages7010060
- Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245
- R Development Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychol. Bull.* 124, 372–422. doi: 10.1037/0033-2909.124.3.372
- Roberts, L., and Felser, C. (2011). Plausibility and recovery from garden paths in second language sentence processing. *Appl. Psycholinguist.* 32, 299–331. doi: 10.1017/S0142716410000421
- Rossi, E. (2021). Individual differences in pronoun processing in heritage speakers of Spanish: Data from ERPs [paper presentation]. International symposium on bilingualism (ISB13), Warsaw, Poland.
- Sekerina, I. A., and Trueswell, J. C. (2011). Processing of contrastiveness by heritage Russian bilinguals. *Biling. Lang. Cogn.* 14, 280–300. doi: 10.1017/S1366728910000337
- Solon, M., Park, H., Dehghan-Chaleshtori, M., Carver, C., and Long, A. (2022). Exploring an elicited imitation task as a measure of heritage language proficiency. *Stud. Second. Lang. Acquis.* 44, 1095–1123. doi: 10.1017/S0272263121000905
- Staub, A. (2007). The parser doesn't ignore intransitivity, after all. *J. Exp. Psychol. Learn. Mem. Cogn.* 33, 550–569. doi: 10.1037/0278-7393.33.3.550



OPEN ACCESS

EDITED BY

Theodoros Marinis,
University of Konstanz, Germany

REVIEWED BY

Jeanine Treffers-Daller,
University of Reading, United Kingdom
Chunxuan Shen,
The University of Queensland, Australia
Ethan Kutlu,
The University of Iowa, United States

*CORRESPONDENCE

Aleksandra Tomić
✉ aleksandra.tomic@uit.no

RECEIVED 25 December 2022

ACCEPTED 26 April 2023

PUBLISHED 23 May 2023

CITATION

Tomić A, Rodina Y, Bayram F and De
Cat C (2023) Documenting heritage language
experience using questionnaires.
Front. Psychol. 14:1131374.
doi: 10.3389/fpsyg.2023.1131374

COPYRIGHT

© 2023 Tomić, Rodina, Bayram and De Cat.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Documenting heritage language experience using questionnaires

Aleksandra Tomić^{1*}, Yulia Rodina¹, Fatih Bayram¹ and
Cécile De Cat^{1,2}

¹Department of Language and Culture, UiT The Arctic University of Norway, Tromsø, Norway, ²School of Languages, Cultures and Societies, University of Leeds, Leeds, United Kingdom

Introduction: There exists a great degree of variability in the documentation of multilingual experience across different instruments. The present paper contributes to the “methods turn” and individual differences focus in (heritage) bilingualism by proposing a comprehensive online questionnaire building on existing questionnaires and the experience of using them to document heritage bilingualism: the Heritage Language Experience (HeLEx) online questionnaire. HeLEx is validated against and contrasted to an extended version of the Language and Social Background Questionnaire designed for heritage speakers (HSs), LSBQ-H.

Methods: We compare data elicited with both questionnaires in turn from a group of Turkish HSs ($n=174$, mean age=32). Our validation focuses on traditional language background variables, including language exposure and use, language proficiency, language dominance, as well as a more novel measure of language entropy. The analyses are based on a subset of key questions from each questionnaire that capture language experience for up to five languages, four modalities, and five social contexts. In a subsequent set of analyses, we explore the impact of different types of response scales, response mechanisms, and manners of variable derivation on the informativity of the data they can provide, in terms of the scope, granularity and distributional properties of the derived measures.

Results and Discussion: Our results show that both HeLEx and LSBQ-H are successful at detecting the important distributional patterns in the data and reveal a number of advantages of HeLEx. In the discussion, we consider the impact of methodological choices regarding question phrasing, visual format, response options, and response mechanisms. We emphasize that these choices are not trivial and can affect the derived measures and subsequent analyses on the impact of individual differences on language acquisition and processing.

KEYWORDS

heritage language, individual differences, validation, language experience, questionnaire design, language entropy

Introduction

In the last decades, it has become increasingly clear that the sociocultural and psycholinguistic experiences of multilingual individuals play a central role in shaping the diversity and variability of their linguistic and cognitive performance (cf., [Green and Abutalebi, 2013](#); [Abutalebi and Green, 2016](#); [Titone and Tiv, 2022](#) for an overview). To understand the complexities of multilingualism, research has focused on identifying the key experience factors and their mediating role in characterizing multilingual language use, development, and cognition (e.g., [Marian et al., 2007](#); [Luk and Bialystok, 2013](#); [Grosjean, 2015](#); [Anderson et al., 2018](#); [Bayram et al., 2019](#); [Lloyd-Smith](#)

et al., 2020; Serratrice and De Cat, 2020; De Cat et al., 2023). Operationalizing multilingualism has become a line of research in itself, aiming to optimize the way we document and quantify the parameters of multilingual experiences. A number of instruments are available to inform this process, including ALDeQ (Alberta Language and Development Questionnaire, Paradis et al., 2010), ALEQ (Alberta Language Environment Questionnaire, Paradis, 2011), LEAP-Q (Language Experience and Proficiency Questionnaire, Marian et al., 2007), BiLEC (Bilingual Language Experience Calculator, Unsworth, 2013), BLP (Bilingual Language Profile, Birdsong et al., 2012; Gertken et al., 2014), BSWQ (Bilingual Switching Questionnaire, Rodriguez-Fornells et al., 2012), LSBQ (Language and Social Background Questionnaire, Luk and Bialystok, 2013; Anderson et al., 2018), LHQ 3.0 (Language History Questionnaire, Li et al., 2020), PaBiQ (Parental Bilingual Questionnaire, COST Action IS0804, 2011), and Q-BEx (Quantifying Bilingual Experience, De Cat et al., 2023). To some extent, these tools build on each other (e.g., PaBiQ builds on the ALDeQ and ALEQ, Q-BEx builds on the BiLEC, PaBiQ and ALEQ), reflecting advances in bilingualism research, both from a conceptual as well as a methodological point of view.

As highlighted in some recent review papers, there is a great degree of variability in the documentation of multilingual experience across different tools (Kaščelan et al., 2022; Rothman et al., 2023). While some are designed to estimate experience in early and late childhood (e.g., BiLEC, Q-BEx, and PaBiQ), others target adults (e.g., LSBQ). Crucially, the available instruments vary in how they capture the depth of bilingual experience since the specific components (exposure, domains of use, proficiency, dominance, etc.) are represented and measured with different levels of detail. There is variability in the set of communicative contexts considered, in the granularity of information about interlocutors and activities in each language, in the life periods documented, in whether language mixing is documented, and in whether attitudes are documented.

The recent “methods turn” in bilingualism research has brought to light issues of comparability of supposedly equivalent measures derived from different questionnaires (for in-depth discussion see Surrian and Luk, 2017; de Bruin, 2019; Kaščelan et al., 2022). This is due to variability in how the constructs of interest are operationalized (e.g., does the amount of exposure take into account the amount of time spent with different interlocutors?), but also in how the questions and response options are formulated. For instance, whether the amount of exposure to each language is recorded on a percentage scale or a Likert scale, whether the points on the Likert scale are labeled, and if so, whether they are labeled with numbers or qualifying terms (e.g., “rarely,” “most of the time”).

Beyond issues of documentation (i.e., what do we ask about and in what level of detail?), there is also variability in how the data is processed. For example, BiLEC, ALDeQ, PaBiQ, and Q-BEx propose specific algorithms to generate composite measures of children’s language experience. Quantity-focused measures include current exposure and use (adjusted according to the amount of time the child spends with different interlocutors or in different contexts) and cumulative exposure and use. Quality-focused measures include composite scores reflecting the diversity of the language experience in terms of interlocutors (e.g., the number of native or non-native speakers providing input, as well as different interlocutors with whom the child communicates exclusively in a given language) or contexts of use. For example, BLP offers an algorithm which automatically

calculates the score for language dominance based on 19 questions distributed across four modules (language history, use, proficiency, and attitudes), which ranges from -218 to $+218$. The extreme values represent dominance in one vs. the other language and the middle values represent more balanced bilingualism.

Recently, language entropy has been proposed as a new measure of language experience (Gullifer and Titone, 2020), inspired by previous work on language mode and social diversity of language use (Grosjean, 2001, 2015). The concept originates in Shannon’s (1948) theory which defines entropy as a measure of information content and uncertainty. Entropy was previously used in psycholinguistics (Hale, 2003; del Prado Martín et al., 2004; Levy, 2008) and neurocognition (Gullifer et al., 2018). In the context of multilingualism, language entropy is derived from estimates of exposure to different languages in different social/communicative contexts. It can be interpreted as a measure of social diversity, indexing the level of non-uniformity in the daily usage of two or more languages across contexts: high entropy scores are indicative of high language diversity in a given communicative context and therefore low language certainty, while low entropy scores are indicative of low diversity and comparatively higher language certainty. For instance, a context in which the individual regularly interacts in their multiple languages in a balanced manner would have high entropy. Another context in which the individual predominantly interacts in a single language would have low entropy. Bilinguals with high language entropy experience a greater number of language states across their communicative contexts than bilinguals with low entropy. Therefore, they may experience less certainty about which language they will be exposed to in a given context. Contrary to the measure of language dominance, language entropy is not indicative of *which* language takes precedence in a given context. It therefore is a valuable addition to other, established measures.

The LSBQ focuses on bilinguals’ language usage patterns, in different contexts and with different individuals in daily life. The LSBQ goes further than most tools, as it aims to derive a *unique* composite score estimating the degree of bilingualism (of young adults). The composite score it generates operationalizes an important dimension in recent theorizing about the bilingualism effect, namely the role of interactional context in determining the degree of bilingualism an individual possesses. It is also possible to calculate multiple composite scores based on the distribution of how a bilingual uses each language in different domains of life (home vs. work vs. social contexts). The composite score reflects the extent of proficiency and use of languages other than the societal language, both within and outside of the home. Based on this, bilinguals can be assigned into groups defined along a monolingual-to-bilingual continuum: a composite score of less than -3.13 would categorize one as monolingual, while having a score above 1.23 would be regarded as being bilingual. The most recent version of the LSBQ (Anderson et al., 2018) follows the footsteps of its predecessor version (Luk and Bialystok, 2013) but bears similarities to other existing tools such as the LEAP-Q (Marian et al., 2007) and the LHQ 2.0 (Li et al., 2006, 2014). This latest version of LSBQ has been validated across a large sample size ($n=408$) of young adults using exploratory factor analysis (EFA).

This brief review shows that the documentation and operationalization of bi-/multilingualism is an incremental endeavor, reflecting research development in terms of scope and in terms of methods. New and more precise measures become necessary (e.g., various aspects of language mixing, language entropy), and existing measures are revisited to enhance their reliability.

The current paper fits within this incremental tradition, by presenting a new questionnaire to inform heritage bilingualism research: the Heritage Language Experience questionnaire (HeLEx). Our initial intention had been to “simply” adapt LSBQ for online data collection, as it is a validated, established, fairly exhaustive, and one of the most commonly used questionnaires to qualify and quantify heritage bilingual experience. However, we found ourselves adding and modifying questions in an attempt to augment and optimize LSBQ to meet our research needs: we added several components focusing on language attitudes, code-switching attitudes and behavior, decided to separate the use of and exposure to languages, and to ask about quantity and quality of language experience in absolute terms. To minimize the data wrangling required to derive language entropy measures and to facilitate the derivation of other composite measures, questions are asked in relation to the same five social contexts of interaction throughout the questionnaire. The formulation of questions and response scales is informed by the psychometric literature (Dillman et al., 2014, 2016). Furthermore, we attempted to remain as neutral as possible in the question text and response options, to reduce any potential emotional discomfort associated with the often minoritized or stigmatized status of heritage language, within the immigrant-origin community and the larger society. For example, code-switching is a frequent, yet often stigmatized language behavior in heritage language communities and beyond. When probing the frequency of use and exposure to code-switching, the question preamble explains that research shows that it is a frequent and normal behavior in many multilingual communities. When probing code-switching attitudes of our participants and other people in their lives, the negative attitude option was carefully chosen not to attach any strong or objective negative value to code-switching, resulting in the following option list: “It should be avoided,” “I do not have an opinion,” “It’s okay,” “I do not know.”

This study is an empirical evaluation of these modifications to LSBQ. Our first objective is to validate HeLEx against (an augmented version of) the LSBQ (i.e., LSBQ-H), by comparing data elicited with both questionnaires from the same group of HSs (i.e., Turkish HSs living in Germany), first with LSBQ-H and then with HeLEx several months to a year later. The validation focuses on traditional background variables, such as language exposure and use, language proficiency, language dominance, as well as a more novel measure of language entropy. Comparing questionnaire data for the same participants allows us to shed light on the impact of different types of response scales, both on the distribution of the raw data, and on the distribution of derived measures. Our second objective is to explore the informativity of each questionnaire (HeLEx vs. LSBQ-H) in terms of the scope and granularity of the derived measures.

Methodology

LSBQ-H: an extended version of LSBQ for heritage speakers

The original LSBQ comprises three sections: (1) social background/demographic information, including age, education, country of birth, immigration, and parents’ education as a proxy of

SES; (2) information about language background, i.e., questions about which language(s) the participant can understand and/or speak, age of acquisition, etc., as well as questions about self-rated proficiency for speaking, understanding, reading and writing the indicated languages; and (3) information about community language use, including language use in different life stages (infancy, preschool age, primary school age, and high school age), language use and code-switching in specific contexts (with different interlocutors), in different situations (home, school, work, and religious activities), and for different activities (reading, social media, watching TV and browsing the internet).

As a first step, an expanded version of the LSBQ was created (Bayram, 2021, unpublished) to optimize it for the documentation of heritage bilingualism, by adding and expanding questions to document the following aspects of HS’ experience in more detail: (i) HL language training and formal education in HL; (ii) changes in language experience over the lifetime (documenting changes for up to three languages, across several periods); (iii) language profile of partners or cohabitants; (iv) parental language, immigration history and education in each language; (v) visits to the country of HL origin and size of HL community in the current society; (vi) different patterns of code-switching; and (vii) community language attitudes.

The rest of the questionnaire was implemented as in the original¹ with a few minor exceptions: (i) for the frequency of use by modality question, “Of the time you spend engaged in each of the following activities, how much of that time is carried out in [language]?” response options changed from “None,” “Little,” “Some,” “Most,” “All” to “None,” “Very little,” “50–50,” “Most,” “All”; (ii) the response options for the HL use proportion, out of use of HL and another language, in different contexts were changed to acknowledge that the participants likely speak more than two languages, so the proportion of HL use is now estimated out of use of all languages. The LSBQ-H was administered online, using the Gorilla questionnaire builder, as a part of a larger study on Turkish as a HL in Germany. In the transfer to the online version of the questionnaire, the LSBQ-H attempted to replicate the LSBQ-on-paper visual format.

The Heritage Language Experience questionnaire

In creating the Heritage Language Experience questionnaire (HeLEx), three main principles were adopted: (i) expand coverage to capture the multi-faceted HL experience, (ii) adopt recommendations from the psychometric literature to optimize response scales, (iii) keep frames of reference as constant as possible, e.g., inquire about the same language contexts within and across questions.

The HeLEx questionnaire was developed to capture the individual language experience of heritage bilinguals primarily. Therefore, it includes most variables which could potentially affect HL acquisition and processing while minimizing questionnaire completion time. The questionnaire can also be used by immersed bilinguals in general. This

¹ LSBQ-H also narrows the initial subsection on health to language disorders only, as the rest, e.g., medicine currently being taken and handedness, is relevant for screening EEG participants.

is useful since the language experience of the first-generation of immigrants providing input to heritage bilinguals could be captured using the same questionnaire.

The full questionnaire contains the following modules:

- Demographics,
- Visits to HL country,
- Proficiency in five languages in four modalities (speaking, listening, reading, and writing),
- Language experience for five languages in four modalities and in tech-related activities,
- Diversity and quantity of HL and societal language (SL) experience in five different social contexts (Home, External Family, Work/School, Leisure, Community),
- Proportion of HL speaking and listening in five social contexts,
- Code-switching types and frequency in five social contexts,
- Attitudes (both personal and community) on code-switching in five social contexts,
- Historical use of HL and SL in self-defined periods,
- HL literacy training,
- Personal language attitudes,
- Community language attitudes.

Contrary to LSBQ, the questionnaire does not assume the existence of particular interlocutors (e.g., parents, siblings), and does not require making any assumption about household composition or any other context. We introduced these indirectly grouped and consistent frames of reference for contexts across questions to reduce thinking time when filling the questionnaire. It also allows the straightforward combination of information across questions during data wrangling. For example, to probe the diversity and quality of the HL input, HeLex asks, for each context, how many interlocutors the participant spends any time within a typical week, how many of them have good HL proficiency, and how many of them are dominant in HL. Helping respondents maintain the same context/inquiry focus in mind while responding was achieved by using matrices (tables) for the majority of questions. The matrix questions target sets of contexts and/or sets of languages, with each field in the matrix usually containing a dropdown menu with response options (see Figure 1 for an illustration). This would be difficult to achieve in a paper version, as the response options could not fit on a page.

Another affordance of the online interface is the use of sliders for answers expressing proportions (e.g., HL use) or level of agreement (e.g., attitudes). This was intended to be more visually intuitive by avoiding overt quantization, hence reducing cognitive burden. The potentially more fine-grained responses might also capture more accurately the individual reality of HL experience.

As in LSBQ (Dunning et al., 2004; Anderson et al., 2018), we allowed some level of redundancy in some questions probing key concepts for triangulation (i.e., questionnaire-internal validation). For instance, to probe language experience in different social contexts, HeLex uses both response scales based on natural metrics (number of days, number of hours) and proportion-based response scales (e.g., sliders ranging from “HL only” to “other languages only”). In contrast, the LSBQ primarily uses ratios (e.g., HL vs. other most used/proficient languages). HeLex uses both a question in absolute time terms (“How many days per week do you meet speakers

of [HL]/[SL],² at least some of them?” “How many hours do you typically spend together with them?”) as well as relative time indicators of HL and other language use using slider scales ranging from “HL only” to “other languages only” spoken and heard in the five contexts.

Since it targets heritage bilinguals, the questionnaire implementation in Gorilla is optimized to be translated and offered in both the HL and SL.³ The Gorilla HeLex questionnaire is freely available for use at the Gorilla Open Materials page.⁴ It is also developed for use in Qualtrics, to ensure wide availability. The Gorilla implementation of the HeLex questionnaire is accompanied by an R script (available at the OSF repository <https://osf.io/mkjax>) which provides automatized numerical transformation of textual responses and calculation of derived variables, including, among others, language entropy (using the R package `languageEntropy`, release v1.0.1c, Gullifer and Titone, 2018). This set of derived variables captures the multi-faceted HL experience.

Each dropdown menu in HeLex began with the prompt “select” and many menus included “does not apply” option at the end, to easily identify non-responses (as dropdown menu widgets in Gorilla Task Builder cannot be set to require a response). Instructions at the beginning of the questionnaire stated that participants should always select an appropriate dropdown option even if they believe the question does not apply to them, and that questions left on “select” will be considered not responded to. Questions with the “select” as the response were quantized, i.e., numerically transformed, to NaNs (“not a number”).

Despite many additional questions, the average completion time for HeLex was 10.5 min (sd = 7.02) for the 227 Turkish-German bilinguals who took the questionnaire, and around 11 min (sd = 7.61) for the 174 participants whose data was analyzed, as opposed to the average LSBQ-H completion time of 5.45 min for the 174 participants included in the analysis. We believe the affordances of the online implementation and keeping fewer contexts consistent across the questionnaire contributed to the relatively short completion time for HeLex considering the number of questions, but we cannot exclude factors such as population characteristics (uniform or extreme experience, clear intuitions on language use and attitudes) and previous engagement with language experience questionnaires. In any case, the focus on individual variables in language processing means that questionnaires now must be on equal footing with tasks in terms of importance and therefore time commitment for participants (within reason).

HeLex validation methods

Our aim is to validate HeLex by comparing its derived measures with those from the LSBQ-H, using data collected from the exact same

² Square brackets [] are used as placeholders in this text. They had been replaced with the appropriate language names in the implemented questionnaire versions.

³ The questionnaire consists of referenced fields created in the Gorilla task builder to which actual question text is fed through a simple Excel or .csv sheet, which in turn can be easily translated and fed back into the fields to create a different language version.

⁴ <https://app.gorilla.sc/openmaterials/605087>

A

Lütfen her dilde ne kadar iyi konuştuğunuzu, anladığınızı, okuduğunuzu ve yazdığınızı değerlendirin. Türkçe ve Almanca dışında başka diller konuşuyorsanız, boş kutulara diğer dillerin adını girin.

	Türkçe	Almanca			
Bu dilde ne kadar iyi konuşuyorsunuz? (Dilin yetisi bir formu yoksa, lütfen "İlgili değil" seçeneğini seçin.)	seçin	seçin	seçin	seçin	seçin
Bu dilde ne kadar iyi yazabiliyorsunuz? (Dilin yazısı bir formu yoksa, lütfen "İlgili değil" seçeneğini seçin.)	seçin	seçin	seçin	seçin	seçin
Bu dili ne kadar iyi konuşabilirsiniz?	seçin	seçin	seçin	seçin	seçin
Bu dili ne kadar iyi anlayabilirsiniz?	seçin	seçin	seçin	seçin	seçin

B

Türkçe Yeterlik / Deneyim

Oldukça yetkin bir konuşmacının performansına göre, aşağıdaki faaliyetler için Türkçe yeterlik seviyenizi 0 (Yeterlik yok) - 10 (Yüksek yeterlik) arasında derecelendirin

Konuşma

Yeterlik yok
5
 Yüksek yeterlik

Dinleme

Yeterlik yok
5
 Yüksek yeterlik

Okuma

Yeterlik yok
5
 Yüksek yeterlik

Yazma

Yeterlik yok
5
 Yüksek yeterlik

C

Aşağıdaki faaliyetlerin her birinde harcadığınız zamanın ne kadarını Türkçe kullanılarak yapılır?

Konuşma

hiç çok azı yarı yarıya çoğu hepsi

Dinleme

hiç çok azı yarı yarıya çoğu hepsi

FIGURE 1

(A; top left). An example of a matrix visual format question with dropdown menu response mechanism from HeLex, probing the proficiency in HL (Turkish), SL (German), and 3 additional languages (columns), in speaking, understanding, reading, and writing (rows). (B; top right). Single column visual format example with sliders as response mechanism from LSBQ-H, probing the proficiency in Turkish in 4 modalities. (C) Example of an LSBQ-H question using clicking on a button response mechanism probing the relative frequency of HL use.

population of heritage speakers using each questionnaire in turn. Both questionnaires are available in spreadsheet format from the OSF repository at <https://osf.io/mkjax>.

The first set of analyses aims to ascertain whether the two questionnaires reliably capture the same reality, insofar as the distribution of the resulting measures is sufficiently similar. The objective threshold we use for “sufficient similarity” is the absence of statistically significant questionnaire effect on the distribution of key variables. These key variables are those frequently used as predictors in the bi-/multilingualism literature: language exposure and use, language proficiency, language dominance, and language entropy. The second set of analyses explores the informativity of each questionnaire, in terms of scope and granularity of the derived measures. Both sets of analyses consider methodological choices, in terms of question phrasing, visual format, response options, and response mechanism, and their effect on measures. We then discuss the implications of our findings.

Participants

Two hundred and twenty-seven (227) Turkish-German HSs took both LSBQ-H and HeLex. The LSBQ-H data for 13 participants who took HeLex was not available, so they were excluded. Out of these 214 participants, 40 participants were excluded due to not having data on language use for most social contexts. This was likely due to a glitch, as the same sequence of questions was missing across participants. The mean age of the participants included in the analysis (66 men, 108 women) was 32.08 (sd = 4.67, range 23–47). The vast majority (168) was born in Germany, whereas six moved

to Germany at or before the age of 3. Out of 116 participants who reported living with someone, only two lived with partners who did not speak Turkish. In most other cases, the partner spoke Turkish as their first or native language. When asked by LSBQ-H on the size of the HL-speaking community, most of the participants (144) reported having an intermediate to massive community. Twenty-one participants reported having reading and/or writing lessons in Turkish in mainstream German public schools and six reported having had additional reading and/or writing lessons in Turkish. Thus, it seems that our participants belong to a thriving, connected HS community, with many opportunities for HL acquisition, use, and maintenance. It is likely that snowball sampling and self-selection further ensured that the sample includes participants with high use and proficiency of HL.

Procedure

The questionnaires were administered in Turkish. The English versions used in this paper consist in (i) the original English version of HeLex and (ii) a back-translation into English of the Turkish LSBQ-H. LSBQ-H was administered first, within a larger study, and HeLex around half a year to a year later.

Questions used to derive the measures of interest

The subset of questions used in this validation analysis are those required to derive the variables of interest, as explained above. These questions were asked differently by the two questionnaires, in terms of content of the questions, response elicitation mechanism, and visual format. The main visual format of the selected questions was either a Matrix or a Single column. Unlike the single column question (Figure 1B),

the matrix question presents the same questions, usually arranged in the rows of the first column, for several different languages or contexts, with one column reserved for each language or context (Figure 1A).

The response mechanism could either be selecting an option from a dropdown menu (Figure 1A), moving the slider tip to the desired point on scale (Figure 1B), or clicking on a button (Figure 1C). Different visual configurations could be combined with different response mechanisms.

Another point of difference is the number of options, whether the options were presented verbally and numerically (or both), and whether the response options included absolute or relative quantities (usually in relation to the other language or languages). The main characteristics of the questions selected to measure the concepts of interest are summarized in Table 3.

Tables 1, 2, 4, 5 below summarize and contrast the questions across questionnaires.

Overall experience in each language in different modalities

HeLex and LSBQ-H versions of the question on frequency of use of five languages in four modalities. As shown in Table 2.

Proportion of HL use by social context

The specifics of the HeLex matrix questions on the quantity, quality, and diversity of HL and SL experience. As shown in Table 4. HeLex and LSBQ-H version of the question(s) on the proficiency in five languages in four modalities. As shown in Table 5.

Proficiency in each language in 4 modalities

The comparison of HeLex and LSBQ-H questions used in the analysis and their characteristics. As shown in Table 1.

Derived measures

HL experience and proficiency in four different modalities

Calculating the scores for HL experience and proficiency in four modalities required minimal derivation, i.e., simple numerical transformation, presented in question summary tables, in case the response options were presented verbally.

Proportion of HL use in different social contexts

HeLex data selection and preparation

HeLex probes the following five contexts:

- Home (including whoever lives in the household)
- External family (family outside the household)
- Work or school
- Local community (shops, organizations, restaurants etc.)
- Leisure (hanging out with friends, roommates, hobbies).

To derive the proportion of HL use in different contexts, we used two questions, detailed in Tables 4, 5. HL use was probed in two ways to compare the effects of different response mechanisms (slider vs. dropdown) and different ways of calculating proportions of HL use (directly from responses in the case of sliders, and by deriving proportion of time of HL exposure from absolute time responses on the quantity of use of HL and SL). The slider question on HL use readily provides the proportion of HL speaking and listening out of all language use in each context, with minimal derivation (0–100 to 0–1.0 transformation). The proportion of the (potential) HL exposure was also calculated from the questions on the diversity and quantity of

TABLE 1 HeLex and LSBQ-H version of the question(s) on the proficiency in up to five languages in four modalities.

	HeLex	LSBQ-H
Instructions	Please rate how well you speak, understand, read and write in each language. Enter the name of other languages in the boxes, if you speak other languages than [HL] and [SL].	Rate your [HL] proficiency for the following activities, based on a highly competent speaker's performance level from 0 (No qualification) to 10 (Higher proficiency) for the following activities. Rate your proficiency level in your most familiar/used language outside of [HL] for the following activities, based on a highly competent speaker's performance level from 0 (No qualification) to 10 (Higher proficiency) for the following activities.
Languages probed	[HL], [SL], Language 1, Language 2, Language 3	[HL], most familiar/used language outside of [HL]
Question structure	Matrix: languages across columns; modalities across rows	Two separate questions for each language, not following each other, with modalities vertically ordered
	How well do you speak it?	Speaking
	How well do you understand it?	Listening
	How well do you read in it?	Reading
	How well do you write in it? If the language does not have a written form, please select "not relevant."	Writing
Response mechanism	Dropdown menu	Slider scale
Response options	hardly at all, not very well, pretty well, very well, does not apply	No proficiency (0) – high proficiency (10)
Response options, quantized	1, 2, 3, 4, 0	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

TABLE 2 HeLex and LSBQ-H versions of the question on frequency of use of up to five languages in four modalities.

	HeLex	LSBQ
Instructions	For all languages you use, rate how frequently you use them. Enter additional languages you might speak in addition to [HL] and [SL].	How much of the time you spend doing the following activities is spent using [HL]? How much of the time you spend doing the following activities is spent using other most proficient language?
Languages probed	[HL], [SL], Language 1, Language 2, Language 3	[HL], other most proficient language
Question structure	Matrix: languages across columns; modalities across rows	Two separate questions for each language, not following each other, with modalities vertically listed
	How often do you hear it?	Listening
	How often do you speak it?	Speaking
	How often do you read it?	Reading
	How often do you write in it?	Writing
	How often do you do any computer/technology-related activities in each language? E.g., TV, radio, music, films, websites, games, apps.	This is probed in a different series of questions in LSBQ.
Response mechanism	Dropdown menus	Horizontally ordered buttons
Response options	1. (Almost) never, 2. A few times per year, 3. Once a month, 4. Once a week, 5. A few times per week, 6. Once per day; most days, 7. Several times per day; most days	Never, very little, 50–50, ¹ most, all
Response options quantized	1, 2, 3, 4, 5, 6, 7	1, 2, 3, 4, 5

¹The original response option list from LSBQ is *None, Little, Some, Most, All*.

TABLE 3 The comparison of HeLex and LSBQ-H questions used in the analysis and their characteristics.

Concept	Questionnaire	Visual format	Response mechanism
SL and HL proficiency	HeLex Proficiency	Matrix	Dropdown menus
	LSBQ-H Proficiency	Single column	Sliders
SL and HL experience in diff. modalities	HeLex Experience in diff. modalities	Matrix	Dropdown menus
	LSBQ-H Experience in diff. modalities	Single column	Buttons
SL and HL experience in different social contexts	HeLex Proportion of HL use in social contexts	Matrix	Sliders
	LSBQ-H Proportion of HL use in social contexts	Single column	Buttons
	HeLex Diversity and quantity of input/exposure to HL and SL	Matrix	Dropdown menus

language exposure which probe the time spent with HL and SL speakers in a typical week in absolute terms: number of days spent with HL speakers and SL speakers in each context and the amount of time in hours spent with them in total on a typical day. We calculated the proportion of time that participants spent with HL speakers out of all time spent with speakers of any language in a typical week in the following way:

1. For each context in each language (HL, SL), we multiplied the total number of daily hours spent with speakers of that language by the number of days per week spent with these speakers in each context to obtain interaction hours (In the case there were no HL or SL speakers met in a context the interaction time in that particular language and context was set to 0.). For each language, we then calculated the sum of interaction time in hours across contexts.

2. We calculated the total amount of hours of HL and SL interaction per week, by summing the time spent with HL and SL speakers across contexts.⁵
3. We then calculated the proportions of HL exposure per week by dividing the hours of the exposure to HL by the total hours spent with any speakers per week.

LSBQ-H data selection and preparation

LSBQ-H documents language experience through a mix of by-person and by-context questions (Table 4). While this provides

⁵ This assumes that the speakers of HL and SL are separate speakers and not necessarily bilinguals.

TABLE 4 HeLex and LSBQ-H questions on the proportion of HL experience in social contexts.

Questionnaire	HeLex	LSBQ-H	
Instructions/questions	Think of all interactions in a typical week. For each context consider face to face and online communication. How much do you speak in each language in each context? How much do you hear each language in each context? The more you speak one language the closer you should put the slider to it. If the slider is in the middle that means you speak [HL] and [SL] in equal amounts. Put the slider all the way to the left if you only speak [HL] and no [SL] in this context. Put the slider to the left but not all the way if you mainly speak [HL] but sometimes use [SL] in this context.	Please indicate which languages you speak generally with the following people. (Please leave relevant columns empty if they do not apply)	Please indicate which language(s) you use generally in the following situations. (Please leave relevant columns empty if they do not apply)
Visual format	Matrix: 2 columns (modalities) by 5 rows (contexts)	Single column	
Modalities	Speaking, Hearing	Use (not explicitly defined)	
Contexts	Family in the household; family outside the household; work or school; local community (shops organizations etc.); leisure (hanging out with friends or roommates, hobbies)	Mother–Father; Siblings; Grandpa(s)-Grandma(s); Other relatives; Friends; Partner; Housemates; Neighbors	Home; School; Work; Social activities (spending time with friends, watching movies, etc.); Religious activities; Out-of-school activities (hobbies, sports, volunteer activities, computer games, etc.); Shopping/Restaurant/ Other commercial activities; Health services/Government-public institutions/Banks
Response mechanism	Slider	Buttons (horizontally ordered)	
Response options	0–100 in steps of 1 (slider tip initially presented at 50 mark)	Only [HL], Mostly [HL], Half [HL] half other language(s), ¹ Mostly other language(s), Other language(s) only	
Response options, quantized	0–1 (in steps of 0.1)	1, 0.75, 0.5, 0.25, 0	

¹The original LSBQ response option list is All [language], Mostly [language], Half English half other language, Mostly the other language, Only the other language.

precision and granularity, it is difficult to group the measures into larger meaningful contexts. The proportions of HL vs. SL use in the Home and Work/School contexts in HeLex are comparable to the following LSBQ-H context questions:

- Home HL experience = quantized proportion of HL use in the home out of other languages (0, 0.25, 0.5, 0.75, 1.0).
- Work or School HL experience = the highest value out of Work HL use proportion and School HL use proportion, or simply the one that was responded to.

The External Family, Community, and Leisure contexts were more difficult to reconstruct using the collected LSBQ-H responses. When this question is asked in HeLex for each context, e.g., Leisure, it is understood that participants do not have to have each of the subcontexts represented in their lives equally, e.g., for Leisure, hanging out with friends, roommates, hobbies. Because the wording of the question and the entire questionnaire is geared towards the HL experience, they are simply giving an answer on the use of HL vs. SL in each context as a whole, and likely choosing subcontexts with the highest HL experience representation. In eliciting responses this way, we maximally avoid researcher-imposed definitions of each context.

To group LSBQ-H responses on the use of HL vs. other language(s) with individual persons and different situations into a smaller number of more meaningful subcontexts, we could take the mean of several responses on the use with specific speakers/in specific situations corresponding to a particular subcontext. For example, we could average the response to the question on Social activities (spending time with friends, watching movies, etc.), Religious activities, and Out-of-school/work activities (hobbies, sports, volunteer activities, computer games, etc.), to get at the HL use in the Leisure context. Nevertheless, Religious activities might not be a significant part of each participant's experience during leisure and might not be taken into account while responding to the same question (Leisure) in HeLex. Whatever such participants respond to the question on HL experience during religious activities, unless they skip it, it will distort the participant's social context reality when calculating the HL experience mean. We thus took a conservative approach and reconstructed the HeLex contexts in LSBQ-H calculations using the least controversial subcontexts and transformations, and not penalizing for non-responses for subcontexts.

We attempted to reconstruct the External Family HL experience, a monolithic context defined as “family outside the home” in HeLex, by taking the mean of HL experience with grandparents and other relatives from LSBQ-H, the family members most likely to live outside

TABLE 5 The specifics of the HeLex matrix questions on the quantity, quality, and diversity of HL and SL experience.

Introduction	Think of all the people you interact with in [HL]/[SL] in a typical week in different contexts including face-to-face and online interaction.				
Visual format	Matrix: contexts in each column, question in each row				
Contexts	family in the household; family outside the household; work or school; local community (shops, organizations, etc.); leisure (hanging out with friends, roommates, hobbies)				
Response mechanism	Dropdown menu				
Questions	How many people do you use [HL]/[SL] with?	How many days per week are you with these people (at least some of them)?	On a typical day when you are with these people, how many hours do you spend together in total?	How many of these people speak [HL]/[SL] very well?	For how many of these people is [HL]/[SL] their best language?
Response options	0, 1, 2, 3, 4, 5, 6–8, 9–11, 12–14, 15–17, 18–20, more than 20	not every week, 1, 2, 3, 4, 5, 6, 7	less than 1, 1, 2, 3, 4, 5, 6, 7, 8, more than 8	0, 1, 2, 3, 4, 5, 6–8, 9–11, 12–14, 15–17, 18–20, more than 20	0, 1, 2, 3, 4, 5, 6–8, 9–11, 12–14, 15–17, 18–20, more than 20
Response options, quantized	0, 1, 2, 3, 4, 5, 7, 10, 13, 16, 19, 22	0.5, 1, 2, 3, 4, 5, 6, 7	0.25, 1, 2, 3, 4, 5, 6, 7, 8, 9	0, 1, 2, 3, 4, 5, 7, 10, 13, 16, 19, 22	0, 1, 2, 3, 4, 5, 7, 10, 13, 16, 19, 22

the home. Nevertheless, this is a reach in conclusion, since we cannot for certain know which family members cohabit with participants and which do not. Also, the data for this context behaved differently to others in early plots, suggesting that the approximation was likely not successful. Therefore, we present the comparisons of measures for External Family within HeLex only and not between HeLex and LSBQ-H.

The Local Community HL experience was reconstructed by using the mean HL experience with neighbors and while shopping (key subcontexts the majority of participants should have):

- Local community HL experience = mean of HL use ratio with neighbors and in the local community, i.e., while shopping, visiting restaurants, and other commercial activities.

The Leisure HL experience was reconstructed by using the mean of HL experience during extracurricular and social activities:

- Leisure HL experience = mean of HL use ratio during extracurricular activities, i.e., hobbies, sports, volunteering, playing games, and social activities, i.e., hanging out with friends.⁶

Dominance

We derived two Dominance variables from the data reviewed so far: one based on experience in each language in four modalities, and the other based on proficiency in each language in four modalities. Dominance in each modality was operationalized as a ratio, by dividing the relevant HL measure by the relevant SL measure for each modality.⁷ The overall dominance for both measures was calculated as

the mean of the HL over SL score ratios for the four modalities (i.e., speaking, listening, reading, writing). For the ratio calculations, a value of around 1 indicates that the participant is balanced overall in terms of HL vs. SL (exposure or use), when all modalities are considered. A value above 1 indicates dominance in HL in terms of experience or proficiency. It is important to note that the overall dominance scores for proficiency and experience may hide variation across modality-specific dominance scores.

Language entropy

We used the data on Proportion of HL use in different social contexts derived from HeLex and LSBQ-H responses (section *Derived measures: Language entropy*) and the R package languageEntropy (release v1.0.1c, Gullifer and Titone, 2018) to calculate language entropy for the following contexts:

- Home: Family in the household
- Work or School
- Local community (shops, organizations, restaurants etc.)
- Leisure (hanging out with friends, roommates, hobbies)

For HeLex, we exploit two types of context-specific language experience questions: (i) a question probing the proportional use of HL and other language(s) using sliders (Table 4), and (ii) questions probing experience of SL and HL separately (Table 5). In both cases, the questions are asked about the following contexts: Home, External Family, Work/School, Leisure, Community. For LSBQ-H, the five macro-social contexts were reconstructed from ratios of HL use with individual speakers/in specific situation, as detailed in the section *Derived measures: Proportion of HL use in different social contexts*. The same proportions of use were used to calculate language entropy.

Accounting for the actual proportion of time spent in each context

Traditionally, language experience questionnaires such as LHQ and LSBQ have measured proportions of exposure to the HL language (or use) with a specific individual or in a specific context out of the

⁶ We do recognize there is some overlap between the Local Community and Leisure contexts, yet Leisure presupposes more involved relationships, whereas Local Community refers to more brief, everyday, surface engagement.

⁷ We also derived dominance score by subtracting SL responses from HL responses and dividing them by the sum of HL and SL responses. This way of calculating dominance gives a similar distribution as do ratios, so we do not include it here.

total use or compared to the SL. However, participants may be spending different amounts of time in each context/with each individual. For example, someone might report that they use the HL 45% of the time at home, whereas another respondent may report 90% use in that context. Imagine the first respondent actually spends 8 h per day at home, whereas the second respondent spends only four. This needs to be taken into account to calculate the total amount of HL experience contributed by the home context (which is equivalent for respondents 1 and 2).

To account for this, we developed weights based on the proportion of time spent with speakers of either language in each context out of the total time spent with speakers during a typical week (Diversity and Quantity of HL and SL input question). This can only be done with responses from HeLEx, as LSBQ or LSBQ-H do not provide absolute time response options for language use.

To derive these weights, we first calculated the total amount of time spent with any speakers. We multiplied the typical number of total hours participants spend with HL speakers on a typical day in each context by the number of days per week they meet with these speakers in each context. We then repeated the process for the SL speakers. To get the number of hours spent with speakers of either language per week, we summed the number of hours spent with HL and SL speakers across contexts (this does assume that the HL and SL speakers are separate speakers). We then calculated the proportion of time spent in each context out of the total time spent with anyone per week, adding up to 1. These context weights were then multiplied with the proportions of the HL vs. SL use in each context.

For unweighted scores, the total HL exposure would be approximated by the average of HL exposure across contexts (i.e., the sum of HL proportions per context divided by the number of contexts). For weighted estimates, the total HL exposure approximation would be the across-context sum of HL exposure in each context multiplied by the weight for that context. Therefore, we compare averages of unweighted scores for HL use proportion across contexts and *sums* of weighted scores across contexts.

Diversity of HL interlocutors

HeLEx provides information about the number of interlocutors in each context, the number and proportion of speakers with good proficiency in HL in each context, and the number and proportion of speakers who are dominant in HL in each context (see Table 5). These measures are useful to approximate the quality and diversity or variation in the input. For each context, the proportion of interlocutors with good HL proficiency is calculated by dividing the number of such interlocutors by the total HL interlocutors in that context. The same procedure was used to derive the proportion of interlocutors who are dominant in HL in each context, as well as overall proportions across contexts. Importantly, the calculations include a data-validation check ensuring that the number of HL-dominant or HL-proficient interlocutors in each context does not exceed the total number of HL interlocutors in the context reported by the respondent (this resulted in a negligible data loss for this sample).

Statistical analysis methods

We employed linear regression models for all relevant statistical comparisons (probing main effects of questionnaire, type of

calculation, contexts, modalities, and interactions between them), using `lmer()` (lme4 package, version 1.1-31, Bates et al., 2015) or `lm()` function (stats package, version 3.6.2) in R (version 4.2.2, R Core Team, 2022). Random intercepts for participants were included where supported by the data. In the linear regression models, we applied dummy contrast coding to the questionnaire variable, where HeLEx was the reference level. In the case of Context and Modality variables, which had four or more levels, we applied deviation coding, where the estimate for each context or modality level was made in reference to the mean of means of HeLEx values across contexts or modalities. The variables were recoded and models were rerun where necessary to obtain the estimates for the contexts or modalities initially coded to -1 (e.g., for Work and Writing). The model outputs can be found in the [Supplementary material](#).

Results

Validation

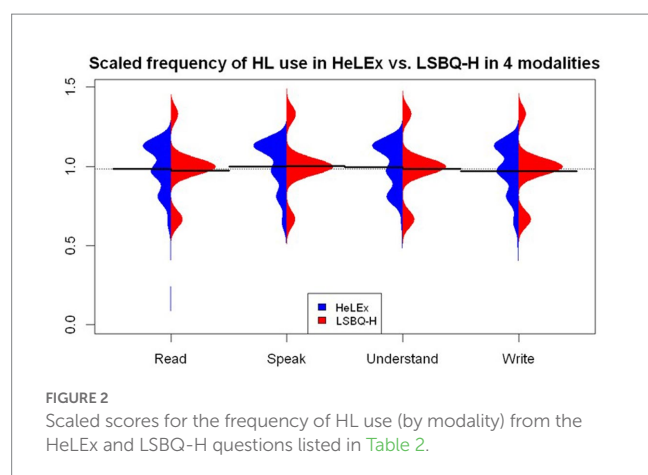
To assess the similarity across the two questionnaires, for each variable of interest, (i) we illustrate the distribution of the variable as per each questionnaire, (ii) we fit a regression model to ascertain if there is a statistically significant questionnaire, context, modality, or manner of calculation effect for the variable in question.

Overall experience in HL in different modalities

The relevant HeLEx and LSBQ-H questions probing the frequency of use of languages in four modalities (reading, writing, speaking, and listening/understanding) are shown in Table 2. LSBQ-H elicits a ratio/percentage of the use of HL vs. “the other most proficient language” which is established earlier in the questionnaire and is usually the SL. HeLEx probes the use of HL and SL, as well as three additional languages in absolute frequency terms (cf., response options). HeLEx also includes a sub-question on the use of the languages in tech-related activities, as they are often multimodal and therefore eschew quantification when probed by using the terms such as writing/reading/listening, etc. LSBQ-H contains a section on the use of HL relative to SL during separate tech-related activities, e.g., social media, TV, etc.

The responses were gathered by the two questionnaires using different Likert scales: different number of points on each scale (7-point scale for HeLEx and 5-point scale for LSBQ-H), and different labels for each scale. To facilitate the comparability of responses, we quantized scores and then scaled them: each score was divided by the standard deviation of the frequency scores distribution from the relevant questionnaire, using the `scale()` base R function. This is a common standardization method, often performed on independent variables with differing scales before entering them into regression models as predictors. A scaled score of 1 denotes that the unscaled score is equal to the standard deviation of the corresponding distribution. The value of standard deviation will of course be directly dependent on the scale. Nevertheless, transforming the scores in terms of the number of standard deviations for each distribution makes the two scores directly comparable. The scaled distributions are shown in Figure 2.

The spread of the distributions seen in Figure 2 is determined by the distribution of actual responses (rather than the range of the



original response scales). As the top and bottom ends of the scale were not used by LSBQ-H respondents, the scores are distributed across three scores, corresponding to the red peaks in Figure 2. By contrast, the HeLex scores feature a more continuous distribution. We entered the main effects of the Questionnaire and Modality, as well as their interactions, into the linear mixed effects model as potential predictors of HL experience scores, with random intercepts for participants.

The results of the linear regression model (Supplementary Table S1, Supplementary material) indicate no statistically significant difference between the estimates obtained from the two questionnaires in each modality when controlling for different scales. Compared with the mean of means of exposure/use across modalities and questionnaires (i.e., the model intercept) in HeLex, writing in the HL is significantly less frequent, and speaking in the HL is significantly more frequent. There was no significant interaction between the questionnaire and modality. We can conclude that the two questionnaires provide a similar distribution of HL experience in different modalities, albeit with a different level of granularity.

Proportion of HL use across social contexts

For comparing the two questionnaires on the measure of the proportion of HL use overall per social context, the question on the proportion of HL speaking out of all languages elicited with sliders from HeLex (0–100 scale, Table 4) was chosen as the closest equivalent to the LSBQ-H question on the proportion of HL use elicited with horizontally ordered buttons (5-point scale, Table 4). Visually (Figure 3), it seems that the proportion of HL use was highest in the Home context, and lowest in the Work or School context. LSBQ-H seems to return lower estimates of HL use, compared with HeLex. We entered the main effects of Questionnaire and Context, as well as their interactions, into the linear regression model as potential predictors of HL use proportions.

The linear regression model (Supplementary Table S2) confirmed that the LSBQ-H-derived estimates of the proportion of HL use are significantly lower than those of HeLex, albeit with a small estimate value. This is likely due to this specific choosing the LSBQ/LSBQ-H “Mostly [HL]” response option when the reality of their experience was between the “Only [HL],” quantized as 1, and “Mostly [HL],” quantized as 0.75. In other words, the participants an option higher than 0.75 and lower than 1, but chose the lower 0.75 to avoid the beginning point of the scale.

In terms of cross-context comparison, the HL use seems to be significantly lower in the Community, Leisure, and Work or School contexts, whereas it is significantly higher in the Home context compared to the mean of means of the HL use proportion across all contexts as measured by HeLex. There was also a significant interaction between the questionnaire (LSBQ-H) and context (Work or School), such that LSBQ-H provided even lower estimates for the Work or School context.

Proficiency in the HL in four modalities

For proficiency, the LSBQ-H response scale is more granular (i.e., 11 points) than that of HeLex (4 points). The LSBQ-H also presents the options numerically, with minimal use of evaluative language, unlike HeLex. Importantly, the LSBQ-H online adaptation includes the slider as the response mechanism, as the most appropriate equivalent of a visual scale with a pronounced mid mark in the paper LSBQ version. The numeric, more granular LSBQ-H slider scale resulted in a more exponential distribution with a high concentration of top-of-the-scale responses compared to the evaluative dropdown menu, as seen in Figure 4. We entered the main effects of the Questionnaire and Modality (i.e., speaking, reading, understanding, and writing), as well as their interactions, into the linear mixed effects model as potential predictors of HL proficiency scores, with random intercepts for participants.

The results of the linear mixed effects model reveal that, despite the visual differences in the distribution, no statistically significant difference was found between the LSBQ-H and HeLex-derived measures of proficiency when controlling for the differences in scale, illustrated in Supplementary Table S3. Expectedly, the sample reports a higher proficiency in HL Speaking and lower proficiency in Writing compared to the mean of means of HL proficiency across all modalities as recorded by HeLex. There was no significant interaction between the questionnaire and modality.

Dominance

Experience-based dominance scores

Figure 5 shows that most people in the sample are balanced bilinguals. The measure calculated with LSBQ-H responses is possibly more discriminatory (with less clustering around 1), but a more diverse population would be needed to assess this. We fitted a linear regression model to assess whether the estimates of Experience-Based Dominance were predicted by the questionnaire used. There was a small but significant effect of the questionnaire, such that LSBQ-H Experience-based HL dominance estimates seem to be higher than those of HeLex (Supplementary Table S4).

Proficiency-based dominance scores

The proficiency-based dominance distribution using ratios is smoother for the HeLex measures than the LSBQ-H measures, despite LSBQ-H providing more response options (Figure 6A). There is a strong concentration of balanced proficiency values (around 1) for LSBQ-H results. We entered the main effects of Questionnaire into the linear mixed effects model as potential predictors of HL proficiency-based dominance scores calculated using ratios.

The results of a linear regression model in Supplementary Table S5 suggest that there is no significant difference between the

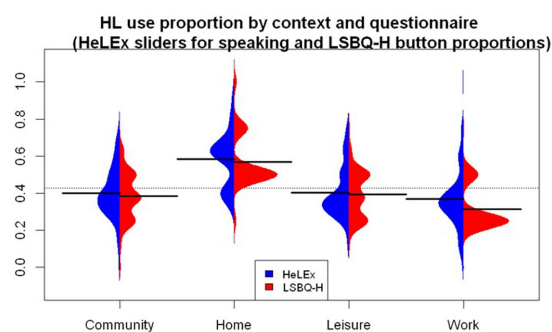


FIGURE 3

The distribution of the use of HL (speaking modality) out of all languages in each context, derived from HeLEx and LSBQ-H (for reasons mentioned in the section *Derived measures: Proportion of HL use in different social contexts*, we do not compare the External Family context across questionnaires).

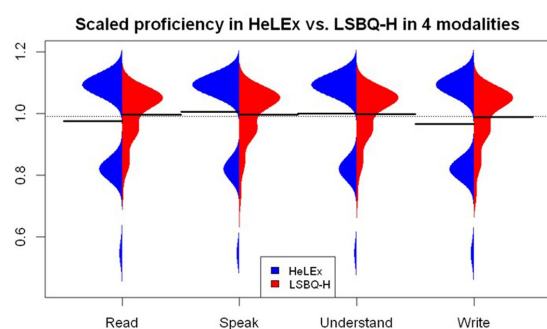


FIGURE 4

Response distribution for HL proficiency by modality in HeLEx and LSBQ-H, scaled.

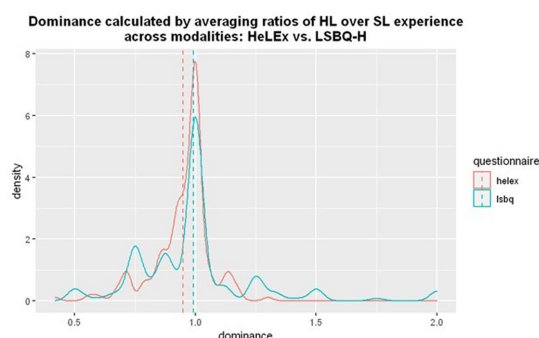


FIGURE 5

The density plot for the distribution of dominance calculated by averaging ratios of HL over SL experience in different modalities from HeLEx (red) and LSBQ-H (blue). The density on y axis represents a kernel density estimate, a smoothed version of frequency on y axis in a regular histogram.

proficiency-based ratio dominance scores derived from HeLEx and LSBQ-H, despite visual differences in distribution. The distributions are affected by differences in the original response scales. A 1-point difference is smaller on a 10-point scale than on a 4-point scale: a

relatively “balanced” participant with high proficiency in both languages might have a 9:10 dominance ratio ($=0.9$) as per LSBQ-H and a 3:4 ratio ($=0.75$) as per HeLEx. The resultant “bunching” of balanced scores is therefore more marked for LSBQ-H than for HeLEx, as seen in Figure 6A. As an alternative method, we also derived a difference score for dominance-by-proficiency, by subtracting SL proficiency from HL proficiency. For this type of calculation, the value of 0 (no difference) would indicate balanced proficiency.

Figure 6B shows that the difference scores provide much more similar distributions of dominance-by-proficiency between the two questionnaires. Here, despite the differences in the original response scales, the resulting scores are more similar across questionnaires. The same relatively “balanced” proficient participant (as discussed above) would have a dominance difference score of -1 (i.e., 9 for HL proficiency minus 10 for SL proficiency) as per LSBQ-H, and a dominance difference score of -1 (i.e., 3 for HL proficiency minus 4 for SL proficiency) as per HeLEx. In a less balanced population sample, the distributions of difference scores would differ more across questionnaires (as the maximum difference score is 9 for LSBQ-H vs. 3 for HeLEx). Both methods nonetheless concur in showing that the vast majority of the participants in our sample are balanced bilinguals, with a slight leaning to SL dominance: the mean dominance score from both questionnaires is to the left of the balance score (i.e., 1 for ratio calculations and 0 for difference score calculations) in each plot.

Language entropy by context

Figure 7 plots the language entropy measures by context for each questionnaire. Both questionnaires seem to return a slightly higher language entropy for the Home and Leisure contexts. The lowest entropy is found in the Work or School context. We entered the main effects of the Questionnaire and Context as well as their interactions into the linear regression model as potential predictors of language entropy scores.

The model summary (Supplementary Table S6) shows that there is no significant difference in entropy estimates between the two questionnaires. The entropy in the Work or School context was significantly lower than the mean of entropies across all contexts as recorded by HeLEx. Focusing on this particular context reveals marked differences in the distribution of scores across questionnaires, in spite of similar means (HeLEx mean $=0.856$, $sd=0.191$; LSBQ-H mean $=0.852$, $sd=0.104$). The data points are concentrated around the highest value for HeLEx, indicating high entropy, whereas the majority of responses are at a lower mark for LSBQ-H. The HeLEx values are overall more distributed due to the slider scale providing more options than the options for the ratio of HL vs. SL use in the LSBQ-H questions where entropy scores only included four possible values.

Interim summary

We considered a range of (mainly standard) measures of HL experience and compared the measures derived from the LSBQ-H data with those derived from the HeLEx data. The measures included HL Experience (across modalities and across contexts), HL proficiency (across modalities), language dominance (based on experience and based on proficiency), and language entropy. Despite some small differences, the results were generally similar across questionnaires, concluding the validation of HeLEx. We now turn to the affordances of HeLEx and discuss their methodological implications.

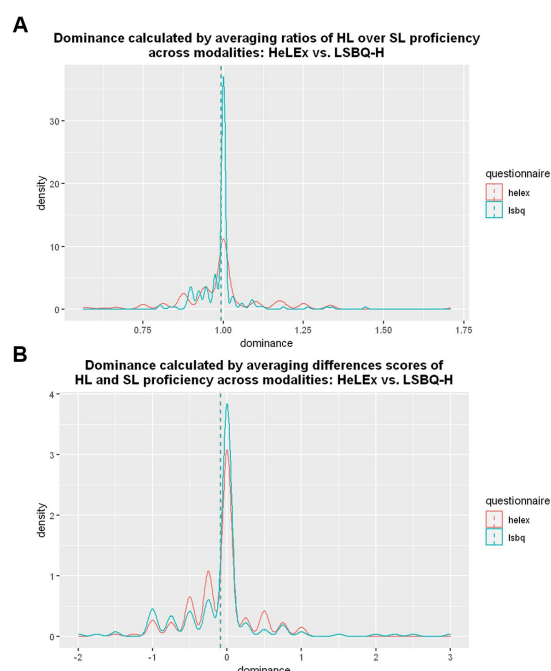


FIGURE 6

(A) The density plot for the distribution of dominance calculated with ratios of HL vs. SL proficiency in different modalities from HeLEx (red) and LSBQ-H (blue). (B) The density plot of the dominance-by-proficiency scores calculated by averaging difference scores between HL and SL proficiency measures across 4 modalities, from HeLEx responses (red) and LSBQ-H responses (blue).

Informativity effects: code-switching and attitudes

HeLEx is characterized by additional informativity compared to LSBQ-H as it includes an extended code-switching (CS) module, an extended section on personal and societal attitudes towards HL, an extensive section on the personal and societal attitudes to CS in five contexts, questions on the number of speakers of HL and SL in each context and their HL proficiency, among others.

The CS module of HeLEx probes the frequency of personal CS use as well as CS exposure in the five social contexts for two directions of code-switching (HL to SL and SL to HL) and for three structural types (one word, two to three words, intersentential CS). When asked how often they use or are exposed to a specific type of CS in each context, the participants had the following general options: “(almost) never, in one or two conversations per week, in one or two conversations per day, in (almost) every conversation, I do not know.”

In the battery of questions on personal and societal attitudes to the HL use and knowledge, participants use sliders (0–100) to indicate how much they agree with specific statements. Personal attitudes statements include, among others, “I identify myself as [a national of the HL matrix country],” “It is important that my children learn [HL] to a high degree,” “It is important to me to speak and understand [HL] like speakers who live in [the HL matrix country],” “I am satisfied with my current overall ability in [HL].” Societal attitudes statements include “I am worried that speaking my home language is not welcome/tolerated in the wider society,” “There is sufficient support

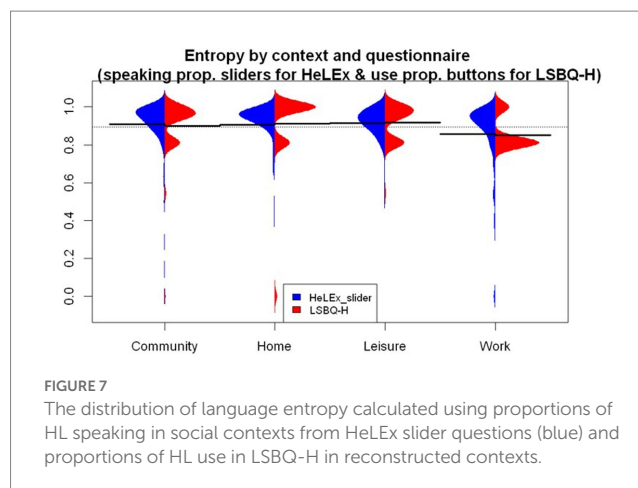


FIGURE 7

The distribution of language entropy calculated using proportions of HL speaking in social contexts from HeLEx slider questions (blue) and proportions of HL use in LSBQ-H in reconstructed contexts.

from the government and society for maintaining my home language,” “I feel external pressure to speak in the dominant language of the society, either by colleagues, friends, etc.” The personal attitudes questions are mostly co-opted from the Bilingual Language Profile (Gertken et al., 2014). These questions can be grouped into several scores (e.g., the importance of HL for self-identification, satisfaction with and perceived importance of HL knowledge, etc.), or they can be averaged to create an index of positive attitudes to HL.

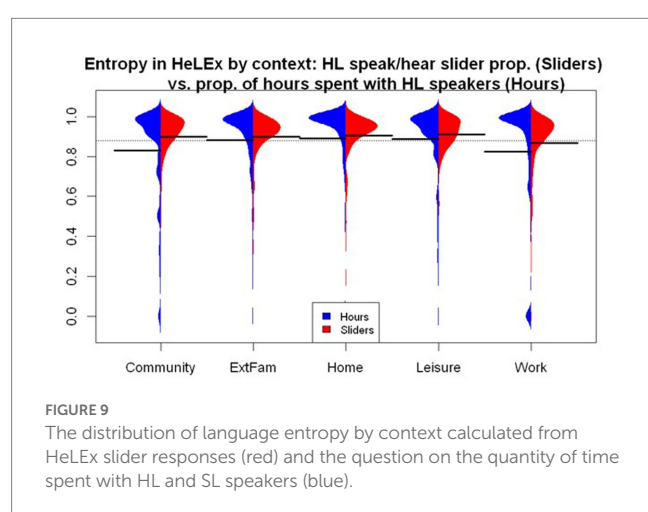
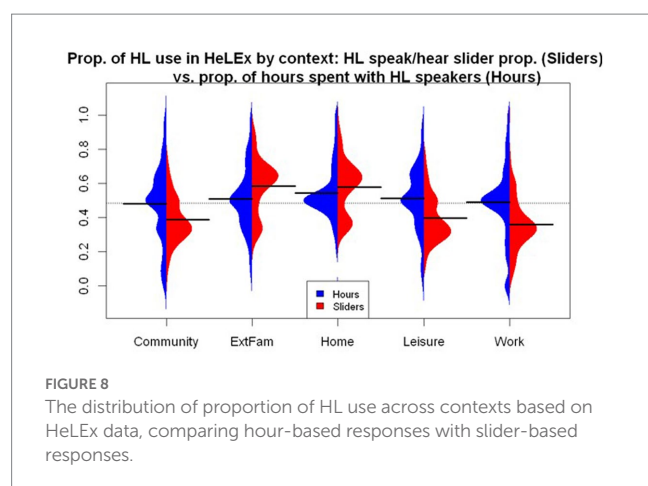
Additionally, HeLEx affords the opportunity for triangulation in relation to the documentation of HL experience, as some aspects are probed by two similar sets of questions. We investigate discrepancies between different types of response scales, and their impact on derived measures. We also illustrate the additional informativity of HeLEx with the results for the questions on the input diversity (number of speakers of each language in each context) and the HL input quality.

Estimates of the proportion of HL experience: natural metrics vs. estimated proportions

The proportion of HL use can be calculated in two ways using HeLEx data. One method, using natural metrics data, divides the hours spent with HL speakers in a particular context (in a typical week) by all hours spent with anyone in that context (in a typical week). The other method, using estimated measures, is based on the proportion of HL vs. other languages (elicited via sliders), averaging the values for speaking and hearing.⁸ The resulting distributions are shown in Figure 8.

Both methods reveal similar trends across contexts. The proportion of HL use in the Home is the highest, whereas it is the lowest in the Community and Work or School contexts. However, the proportions calculated with slider responses seem to exhibit more marked by-context differences than the proportions obtained from

⁸ In a more typically researched heritage speaker population, spending time with HL speakers could likely mean only being exposed to, i.e., hearing, HL. Nevertheless, the proficiency and use patterns from the questions on modalities suggest that these speakers are as likely to speak as they are to hear HL when spending time with other HL speakers.



hour-based responses. We entered the main effects of the Questionnaire and the Manner of the HL Use Proportion Calculation, as well as their interactions, into the linear mixed effects model as potential predictors of HL proficiency scores with random intercepts for participants. The Manner of Calculation variable had Hours as the reference value.

The results of the linear mixed effects regression model (Supplementary Table S7) suggest that the HL proportion calculated using slider responses is significantly lower overall. In terms of contexts, Community has a significantly lower HL use proportion estimate, whereas Home has a higher estimate, compared to the mean of means of HL proportion across contexts for the hour-based calculation. The significant interaction of Context and Manner of calculation suggests that slider-derived estimates are higher for the Home and External Family contexts and lower for the Work or School, Community, and Leisure contexts.

The slider responses might provide more categorical estimates: they further amplify the HL proportion trends for the Home and Community context. They also might be more reliable, as they give direct estimates of the proportion of language experience in each context and should also reflect the proportion HL use during language mixing. By contrast, the hour-based data is a derived measure with more steps, and it only reflects the time spent with HL interlocutors in each context (irrespective of the actual HL use proportion with these people, in case they are bilinguals).

Deriving language entropy from interaction hours vs. slider data

We derived language entropy scores from the two estimates of HL experience we have just compared, yielding the distributions shown in Figure 9.

We entered the main effects of the Manner of Calculation and Context, as well as their interactions, into the linear mixed effects model as potential predictors of entropy scores, with random intercepts for participants. The Manner of Calculation variable had Hours as the reference value.

The linear mixed effects model summary in Supplementary Table S8 shows that, compared to the mean of means of entropy for all contexts as calculated using hours, language entropy is significantly lower in the Community and Work or School contexts and higher in the Home and Leisure contexts. The significant main effect of the Manner of Entropy Calculation suggests that entropy calculated using slider responses is significantly higher, possibly due in part to the inclusion of potential non-responses (sliders left on 0.5 translating to high entropy). The interaction between the Manner of Calculation and context suggests that the estimates for Community entropy are significantly higher for the slider-derived calculation.

Considering the time spent in each context

One of the HeLex features not available in the LSBQ(-H) is that it documents the estimated amount of time spent in each context. As seen in Figure 10A, this varies substantially both across contexts and within contexts. In general, respondents report spending most time in the Home and Work/School environments.

The proportion of HL use also varies substantially across contexts (highest in the Home and with External Family), as shown in Figure 10B.

When calculating the overall proportion of HL use across contexts, it is important to take into account the actual proportion of the time spent in each context. Figure 10C compares overall proportions with vs. without weighing by time-in-context (based on the calculations explained in the section *Derived measures: Accounting for the actual proportion of time spent in each context*).

The weighted scores seem to be slightly higher than the unweighted scores, likely due to the overlap between the contexts in which participants spend a lot of time in and contexts in which there is a high proportion of HL use, such as Home. The difference did not prove statistically significant (Supplementary Table S9). This will need to be replicated using more diverse groups of Heritage Speakers. In this case, the high proportion of HL use in the work context appears to have balanced out the small amount of time spent with Extended Family.

HL input quality and diversity

Two important dimensions of the richness of HL experience are the number and diversity of interlocutors and their level of proficiency in the HL. HeLex is particularly informative in these respects: it quantifies and “qualifies” HL speakers in each context.

The present participant sample seems to get the most diverse HL input in the Family outside of the home context, i.e., ExtFam, judging by the number of HL speakers they spend time with in the context (Figure 11A).

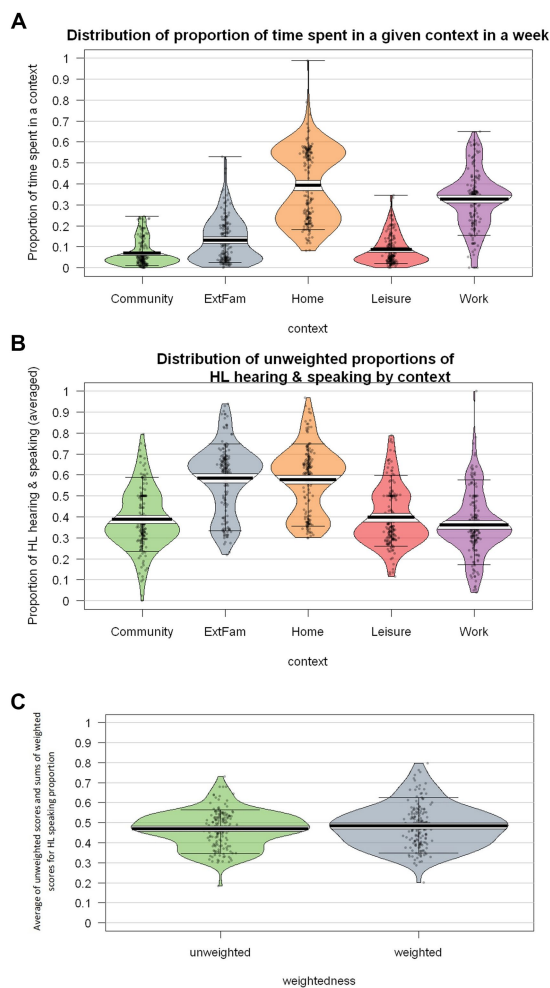


FIGURE 10

(A) Distribution of proportion of time spent in 5 contexts. (B) Distribution of unweighted proportion of HL vs. SL hearing and speaking (averaged). (C) The distribution of cross-context averages of unweighted HL speaking proportion scores and sums of weighted HL speaking proportion scores.

The proportion of speakers with good HL proficiency and HL-dominant speakers follows a broadly similar distribution pattern across contexts, with expectedly higher estimates for the proportion of speakers with good HL proficiency, so we only present the distribution of proportions of HL-dominant speakers (Figure 11B). Interestingly, though, the external family does not seem to have the highest proportion of HL-dominant or HL-proficient speakers. Rather, such speakers are most represented in the Home context.

The results of a linear regression model (Supplementary Table S10) confirm that Home has the highest proportion of HL-dominant speakers, whereas External Family has the lowest proportion, compared to the mean of means of proportion values at every level of the context variable.

Discussion

In the present study, we analyzed the language experience data from 174 Heritage Speakers of Turkish living in Germany using both a

slightly extended version of LSBQ, and HeLEx (“Heritage Language Experience questionnaire”: a new questionnaire amalgamating, modifying, and building on LSBQ and other questionnaires, e.g., the Bilingual Language Profile). We carried out two sets of analyses. The first aimed to ascertain whether the two questionnaires reliably capture the same reality, insofar as the distribution of the resulting measures is sufficiently similar. The second explored the informativity of each questionnaire, in terms of scope and granularity of the derived measures.

Group-level analyses reveal that, despite the distributional differences due to different response scales (see Figures 2, 4), the key variables obtained from each questionnaire are nonetheless sufficiently similar, in that no statistically significant difference was detected in linear regression models probing questionnaire effect on the scaled variables of interest. This was shown in turn for language experience by modality (speaking, listening, reading, writing) and by context (Home, Work or School, Leisure, Community), self-estimated proficiency, experience-based dominance, proficiency-based dominance, and language entropy. The only between-questionnaire difference observed was that LSBQ-H estimates of HL experience across social contexts are significantly lower than the HeLEx ones, especially in the Work/School context. We conclude that the two questionnaires are overall similarly successful at detecting the important distributional patterns in the data.

In terms of informativity, our analyses brought to light several issues regarding response scales (e.g., scales with 4 vs. 7 options, numerical vs. qualitative labels) and response mechanisms (e.g., sliders vs. buttons), which will need to be taken into account in further developments of these and other language experience questionnaires.

First, the minimum and maximum values allocated by design to response scales documenting language experience (i.e., exposure and use) need to take into account the fact that equivalents to 0% (e.g., “never”) and 100% (e.g., “all the time”) will mostly not apply to bilinguals, as even the most dominant ones will still experience their weaker language to some extent. In LSBQ-H, what was by design a 5-point scale effectively turned out to be a 3-point scale as the extremes did not apply. The implication for future questionnaires is that, if unrealistic absolute values are used, the granularity of the scale needs to be adapted accordingly to allow the desired level of detail.

Second, the choice of whether to assign qualitative labels to points on a Likert scale needs careful consideration. Recall that to capture self-reported proficiency, the LSBQ-H employed an 11-point numeric scale with qualitative labels attached to the extreme ends (0, 10) only. In contrast, HeLEx used a 4-point scale with qualitative labels for each of the points (Supplementary Table S5). Our comparative analysis of self-reported proficiency data across questionnaires reveals that the use of qualitative labels such as “pretty well” and “very well” for the top half of a 4-point scale (in HeLEx) returned a less positively skewed distribution than an 11-point numeric scale with qualitative labels attached to the extreme ends (0, 10) only (in the LSBQ-H). Note however that the positive skew of LSBQ-H proficiency responses could have partly stemmed from the response mechanism, sliders, whose potential impact is discussed below. Numeric scales are not necessarily more objective, however: one respondent’s “9” could effectively equate to another’s “7.” The psychometric literature calls for caution in the choice of response scales. Following Dillman et al. (2014), in HeLEx we limited the number of categories on the scale, used symmetrical categories at each end of the scale, and labeled the categories verbally rather than

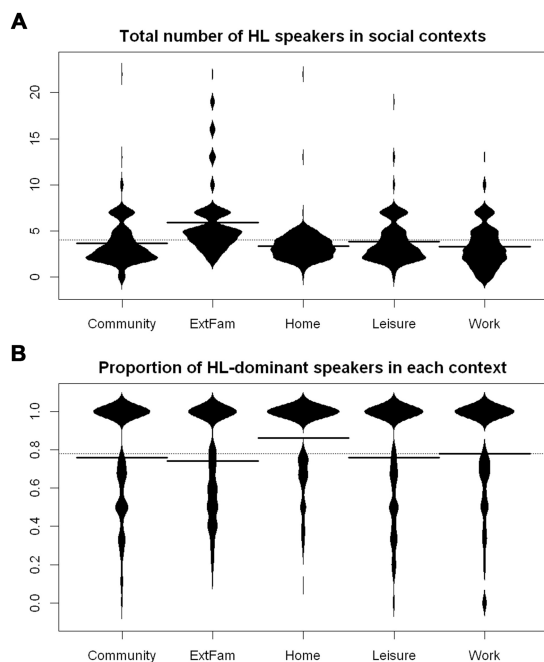


FIGURE 11

(A) Total number of HL speakers in each context. Full lines represent means by contexts, whereas the dotted line represents the overall mean. The clustering of responses into “knots” is an artifact of quantization (cf., Table 5). (B) Proportion of speakers who are dominant in HL in each context. The calculations included a check for a number of speakers dominant in HL higher than the total number of speakers, so these were excluded.

numerically. We believe this led to more consistency across respondents, as it reduces the possibility of different interpretations of what a numerical score of 3 or 7 means in terms of proficiency. A comparison of the scores (derived from each questionnaire) with an objective measure of language proficiency will be needed to settle the issue, but it is beyond the scope of the present paper.

Third, the granularity of ordinal data obtained from Likert scales combined with population characteristics has an impact on the distribution of variables derived from these ordinal data. In this highly balanced population sample, ratio-based dominance-by-proficiency scores (dividing HL proficiency by SL proficiency) featured less variance than difference-based scores (subtracting SL proficiency from HL proficiency; see Figure 6). Further research will need to investigate the informativity of each type of measure when used as predictor variables.

Fourth, the response elicitation mechanism seems to play a role in the response distribution. Slider scales in both questionnaires (to measure proficiency in the LSBQ-H and the proportion of HL use in social contexts in HeLex) seem to amplify intuitive, categorical choices, where the beginning, mid-point, and end of the slider scale seem to be “hot-spots,” depending on whether the participant considers themselves to be balanced, HL-dominant, or SL-dominant in their language use. This is likely due to the motoric nature of filling in the slider scales, as well as visual presentation. Filling out proficiency scales has different motoric requirements on paper (LSBQ) vs. online (LSBQ-H), and compared to selecting a button response or an option from a menu. On paper, the proficiency scale is filled out by placing a mark on the scale, a movement considered and planned in advance. Its most

obvious equivalent in the online questionnaire widget selection is the slider scale. Not to confuse participants and to ensure a consistent point of departure for all responses and participants, slider scales in both HeLex and LSBQ-H included an initially visible slider tip in the middle of the scale (the tip could have also been initially hidden).⁹ Participants could similarly just click on the desired point on the scale, and the slider tip would appear there. Nevertheless, most participants are likely to have clicked on the tip and dragged it to the desired position. This movement might and more likely to be executed by pulling the slider all the way to the movement limit (beginning or the end point, depending on the participant's experience), and adjusted slightly from there, or simply leaving it in the middle in case they believe this is the appropriate choice. The “slipperiness” of sliders when choosing a response close to the end of scales could thus cause exponential distribution with a concentration of responses at the end or the beginning of the scale, whereas the inertness of the slider tip when the participant feels a balanced 0.5 response is in order could cause overestimation of balanced scores. In a maximally representative sample of HSSs, the slider effect would likely manifest as a trimodal distribution. With the increased necessity for online data collection and translating questionnaires from paper to online platforms, it is important to consider whether the “obvious” online equivalents to paper question formats, e.g., slider scales, are indeed filled out in the same way. As results suggest, this difference in motoric execution of filling in responses is not negligible and could affect the results, in addition to factors such as level of measure derivation or number of response options.

Finally, the treatment of non-responses is not a trivial issue. It is important to distinguish between meaningful non-responses (implying the question does not apply to the participant, or the probed quantity is 0) and non-meaningful non-responses (due to fatigue or non-willingness to respond). There were interactions between the manner of elicitation/calculation of HL use proportion and context, such that slider-derived estimates for HL use proportion for Work/School, Community, and Leisure were significantly lower than hour-derived estimates. The hour-derived estimates particularly for Work/School, Community, and Leisure could have been artificially inflated by excluding quantity non-responses which should have been treated as zeros, since the language experience in question might not have been relevant to some participants. For example, the number of days a participant spends with HL speakers in the Work/School context could have been 0, but the participant left the question on “select” instead of choosing 0 from the menu, which was quantized as NaN and thus technically excluded from plots and statistical models. If these non-responses were turned to zeros to imply that there is no HL use in these contexts, the mean HL use proportions for these contexts would decrease, bringing them closer to slider estimates. We conclude that triangulation or probing similar constructs with several questions of different format and formulation, is highly useful for overcoming such difficulties in interpretation.

While it is not yet established as a standard predictor in bilingualism research, we decided to include Language Entropy in the set of derived variables of interest, as we believe this measure provides

⁹ The LSBQ paper version also has a salient mark and a label (5) at the mid-point of the scale, which could also potentially draw more balanced responses.

an objective estimate of linguistic diversity by context of language experience (Gullifer and Titone, 2020). It might be a reliable proxy for the level of (between-speaker) language mixing (though we leave this for future research). The consistent use of the same five contexts throughout the HeLex questionnaire facilitates entropy calculations, whereas we show that it is more complex and error-prone to group speakers and derive contexts from various questions in the LSBQ.

We believe the HeLex questionnaire has a number of advantages as a tool documenting language experience in adult Heritage Speakers. First, language experience questions are all asked in relation to the same set of 5 contexts. This avoids having to reconstruct contexts from by-interlocutor data and avoids having to make assumptions about who the key interlocutors might be in each context (e.g., composition of the homes of young adults). It allows the straightforward combination of information about each context from different questions (e.g., in order to adjust by-context quantities for the actual amount of time spent in each context). We assume that maintaining the same contexts as frame of reference across questions helped reduce the cognitive burden of the questionnaire. Independent evidence would however be required to ascertain that this was the case. Second, we followed the recommendations from the psychometric literature (Dillman et al., 2014, 2016) by systematically using qualitative labels on Likert scales, and by relying on natural metrics (e.g., number of people, hours, days) instead of more ambiguous adverbs of quantification.

One limitation of the study was that the test–retest reliability was not estimated for HeLex prior to the comparison with LSBQ-H, as we were presented with a unique opportunity to compare HeLex against LSBQ before completing this step, with a large accessible sample who had recently completed LSBQ-H. Test–retest reliability and (confirmatory) factor analysis should be conducted. Another consequence of “inheriting” data, from a study which did not ensure diversification or representativity across HL populations, was the relatively high homogeneity of the participant sample in terms of language experience and proficiency in both the HL and the SL language. Also, one possible limitation is the time elapsed between the completion of the two questionnaires: several months to a year, leading to potential changes in language experience. However, we believe this is not a cause for concern, as the group-level comparisons reveal a consistent picture.

The current paper suggests that HeLex is successful in capturing the same constructs as previous questionnaires, namely LSBQ(-H), and provides additional above-mentioned affordances. The findings underline the importance of careful consideration of methodological choices regarding individual difference data elicitation and derivation, and their potential impact on subsequent analyses. The next step of our research is to use HeLex to document language experience in a highly diverse population of heritage language speakers, and to identify the key language experience variables that predict individual differences in language outcomes (both in terms of language processing and language proficiency).

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <https://osf.io/mkjax/>.

Ethics statement

The studies involving human participants were reviewed and approved by NSD - Norsk AS Dataforskningssenter (personverntjenester@nsd.no). The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

CDC and AT co-designed HeLex, with CDC taking the conceptual and methodological lead. AT implemented the new questionnaire online, created scripts for response quantization and variable derivation, conducted statistical analysis and interpretation, and wrote the first draft of methods, results, and discussion, with assistance and guidance from CDC. CDC co-wrote and edited the manuscript. YR consulted on questionnaire design, wrote introduction, and edited the manuscript. FB consulted on the design of HeLex, recruited participants, collected the data, and participated in writing. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by the UiT The Arctic University of Norway Aurora Center for Language Acquisition, Variation & Attrition; The Dynamic Nature of Languages in the Mind (project code 2062165). The publication charges for this article have been funded by a grant from the publication fund of UiT The Arctic University of Norway. This project also received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 799652.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1131374/full#supplementary-material>

References

- Abutalebi, J., and Green, D. W. (2016). Neuroimaging of language control in bilinguals: neural adaptation and reserve. *Biling. Lang. Cogn.* 19, 689–698. doi: 10.1017/S1366728916000225
- Anderson, J. A. E., Mak, L., Keyvani Chahi, A., and Bialystok, E. (2018). The language and social background questionnaire: assessing degree of bilingualism in a diverse population. *Behav. Res. Methods* 50, 250–263. doi: 10.3758/s13428-017-0867-9
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bayram, F., Rothman, J., Iverson, M., Kupisch, T., Miller, D., Puig-Mayenco, E., et al. (2019). Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *Int. J. Biling. Educ. Biling.* 22, 919–939. doi: 10.1080/13670050.2017.1324403
- Birdsong, D., Gertken, L., and Amengual, M. (2012). *Bilingual language profile: An easy-to-use instrument to assess bilingualism*. COERLL. University of Texas.
- COST Action IS0804 (2011). Parents of bilingual children questionnaire (PaBiQ). A part of the LITMUS battery (COST IS0804). Available at: <http://www.bi-sli.org> (Accessed December 20, 2022).
- de Bruin, A. M. T. (2019). Not all bilinguals are the same. A call for more detailed assessments and descriptions of bilingual experiences. *Behav. Sci.* 9:33. doi: 10.3390/b9030033
- De Cat, C., Kaščelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., et al. (2023). How to quantify bilingual experience? Findings from a Delphi consensus survey. *Biling. Lang. Cogn.* 26, 112–124. doi: 10.1017/S1366728922000359
- del Prado Martín, F. M., Kostić, A., and Baayen, R. H. (2004). Putting the bits together: an information theoretical perspective on morphological processing. *Cognition* 94, 1–18. doi: 10.1016/j.cognition.2003.10.015
- Dillman, D. A., Hao, F., and Millar, M. M. (2016). “Improving the effectiveness of online data collection by mixing survey modes” in *The sage handbook of online research methods*. eds. N. G. Fielding, R. M. Lee and G. Blank. 2nd ed (London: Sage Publications)
- Dillman, D. A., Smyth, J. D., and Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* John Wiley & Sons.
- Dunning, D., Heath, C., and Suls, J. M. (2004). Flawed self-assessment implications for health, education, and the workplace. *Psychological Science in the Public Interest*, 5, 69–106. doi: 10.1111/j.1529-1006.2004.00018.x
- Gertken, L. M., Amengual, M., and Birdsong, D. (2014). “Assessing language dominance with the bilingual language profile” in *Measuring L2 proficiency: Perspectives from SLA*. eds. P. Leclercq, A. Edmonds and H. Hilton (Bristol: Multilingual Matters), 208–225.
- Green, D. W., and Abutalebi, J. (2013). Language control in bilinguals: the adaptive control hypothesis. *J. Cogn. Psychol.* 25, 515–530. doi: 10.1080/20445911.2013.796377
- Grosjean, F. (2001). “The bilingual’s language modes” in *One mind, two languages bilingual language processing*. ed. J. Nicol (Malden, MA: Blackwell Publishers), 1–22.
- Grosjean, F. (2015). “The complementarity principle and its impact on processing, acquisition, and dominance” in *Language dominance in bilinguals: Issues of measurement and operationalization*. eds. C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press), 66–84.
- Gullifer, J. W., Chai, X. J., Whitford, V., Pivneva, I., Baum, S., Klein, D., et al. (2018). Bilingual experience and resting-state brain connectivity: impacts of L2 age of acquisition and social diversity of language use on control networks. *Neuropsychologia* 117, 123–134. doi: 10.1016/j.neuropsychologia.2018.04.037
- Gullifer, J. W., and Titone, D. (2018). Compute language entropy with {languageEntropy}. Available at: <https://github.com/jasongullifer/languageEntropy> (Accessed December 20, 2020).
- Gullifer, J. W., and Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Biling. Lang. Cogn.* 23, 283–294. doi: 10.1017/S1366728919000026
- Hale, J. (2003). The information conveyed by words in sentences. *J. Psycholinguist. Res.* 32, 101–123. doi: 10.1023/A:1022492123056
- Kaščelan, D., Prévost, P., Serratrice, L., Tuller, L., Unsworth, S., and De Cat, C. (2022). A review of questionnaires quantifying bilingual experience in children: do they document the same constructs? *Biling. Lang. Cogn.* 25, 29–41. doi: 10.1017/S1366728921000390
- Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006
- Li, P., Sepanski, S., and Zhao, X. (2006). Language history questionnaire: a web-based interface for bilingual research. *Behav. Res. Methods* 38, 202–210. doi: 10.3758/BF03192770
- Li, P., Zhang, F., Tsai, E., and Puls, B. (2014). Language history questionnaire (LHQ 2.0): a new dynamic web-based research tool. *Biling. Lang. Cogn.* 17, 673–680. doi: 10.1017/S1366728913000606
- Li, P., Zhang, F., Yu, A., and Zhao, X. (2020). Language history questionnaire (LHQ3): an enhanced tool for assessing multilingual experience. *Biling. Lang. Cogn.* 23, 938–944. doi: 10.1017/S1366728918001153
- Lloyd-Smith, A., Bayram, F., and Iverson, M. (2020). “The effects of heritage language experience on lexical and morphosyntactic outcomes” in *Studies in Turkish as a heritage language*, 60. ed. F. Bayram (John Benjamins Publishing), 63–86.
- Luk, G., and Bialystok, E. (2013). Bilingualism is not a categorical variable: interaction between language proficiency and usage. *J. Cogn. Psychol.* 25, 605–621. doi: 10.1080/20445911.2013.795574
- Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)
- Paradis, M. (2011). Principles underlying the bilingual aphasia test (BAT) and its uses. *Clin. Linguist. Phon.* 25, 427–443. doi: 10.3109/02699206.2011.560326
- Paradis, J., Emmerzael, K., and Sorenson, D. T. (2010). Assessment of English language learners: using parent report on first language development. *J. Commun. Disord.* 43, 474–497. doi: 10.1016/j.jcomdis.2010.01.002
- R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: <https://www.R-project.org> (Accessed December 20, 2022).
- Rodriguez-Fornells, A., Krämer, U. M., Lorenzo-Seva, U., Festman, J., and Münte, T. F. (2012). Self-assessment of individual differences in language switching. *Front. Psychol.* 2:388. doi: 10.3389/fpsyg.2011.00388
- Rothman, J., Bayram, F., DeLuca, V., González Alonso, J., Kubota, M., and Puig-Mayenco, E. (2023). “Defining bilingualism as a continuum: consequences for the study of bilingual mind and brain effects” in *Understanding language and cognition through bilingualism: In honor of Ellen Bialystok*. eds. G. Luk, J. A. E. Anderson and J. G. Grundy (John Benjamins Publishing Company)
- Serratrice, L., and De Cat, C. (2020). Individual differences in the production of referential expressions: the effect of language proficiency, language exposure and executive function in bilingual and monolingual children. *Biling. Lang. Cogn.* 23, 371–386. doi: 10.1017/S1366728918000962
- Shannon, C. E. (1948). The mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x
- Surrian, S., and Luk, G. (2017). *Describing bilinguals: A systematic review of labels and descriptions used in the literature between 2005–2015*. Bilingualism: Language and Cognition, 1–15.
- Titone, D. A., and Tiv, M. (2022). *Rethinking multilingual experience through a systems framework of bilingualism*. Bilingualism: Language and Cognition 1–16.
- Unsworth, S. (2013). Assessing the role of current and cumulative exposure in simultaneous bilingual acquisition: the case of Dutch gender. *Biling. Lang. Cogn.* 16, 86–110. doi: 10.1017/S1366728912000284



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY

Mikael Roll,
Lund University, Sweden
Svenja Schmid,
University of Konstanz, Germany

*CORRESPONDENCE

Nuria Sagarra
✉ nuria.sagarra@rutgers.edu

RECEIVED 10 January 2023

ACCEPTED 04 May 2023

PUBLISHED 12 June 2023

CITATION

Sagarra N and Casillas JV (2023) Practice beats age: co-activation shapes heritage speakers' lexical access more than age of onset.
Front. Psychol. 14:1141174.
doi: 10.3389/fpsyg.2023.1141174

COPYRIGHT

© 2023 Sagarra and Casillas. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Practice beats age: co-activation shapes heritage speakers' lexical access more than age of onset

Nuria Sagarra* and Joseph V. Casillas

Department of Spanish and Portuguese, Rutgers University, New Brunswick, NJ, United States

Probabilistic associations make language processing efficient and are honed through experience. However, it is unclear what language experience factors explain the non-monolingual processing behaviors typical of L2 learners and heritage speakers (HSs). We investigated whether AoO, language proficiency, and language use affect the recognition of Spanish stress-tense suffix associations involving a stressed syllable that cues a present suffix (*SALta* "s/he jumps") and an unstressed syllable that cues a past suffix (*SALtó* "s/he jumped"). Adult Spanish-English HSs, English-Spanish L2 learners, and Spanish monolinguals saw a paroxytone verb (stressed initial syllable) and an oxytone verb (unstressed initial syllable), listened to a sentence containing one of the verbs, and chose the one they heard. Spanish proficiency measured grammatical and lexical knowledge, and Spanish use assessed percentage of current usage. Both bilingual groups were comparable in Spanish proficiency and use. Eye-tracking data showed that all groups fixated on target verbs above chance before hearing the syllable containing the suffix, except the HSs in the oxytones. Monolinguals fixated on targets more and earlier, although at a slower rate, than HSs and L2 learners; in turn, HSs fixated on targets more and earlier than L2 learners, except in oxytones. Higher proficiency increased target fixations in HSs (oxytones) and L2 learners (paroxytones), but greater use only increased target fixations in HSs (oxytones). Taken together, our data show that HSs' lexical access depends more on number of lexical competitors (co-activation of two L1 lexica) and type (phonotactic) frequency than token (lexical) frequency or AoO. We discuss the contribution of these findings to models in phonology, lexical access, language processing, language prediction, and human cognition.

KEYWORDS

heritage speakers, stress, age of onset, proficiency, use, eye-tracking, lexical access, co-activation

1. Introduction

Monolinguals use multiple cues to predict what a speaker will say, but L2 learners struggle when making predictions based on L2 cues absent in their L1. However, it is unclear what causes this struggle. The study of heritage speakers (HSs) allows us to examine the role of age of onset (AoO) and language experience on L2 processing. These variables advance our understanding of why HSs differ from monolinguals and first-generation immigrants. HSs refer to "individuals from minority language groups who grow up exposed to a minority language in the home and the majority societal language" (Montrul, 2016, p. 16). HSs' uniqueness is attributed to representational differences (Montrul, 2008), limited quality input (Pires and Rothman, 2009), gradual attrition (Polinsky, 2011), or reduced current activation of their heritage language

(Putnam and Sánchez, 2013). We explored alternative explanations in terms of reduced knowledge of their heritage language (proficiency), as well as increased lexical competition due to co-activation of their two L1 lexica (use: current amount of input, output, and interaction in a language on a weekly basis). We employed an auditory implicit eye-tracking task and investigated whether AoO, language proficiency, and language use influence how monolinguals, HSs, and L2 learners form stress-suffix lexical associations during spoken word access. Probabilistic associations are crucial in making spoken language processing efficient (Romberg and Saffran, 2010), and are refined through experience. L2 studies show that higher language proficiency facilitates tone-tense and tone-number suffix associations in L2 Swedish (Schremm et al., 2016; Gosselke Berthelsen et al., 2018, 2020) and stress-tense suffix associations in L2 Spanish (Sagarra and Casillas, 2018), and that novice learners only recognize L2 tone-suffix associations if their L1 is tonal (Gosselke Berthelsen et al., 2021).

While research comparing monolinguals to both L2 learners and HSs could tease apart AoO from language experience, these studies are often inconclusive. Written mode studies (e.g., Foote, 2011; Keating, 2022; Parshina et al., 2022) are problematic because HSs perform auditory tasks better than written ones (Bowles, 2011). Single-proficiency studies are unable to determine whether non-native processing is due to late AoO, low proficiency, or both (e.g., Sekerina and Trueswell, 2011; Montrul et al., 2013; Jegerski and Sekerina, 2020). L2 studies without HSs (e.g., Nichols and Joanisse, 2016), HS studies without monolinguals (Lemmerth and Hopp, 2019), or HS studies with a composite score merging AoO and proficiency (Hervais-Adelman et al., 2018) are incapable of assessing AoO effects. Some studies combined AoO and proficiency (Wartenburger et al., 2003; Hervais-Adelman et al., 2018; Sagarra and Rodríguez, 2022), AoO and use (Lloyd-Smith et al., 2019), or proficiency and use (Di Pisa and Marinis, 2022), and the only study examining AoO, proficiency and use separately examined morphosyntax (Sagarra et al., 2021). We investigated the separate effects of AoO, proficiency, and current use on the recognition of Spanish stress-tense suffix associations by Spanish-English HSs, English-Spanish L2 learners, and Spanish monolinguals. Lexical stress is contrastive in English and Spanish, but these languages differ in stress realization and cue weight.

2. Lexical stress

Lexical stress (henceforth stress) refers to the relative prominence of one syllable with regard to the others in a given word. Stress is lexically encoded and contrastive in Spanish (*término* ['ter.mi.no] “term;” *termino* [ter.'mi.no] “I finish;” *terminó* [ter.mi.'no] “s/he finished”) and in English (*produce* ['pɹo.du:s] noun; *produce* [pɹə.'du:s] verb), though it is more productive in Spanish than in English. To wit, few stress minimal pairs exist in English that are not semantically related (see Cutler, 2012). The primary acoustic correlates of stress are f₀, duration, and intensity, although their relative cue-weighting is language-specific (see Holt and Lotto, 2006; Chrabaszcz et al., 2014; Gordon and Roettger, 2017, among many others). Despite native English speakers' familiarity with stress, they typically have trouble producing (Bullock and Lord, 2003; Lord, 2007) and perceiving (Face, 2000, 2005, 2006; Saalfeld, 2012; Ortega-Llebaria et al., 2013) stress differences in L2 Spanish. A possible explanation might be found in language-specific isochrony (Pike, 1945). Whereas

English is often described as a “stress-timed” language, i.e., one with relatively constant intervals between stressed syllables, Spanish is typically described as “syllable-timed,” i.e., each syllable is perceived as having the same duration. Differences such as these may shape how stress is perceived in each language. In English, for example, unstressed vowel reduction—often present in stress-timed languages—may be sufficient for indicating stress (Cutler, 2012; Tremblay et al., 2018), rendering other cues relatively less important for speech perception. Consequently, native English speakers need to adjust their cue-weighting strategies when learning Spanish, a language that does not have vowel reduction. Evidence from cross-modal priming studies indicates that stress is processed differently by native listeners in both languages during lexical access (see Soto-Faraco et al., 2001; Cooper et al., 2002). Extant literature also suggests that native listeners are tuned in to the relevant acoustic cues of their language and take advantage of them to increase processing efficiency. Unsurprisingly, they use the same cue-weighting strategies when learning an L2, which often generates difficulties in the early stages of acquisition (Iverson et al., 2003; Ingvalson et al., 2012). With respect to prediction, there is evidence that monolingual Spanish speakers use lexical stress to predict a word's suffix and that highly proficient L2 learners can also master this skill (Sagarra and Casillas, 2018), but it is unclear whether unique language experiences and earlier AoO modulate spoken word prediction.

3. The role of AoO, proficiency, and use on bilingual language processing and prediction

Hundreds of studies conducted over half a century have yielded mixed findings regarding the effects of AoO on language acquisition in bilinguals (see Mayberry and Kluender, 2017, for a review, and Singleton and Leśniewska, 2021, for an argument that the critical period hypothesis is irrelevant because it is unfalsifiable). Offline studies are inconclusive. Some studies showed that advanced HSs were grammatically more accurate than advanced L2 learners in perception and production tasks (Bowles, 2011), whereas others did not reveal any grammar differences between the two at any proficiency level (Foote, 2011). Relevant to our study, Kim (2020) reported that Spanish-English HSs perceived Spanish lexical stress more accurately than English-Spanish L2 learners, but the two were equally deviant from monolinguals in production. Online studies are equally ambiguous. While several studies concluded that HSs processed morphology more effectively with earlier than later AoO (Veríssimo et al., 2018), others showed no differences between HSs and L2 learners (Wartenburger et al., 2003; Foote, 2011; Rodríguez and Reglero, 2015; Martohardjono et al., 2017). This lack of consensus has led researchers to question if we are missing the point by focusing on AoO rather than the quality and quantity of bilinguals' individual language experiences (Luk and Platsikas, 2015). Although bilinguals and monolinguals are conceived as separate homogeneous groups, the degrees of variability among bilinguals are enormous (De Bruin, 2019). The investigation of language proficiency and use advances our understanding of what factors produce such variability.

L2 proficiency studies showed that low L2 proficiency denoted delayed processing, insensitivity to violations, processing violations as semantic anomalies, reduced attention to cues used by monolinguals,

and less and later fixations on targets (see Ito and Pickering, 2021, for a review of L2 prediction studies examining proficiency effects). Regarding morphology, higher L2 proficiency promoted the processing of L2-derived and inflected words, new valid derivations, and forms combining a real stem with a new suffix (Kimppa et al., 2019). Concerning phonology, higher L2 proficiency inhibited L1 lexical activation (Berghoff et al., 2021), facilitated the distinction of L2 phonemic contrasts (White et al., 2015), and increased monolingual-like pronunciation (Maddah and Reiterer, 2018), intonation (Jun and Oh, 2000) and stress (Konishi et al., 2018). Furthermore, neural representations change with L2 proficiency (see Pliatsikas et al., 2020, for a review) and higher proficiency L2 learners activate the same areas in the brain as monolinguals (Vingerhoets et al., 2003). Though numerous studies investigated the role of proficiency with late bilinguals, to our knowledge, only five online studies examined proficiency in early bilinguals. Bice and Kroll (2021) investigated the role of proficiency and working memory on grammatical judgments in HSs and monolinguals. They found that HS showed smaller P600 and N400 effects (i.e., sensitivity to syntactic and semantic violations) than monolinguals, and that ERP variation for grammatical judgments was mostly caused by proficiency (a fluid variable) in the HSs and by working memory (a stable variable) in the monolinguals: Wartenburger et al. (2003), Hervais-Adelman et al. (2018), Sagarra and Rodríguez (2022), and Sagarra et al. (2021) also reported how proficiency affected the ways that HSs processed their heritage language. Because these studies investigated proficiency and AoO, we cover them at the end of the background section as part of our review of studies that investigate multiple language experience variables. Taken together, L2 and HS studies suggest that higher proficiency facilitates morphosyntactic and syntactic processing. Although many studies examined the role of proficiency during language processing in bilinguals, only a few studies explored usage-based measures (Surrain and Luk, 2019). Next, we will summarize bilingual studies including these measures.

Language use is an important component of language processing and acquisition (Ranta and Meckelborg, 2013). L2 studies showed that greater L2 use facilitated monolingual-like L2 morphosyntactic processing (Faretta-Stutenberg and Morgan-Short, 2018), sensitivity to gender code-switching rules (Beatty-Martínez et al., 2020), L2 grammar development (Isabelli-García and Lacorte, 2016), L2 auditory production (Muñoz, 2014), reduction of foreign accents (Abu-Rabia and Kehat, 2004), and discrimination of consonants (Black et al., 2020) and vowels (Flege and MacKay, 2004). Similarly, HS studies demonstrated that greater language use facilitated monolingual-like syntax (Schmidd, 2022), pronunciation (Lloyd-Smith et al., 2019) and reduction of foreign accents (Yeni-Komshian et al., 2000). Pereira Soares (2022) reported that early AoO and greater language use increased functional brain connectivity in HSs and L2 learners; however, the HSs showed greater connectivity and inhibitory control than the learners. Four online HS studies did not measure language use, but their findings appeared to be attributable to language use and exposure. These HS studies employed written tasks, using self-paced reading (Foote, 2011), eye-tracking (Keating, 2022; Parshina et al., 2022), and ERPs (Caffarra et al., 2017). In Foote, HSs and bilingual native speakers raised abroad were equally sensitive to gender and number agreement violations. In Keating, sequential bilinguals were more perceptive to gender agreement violations than simultaneous bilinguals because sequential bilinguals typically use

their heritage language longer than simultaneous bilinguals. In Parshina et al., HSs and L2 learners predicted the gender of an upcoming noun, while only the HSs predicted its number; importantly, the HSs benefited from higher literacy experience. Finally, Caffarra et al. found that gender to gender agreement violations increased with greater language use for opaque nouns (opaque nouns mark gender lexically), but with higher language dominance mostly for transparent nouns (transparent nouns mark gender morphologically). Because HSs perform worse in written than auditory tasks (Bowles, 2011), it is important to examine the four HS studies employing auditory eye-tracking tasks to investigate syntactic predictions (Sekerina and Trueswell, 2011; Jegerski and Sekerina, 2020) and morphosyntactic predictions (Fuchs, 2021; Sagarra et al., 2021). In Sekerina and Trueswell, HSs were slower in processing contrastive focus than monolinguals, due to the HSs having used their heritage language less than the monolinguals. In Jegerski and Sekerina, HSs and L2 learners raised abroad were equally sensitive to the Spanish object marker *a*, showing that using Spanish for a longer period of time can compensate for a later AoO. In Fuchs (2021, 2022), HSs and native speakers of Spanish and Polish used lexical gender cues to make gender agreement predictions. Considering that native speakers use lexical gender cues even with gender transparent nouns (Zeller et al., 2022) and that L2 learners struggle using these cues (see Lemmerth and Hopp, 2019, for a review), we can conclude that the HS advantage over the L2 learners must be due to the HSs' more extensive experience with their heritage language. However, Fuchs did not measure proficiency or use and could not determine whether their HS advantage was due to an earlier AoO, higher proficiency (proficiency was measured with self-reports and with accuracy in producing nouns with the correct gender), or greater frequency of use. Sagarra and Varela addressed this limitation by teasing apart the effects of AoO, proficiency and frequency of use. We describe this study at the end of the background section.

The studies reviewed thus far investigated the role of AoO, proficiency, or use, on bilingual language processing and learning in separate sample pools. Studies that have examined these variables within the same pool have produced different outcomes. We first review studies with L2 learners. Muñoz (2014) found that higher L2 use promoted L2 auditory production more than AoO. Hartshorne et al. (2018) reported that the effects of age, years of experience, and age of exposure in 680,333 participants revealed a late critical period of 17.4 years old to acquire new syntax. In contrast with these two offline studies, L2 neurocognitive studies offer a consistent picture regarding the benefits of greater language use. For instance, white matter microstructure—linked to improved nerve-impulse conduction and working memory function—changed: (a) with greater L2 use, rather than with earlier AoO or higher L2 proficiency (Del Maschio et al., 2020); and (b) with later AoO, a clear consequence of L2 use (Nichols and Joanisse, 2016; DeLuca et al., 2019). Similarly, subcortical structures associated with language control are shaped by longer L2 use (DeLuca et al., 2019). Relevant to our study, Fedeli et al. (2021) observed different effects of AoO, proficiency, and use on structural adaptations in the brain: AoO and L2 use modulated brain areas related to cognitive control, L2 proficiency affected those linked to word learning and language selection, and L2 use influenced those involved in overall comprehension and production. Taken together, the L2 studies reviewed in this paragraph suggest that offline techniques are not sensitive to all language experience nuances and that AoO, proficiency and use should be investigated separately within

the same sample pool, because they restructure the brain differently. Next, we review studies combining AoO and proficiency (Wartenburger et al., 2003; Hervais-Adelman et al., 2018; Sagarra and Rodríguez, 2022), AoO and use (Lloyd-Smith et al., 2019), proficiency and use (Di Pisa and Marinis, 2022), and AoO, proficiency, and use (Sagarra et al., 2021).

Klein et al. (2014) found that bilinguals from birth had a similar brain structure to monolinguals: bilinguals with onset of 3–4 years and later showed thicker cortex in Broca's area. In the same line, Hervais-Adelman et al. (2018) investigated AoO and proficiency effects in bilinguals speaking three or more languages. Greater “multilingual experience”—a composite variable formed by adding AoO (earlier receiving higher weight) and proficiency (more receiving higher weight)—enlarged brain structures associated with language control processes. Because AoO and proficiency were merged, participants with earlier AoO and greater proficiency were treated the same as those with later AoO and less proficiency. Wartenburger et al. (2003) addressed this limitation when examining the effects of proficiency and AoO on grammatical and semantic judgments in HSs and L2 learners with different proficiency levels. Proficiency and AoO affected the neural substrates of L2 processing, but proficiency shaped semantics whereas AoO modulated grammar. These results applied to an explicit task (judgments). Sagarra and Rodríguez (2022) explored the role of AoO and proficiency using an implicit reading eye-tracking task to assess how monolinguals, HSs, and L2 learners processed adjacent subject-verb number agreement. Monolinguals and HSs used articles to a greater extent than L2 learners regardless of proficiency, monolinguals and L2 learners fixated longer on more salient plural and preterit suffixes than less salient singular and present suffixes, and HSs were immune to plural-singular differences. This study did not measure language use, and the written task may have been too challenging for the HSs, because HSs perform poorly on written tasks. For example, HSs are more sensitive to grammatical violations than L2 learners in reading and speaking tasks, but HSs perform worse than L2 learners when completing writing tasks (e.g., Montrul et al., 2013, 2014).

Lloyd-Smith et al. (2019), Di Pisa and Marinis (2022), and Sagarra et al. (2021) addressed these limitations by measuring language use and by employing an auditory task. In Lloyd-Smith et al., Italian monolinguals, German-Italian L2 learners, and Italian-German HSs completed accent rating tasks in Italian and German. All groups were similar in German, but HSs' perceived accent in Italian laid between the monolinguals and the learners. Majority language use did not affect HSs' majority language or heritage language, and heritage language use did not affect HSs' majority language; however, greater heritage language use clearly increased monolingual-like perception of heritage language accent. In Di Pisa and Marinis, Italian controls and HSs completed an elicited production task and a gender assignment task. Higher proficiency increased monolingual-like gender assignment and agreement, but higher use of the heritage language in the home only facilitated gender assignment. In Sagarra and Varela, Spanish monolinguals, and HSs and L2 learners of Spanish listened to sentences with determiner-noun-adjective gender agreement/disagreement while looking at a masculine and a feminine adjective on the screen. The two bilingual groups differed in AoO (before or after puberty) but were matched in proficiency (based on a Spanish proficiency test) and use (weekly percentage of Spanish input,

output and interaction). Eye-tracking data revealed that monolinguals predicted earlier than bilinguals and HSs earlier than L2 learners, and that only the L2 learners struggled using lexical cues (knowing the gender of opaque-gender nouns) and attending to redundant syntactic cues (i.e., suffixes). While higher proficiency and use—but not earlier AoO—produced more predictions in both bilingual groups, these factors affected predictions differently: higher proficiency produced faster predictions and more attention to lexical and syntactic cues in HSs and L2 learners, whereas higher use yielded earlier predictions, more attention to lexical cues in L2 learners, and less attention to syntactic cues in HSs and L2 learners. These findings suggest that language proficiency is different from language use and call for additional online studies to determine the individual contributions of language proficiency and use on other types of associations. Using a visual world eye-tracking task, our study fills this gap by investigating whether AoO, language proficiency and language use modulate how HSs and L2 learners form stress-tense suffix associations *within* words.

4. The study

Predicting what a person will say facilitates processing efficiency, adaptation, and learning (Kaan and Grüter, 2021). Prediction refers to the unconscious pre-activation of pertinent information before hearing it (Barr, 2008) using multiple linguistic cues (e.g., phonological, morphological, syntactic, and semantic) and non-linguistic cues (e.g., auditory, visual, olfactory). As shown in the background section, most prediction studies investigated AoO and proficiency effects between words (e.g., agreement) in L2 learners using written cues. Studies examining language experience effects on within-word predictions via acoustic cues are rare and show that native speakers use suprasegmental information such as tone or stress to predict word endings, but learners do not always make L2 predictions. There is a growing interest in understanding why this occurs. Is it because the learners began acquiring the L2 later in life? Is it due to insufficient L2 proficiency? Or is it a byproduct of how much the learners currently use the L2?

Bilingual studies on morphophonological associations only investigated the role of L2 proficiency. For instance, higher proficiency was found to facilitate the formation of tone-suffix word associations by L1 German-L2 Swedish learners (Swedish, but not German, is tonal) in both Swedish verbs (low tones cueing present suffixes and high tones cueing past suffixes; Schremm et al., 2016) and Swedish nouns (low tones cueing singular suffixes and high tones cueing plural suffixes; Gosselke Berthelsen et al., 2018). Instead of using a Swedish proficiency test, Schremm et al. employed the university entry placement test score, and Gosselke Berthelsen et al. used self-ratings. Sagarra and Casillas (2018) administered a Spanish proficiency test and an auditory eye-tracking task to L1 English-L2 Spanish learners. Advanced, but not beginning, learners predicted stress-tense suffix associations (lexical stress in English and Spanish differ in realization, functional load, and frequency) in Spanish verbs (stressed initial syllables cueing present suffixes and unstressed ones cueing past suffixes). Similar findings were observed in a gating task containing verbs with noise replacing suffixes.

Despite studies showing the effects of AoO and language use on morphosyntactic and phonological processing, the role of these

variables on morphophonological prediction within words is unknown. Recent studies with L2 learners who are professional simultaneous interpreters suggest that language use and cognitive resources impact stress-suffix predictions in bilinguals. First, interpreters predicted faster than non-interpreters of the same L2 proficiency level (Lozano-Argüelles et al., 2022), due to their extensive experience making predictions while interpreting. Second, verbal working memory facilitated predictions in monolinguals and interpreter L2 learners, but not non-interpreter L2 learners (Lozano-Argüelles et al., 2022). To determine whether language use also affects stress-suffix predictions in early bilinguals, we recorded the percentage of time participants used Spanish on a weekly basis (see Materials for more information about this measure). Additionally, we compared HSs to L2 learners to advance our understanding of AoO effects on bilingual predictions.

Using an implicit auditory eye-tracking task, we investigated whether verb stress (oxytone, paroxytone), AoO (before, after puberty), language proficiency, and language use modulated how Spanish monolinguals, HSs, and L2 learners formed stress-suffix associations. Regarding stress effects, paroxytones are more common in Spanish words (Morales-Front, 2014), but oxytones are more typical in English disyllabic verbs (Chomsky and Halle, 1968). We expect that stress type will not affect the monolinguals' predictions due to ceiling effects, and that HSs' and L2 learners' dominance in English will produce more fixations on targets with oxytones than paroxytones. Concerning AoO effects, we hypothesize that all groups will predict above chance, based on Sagarra and Casillas' (2018) findings with monolinguals and non-beginning L2 learners. But we expect the monolinguals to predict earlier than the HSs and L2 learners, following Sagarra et al. (2021). AoO of English was not included because all the HSs began learning English formally at age 5, when they began kindergarten, and because Lloyd-Smith et al. (2019) found that AoO of the majority language did not affect the majority language or the heritage language. Respecting language proficiency effects, we foresee that higher proficiency will increase fixations to target verbs, considering studies with Spanish L2 learners (Sagarra and Casillas, 2018) and Swedish L2 learners (Schremm et al., 2016; Gosselke Berthelsen et al., 2018). As for language use effects, we anticipate that greater language use will produce more fixations to target verbs. This is in line with studies showing that greater language use facilitates morphosyntactic processing (L2 learners: Faretta-Stutenberg and Morgan-Short, 2018; HSs: Foote, 2011; Caffarra et al., 2017; Keating, 2022) and prediction (HSs: Parshina et al., 2022), as well as L2 sound discrimination (Flege and MacKay, 2004; Black et al., 2020), monolingual-like pronunciation in HSs (Lloyd-Smith et al., 2019), and reduced L2 accent (Guion et al., 2000). Lastly, we postulate that language use will have a stronger impact on prediction than language proficiency in both HSs and L2 learners, but particularly in the HSs. This is because language use, but not AoO or L2 proficiency, changes white matter microstructure (Del Maschio et al., 2020), and because language use restructures brain areas associated with language control (DeLuca et al., 2019; Fedeli et al., 2021).

4.1. Participants

We collected data from 122 individuals: 30 Spanish monolinguals (M; 22 females), 42 HSs (26 females; with Spanish

as the heritage language and English as the majority language), and 50 L2 learners (36 females, L1 English, L2 Spanish). Participants had normal hearing and normal or corrected-to-normal vision and held at least a high school diploma. In addition, they were between 18 and 40 years-old and right-handed. HS data were collected in the U.S. and M and L2 data were collected in Spain. L2 data were gathered in Spain to have L2 learners with high Spanish use comparable to the HSs. The M were born and raised in Madrid, Spain. They spoke English but were not advanced learners, according to self-ratings. Also, they did not speak other languages, and had not lived in a non-Spanish community for more than 2 months. The HS and L2 groups only spoke Spanish and English. HSs were born and raised in the United States, were second generation of immigrants, and had not received formal education in their heritage language, apart from taking Spanish in school. They grew up using Spanish at home and in their neighborhood, and they continued using Spanish in these contexts. Half of them had traveled to their parents' native country. Approximately 30% of the HSs spoke Spanish with friends, 80% listened to music in Spanish, and 40% watched TV in Spanish. The L2 learners began learning Spanish at least 1 h of class per week in middle school and continued in high school and at the university, and had lived in Madrid an average of 38.29 months ($SD = 34.12$).

The bilingual participants completed language *use* and *proficiency* assessments described in the materials section. The *use* and *proficiency* data were fit to separate Bayesian linear models, in order to assess potential group differences.¹ The posterior marginal mean difference between groups on both response variables was compared, using a region of practical equivalence (ROPE) of ± 0.1 . If, for a given measure, the full range of the 95% highest density credible interval (HDI) of the difference estimate fell within the ROPE, the groups were considered to be equivalent. The average Spanish proficiency score was 0.70 ($SD = 0.09$) for the HS group and 0.71 ($SD = 0.14$) for the L2 group, the marginal mean difference was 0.02 [$-0.03, 0.07$], and all the HDI fell within the ROPE. The probability that the effect was positive was 0.77. Regarding Spanish use, the average score was 0.41 ($SD = 0.15$) for the HS group and 0.38 ($SD = 0.16$) for the L2 group, the marginal mean difference was -0.02 [$-0.09, 0.05$], and the HDI fell within the ROPE. The probability that the effect was negative was 0.72. Taken together, we are confident that the groups do not differ in any meaningful way with regard to use or proficiency in Spanish. Table 1 provides descriptive statistics and summarizes the models.

4.2. Materials and procedure

Data collection was conducted individually in a single session. Participants completed four tasks. First, the bilingual groups

¹ In both cases, the response variable, *use* or *proficiency* score, was a proportion. Thus, we used the beta distribution for the model likelihood with a logit linking function. The models included regularizing, weakly informative priors. See the [Supplementary material](#) for full details.

TABLE 1 Language use and proficiency assessments for the HS and L2 bilingual groups.

Metric	HS (n=42)	L2 (n=50)	Contrast	Estimate	ROPE	PD
Proficiency	0.70 (0.09)	0.71 (0.14)	L2 – HS	0.02 [–0.03, 0.07]	1	0.77
Use	0.41 (0.15)	0.38 (0.16)	L2 – HS	–0.02 [–0.09, 0.05]	1	0.72

The table reports the mean and standard deviation, as well as posterior estimates of the marginal mean difference (L2 – HS) and the 95% highest density credible interval (in brackets). The proportion of the posterior density falling within the region of practical equivalence (± 0.1) is reported in the ROPE column. The probability that the effect is of the median's sign is reported in the PD column.

completed a Spanish proficiency test in Qualtrics. The test consisted of a 56-item adapted version of the *Diploma de Español como Lengua Extranjera* (Certificate of Spanish as a Foreign Language) that assessed Spanish grammar and vocabulary knowledge (Sagarra and Herschensohn, 2010). Second, the bilingual groups completed a language background questionnaire with questions regarding age, handedness, languages spoken at home when growing up, AoO, time spent in Spanish-speaking countries, and other languages spoken. Third, the HS and L2 groups filled out a Spanish use questionnaire measuring the percentage of time actively using each language weekly (a combination of input, output, and interaction when talking with friends and family, at work, listening to music, and watching TV).

Lastly, all groups completed an eye-tracking task assessing participants' abilities to use the stress of a Spanish disyllabic verb's first syllable to predict the verb's ending (i.e., the tense suffix) before hearing it. The eye-tracker was an EyeLink 1,000 Plus desktop mount from SR Research (sampling rate: 1 kHz; spatial resolution of 0.32° horizontal and 0.25° vertical; averaged calibration error: 0.25°–0.5°). The task was programmed with SR Research's Experiment Builder software, and the data were extracted with SR Research's DataViewer software. Tracking was monocular (right eye) and followed cyclopean extraction mode. The velocity threshold (the threshold to consider an eye movement a saccade) was 30°/sec, which is the default for cognitive research in Experiment Builder. Shorter eye movements taking place during fixations (e.g., tremors, drifts, and microsaccades) were considered part of the fixation because numerous studies show that they rarely affect the analysis of higher-level structures such as words or phrases (e.g., Ditchburn, 1980). The monitor was a BenQ XL2420TE display monitor at a resolution of 1,920 × 1,080 pixels, and the headphones were Sol Republic 1601-32.

Participants listened to 100 sentences: four practice sentences, 16 experimental sentences, and 80 fillers. Sentences rather than words were used to imitate naturalistic comprehension and increase ecological validity. The practice sentences appeared always in the same order, and the experimental and filler sentences were distributed into 8 blocks. Each block contained six filler sentences and two experimental sentences, one per condition. Sentences were randomized between blocks and pseudo-randomized within blocks to avoid two consecutive experimental sentences of the same condition. We recorded these sentences using a Fostex DC-R302 digital recorder and a Shure SM10A head-mounted microphone in a Whisper room 6,084 E sound booth at a sampling rate of 44.1 kHz and 16-bit quantization. A Castilian Spanish female speaker unaware of the purpose of the study recorded all the sentences three times in three different

pseudo-randomized orders; we chose the clearest pair of the last two repetitions. She used a standard intonation and a consistent rate of 4.37 ($SD = 0.68$) syllables per second and 4.17 ($SD = 1.14$) seconds per sentence. Intensity was normalized to ~75 dB and 100 ms of leading and trailing silence added using Praat (Boersma and Weenink, 2021).

All sentences were grammatical and consisted of 5–14 words. Filler sentences contained anaphora, gender agreement, and idiomatic expressions. Experimental sentences were five words long and followed an SVO word order, with animate noun subjects and inanimate noun objects. Subjects and objects were 2–4 syllables long. Experimental verbs were disyllabic third-person singular regular transitive -ar verbs with a CVC-CV syllabic structure. The mean duration of the verbs was 424 ms ($SD = 42.22$, CI [408.78, 439.22]). Breaking down the verb duration into syllables, the first syllable had a mean duration of 308.38 ms ($SD = 52.03$, CI [284.74, 321, 20]) and the second syllable of 115.63 ms ($SD = 32.74$, CI [108.53, 133.53]). The second syllable disambiguated the tense segmentally. Experimental sentences had two conditions: paroxytone/present and oxytone/preterit (e.g., *El ladrón salta/saltó la valla* “the thief jumps/jumped over the fence”) and only differed in the verb. The visual stimuli consisted of a present and a preterit verb displayed side by side on the screen. Their positions were counterbalanced across participants and trials. We chose words rather than images because (1) it is difficult to illustrate present and past actions, (2) it is uncertain what word participants truly activate when they see an object, and (3) phonological competitor effects are stronger with words than pictures (Huettig and McQueen, 2007; Ito et al., 2017). The written words for the filler sentences consisted of inanimate nouns for the anaphora fillers, descriptive adjectives for the gender agreement fillers, and ending nouns for the idiomatic fillers.

The procedure of the eye-tracking task was as follows: participants were first randomly assigned to one of two versions of the task. Each version contained only one of the two conditions of each verb pair (e.g., if *salta* “s/he jumps” (paroxytone/present) appeared in version 1, then *saltó* “s/he jumped” (oxytone/preterit) appeared in version 2). Both versions had the same number of practice, filler, and experimental trials. Participants rested their heads on a chin rest, completed a 9-point grid calibration task, and received task instructions. Next, participants completed the practice trials, followed by the experimental trials. For each trial, participants saw a + drift correction sign, followed by a 250 ms blank screen, saw two verbs side by side for 1,000 ms, listened to the sentence, and chose the verb on the screen they heard as soon as possible by pressing the left- or right-shift key. Upon pressing either key, a rectangle appeared around the selected verb. Participants did not

receive feedback after completing the task. We set up response recording to register only when the keypress happened at or after the onset of the verb. Key presses did not stop the sound file. After each sentence, a blank screen appeared for 500 ms, and the next trial began. After the eye-tracking task, participants completed a test assessing their knowledge of the meaning of the experimental verbs (e.g., to know that *salta* means to jump) and the tense suffixes (e.g., to know that *salta* is present). Participants saw a list containing the experimental Spanish verbs and a list containing English verbs. Their task was to match each Spanish verb with the correct English translation.

4.3. Statistical analyses

We fit a series of Bayesian regression models to examine the time course data. The primary model was a Generalized Additive Mixed Model (GAMM, Winter and Wieling, 2016; Sóskuthy, 2017). GAMMs are useful for scrutinizing non-linear data, such as that typically associated with eye-tracking.² In subsequent analyses, we summarized the posterior predictive distribution to make inferences about the relationships between speaker groups, lexical stress, language proficiency, and language use. Given the distinct nature of some of these analyses, we provide a brief description of the statistical approach at the beginning of each subsection. For all models, we employed regularizing, weakly informative priors (Gelman et al., 2017).³ In most cases, we used the following formula to establish a region of practical equivalence (ROPE) around a point null value (see Kruschke, 2018):

$$ROPE = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

We report mean posterior point estimates for parameters of interest, along with the 95% highest density credible interval (HDI), the percent of the region of the HDI contained within the ROPE, and the probability of direction for each effect (PD). For statistical inferences, we focus on estimation rather than decision-making rules, though, generally, a posterior distribution for a parameter β in which 95% of the HDI falls outside the ROPE and a high PD (i.e., values close to 1) are taken as compelling evidence for a given effect. We conduct all analyses using R (version 4.2.1) and fit all models using the probabilistic programming language *stan* via the R package *brms* (Bürkner, 2017, 2018).

² GAMMs represent an extension to the linear model framework that allow non-linear functions called factor smooths to be applied to predictors. In this sense, the predictors can be classified into two types: parametric terms (equivalent to fixed effects in hierarchical model terminology) and smooth terms. Random smooths are conceptually similar to random slopes and intercepts in the mixed-effects regression framework (Winter and Wieling, 2016). Thus, they allow the by-subject trajectory shapes to vary as a function of a parametric effect and are essential in avoiding anti-conservative models.

³ See Supplementary material for detailed information.

5. Results

The analyses are divided into three sections. First, we describe the trajectories of the time course. Then, we evaluate suffix prediction at the target syllable offset. Lastly, we consider the effects of language use and proficiency.

5.1. The time course of morphological processing

Our analysis of the time course data from the eye-tracking task models measures how the probability of fixating on target items changes over time and under different suprasegmental conditions. We down-sampled the data to bins of 50 ms which were centered at the offset of the first syllable of target items. The time course of fixation used for analysis ranged from 200 ms before target syllable offset to 600 ms after. We chose this window because it captures the portion of the time course in which target fixations began to steadily increase from chance. Figure 1 illustrates the trajectories of the monolinguals, HSs, and L2 learners as a function of lexical stress. In both panels, we see that the probability of fixating on the target hovers around 0.5 and begins to increase as time increases. Notably, we also observe that the lines are not overlapping. The monolingual group begins to fixate on the target earlier in the time course in both paroxytones and oxytones. Essentially, the HSs and L2 groups are phase shifted to the right, representing later target fixations.

Given the binary nature of the dependent variable (“i.e., fixations on the target word vs. elsewhere”), we assumed that the likelihood was going to be binomially distributed. The model assessed target fixations as a function of the parametric terms *group* (monolingual, HSs, L2), *stress* (paroxytone, oxytone), and a nonlinear function of time. Both *group* and *stress* were set as ordered variables with monolinguals and paroxytones coded as “0.” We implemented cubic regression splines with four basis knots: (a) as a reference smooth to time, (b) as a difference smooth to time conditioned on stress, and (c) as a random smooth for each participant conditioned on time. Thus, the trajectory of the monolinguals’ target fixations to paroxytone words (e.g., *CANta*) served as the baseline, and we could compare it to the trajectories of the other groups. The forest plot in Figure 2 illustrates the model summary (see Supplementary material for the complete summary in table form).

To quantify and assess the between-group differences over time, we used the posterior predictive distribution to calculate posterior pairwise difference smooths. Figure 3 illustrates these pairwise comparisons over the time course in the probability space. Overall, the analysis shows that the monolingual group fixates on targets earlier than the HS and L2 groups in both stress conditions over the time-window we selected. The HS-L2 comparison suggests that the HS group fixates on targets slightly more and earlier in paroxytone condition, but the opposite is true in the oxytone condition.

5.2. Prediction at target offset

In order to assess the participants’ ability to predict suffixes, we used the posterior predictive distribution of target fixations 200 ms after the target syllable offset (i.e., the minimum time necessary to

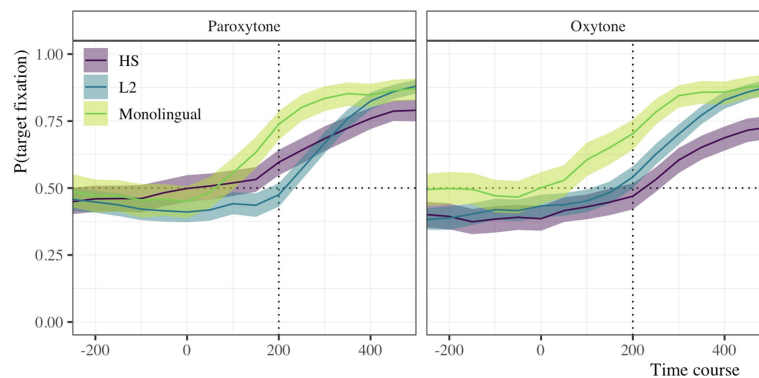


FIGURE 1

Time course of raw target fixation data as a function of stress condition (paroxytone, oxytone) for monolingual, HS, and L2 groups. Transparent ribbons represent 99% bootstrapped confidence intervals.

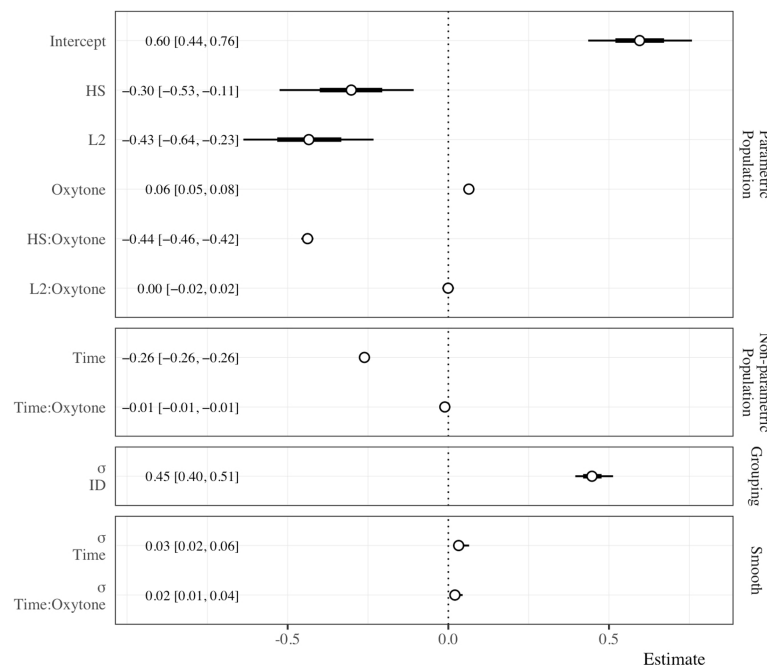


FIGURE 2

Forest plot of the omnibus GAMM. The horizontal axis represents the models estimates in log-odds. The vertical axis lists the terms estimated in the model. The points illustrate the posterior mean along with the 66% and 95% HDI. The vertical faceting separates the estimates into parametric and non-parametric population-level effects, group-level effects, and smooth terms.

plan and launch a saccade, see Fischer, 1992). We considered that the probability that target fixation was greater than chance at this time point for each group in each stress condition and implemented a ROPE of 0.01 around a point null, chance value of 0.5. Figure 4 illustrates posterior distributions of target fixations.

All groups fixated on targets above chance 200 ms after the target syllable offset with the exception of the HS group in the oxytone condition ($\beta=0.48$, $\text{HDI}=[0.45, 0.52]$, $\text{ROPE}=0.97$, $\text{PD}=0.85$). Approximately 97% of the HDI fell below the upper bound of the ROPE and there is an 85% chance that the estimate is below 0.5. Additionally, a small portion of the posterior probability mass of the L2 group in the paroxytone condition fell within the ROPE ($\beta=0.54$, $\text{HDI}=[0.51, 0.57]$, $\text{ROPE}=0.01$, $\text{PD}=1$), though, given the model, the data and our prior assumptions, the effect is nearly certain to be above 0.5.

Subsequently, we assessed the rate of target fixations at the same time point (i.e., 200 ms after the offset of the target syllable). While the previous assessment evaluates *if* participants fixate on targets before hearing a critical suffix, this analysis sheds light on *how fast* target fixations occur by calculating the partial derivative (i.e., slope) of the trajectory at this time point. The top panels of Figure 5 show the marginal slope estimates for each group for paroxytone and oxytone words. The bottom panels of Figure 5 provide pairwise group comparisons in each condition. The monolingual group demonstrates a slower rate of target fixation (i.e., a less steep slope) than the HS group for paroxytones ($\beta=-0.004$, $\text{HDI}=[-0.008, -0.001]$, $\text{ROPE}=0$, $\text{PD}=0.999$) and oxytones ($\beta=-0.006$, $\text{HDI}=[-0.010, -0.004]$, $\text{ROPE}=0$, $\text{PD}=1$). This is also the case when compared with L2 learners (paroxytones: $\beta=-0.006$, $\text{HDI}=[-0.009, -0.003]$, $\text{ROPE}=0$,

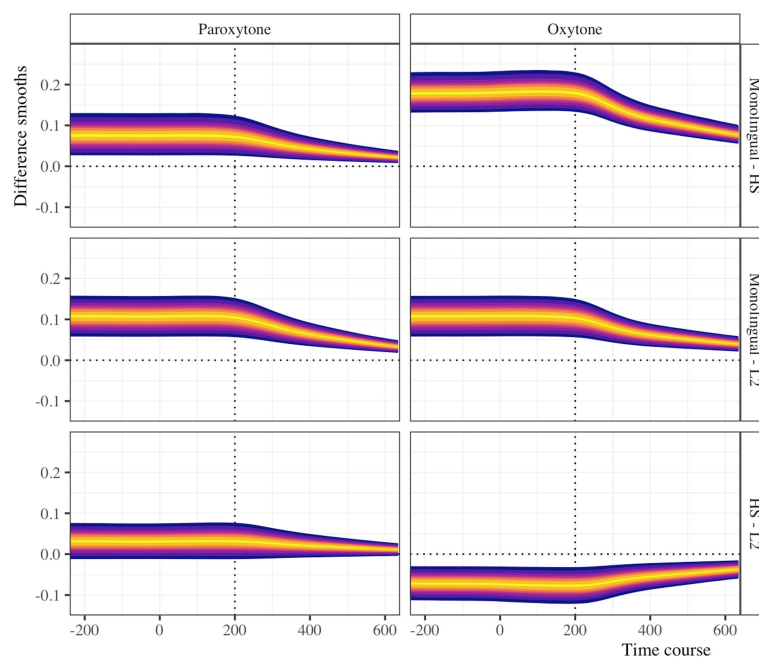


FIGURE 3

Pairwise difference smooths for paroxytone and oxytone items. From dark to light, the colors represent 95%, 80%, 70%, 60%, 50%, 35%, and 10% highest density credible intervals.

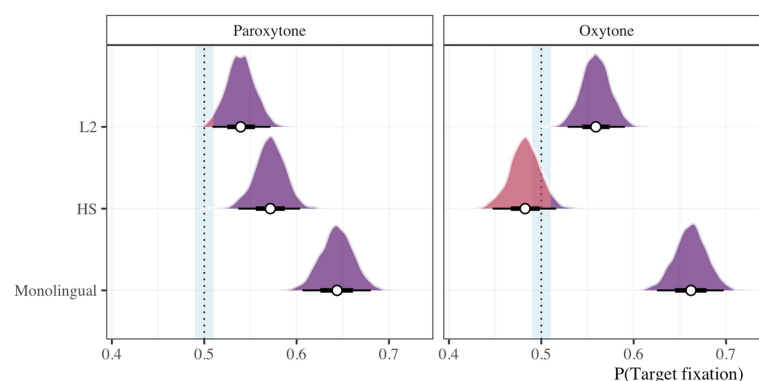


FIGURE 4

Proportion of target fixations 200ms after the offset of 1st syllable for monolingual, HS, and L2 groups in paroxytone (CANTo) and oxytone (canTO) conditions. The vertical dotted line marks chance (50%) surrounded by a $\pm 1\%$ region of practical equivalence (ROPE). The density mass of a posterior distribution that falls below the upper bound of the ROPE is displayed in red and values above this threshold are purple.

PD=1; oxytones: $\beta = -0.006$, $\text{HDI} = [-0.009, -0.003]$, $\text{ROPE} = 0$, $\text{PD} = 1$). Upon evaluating the HS and L2 groups, we do not find compelling evidence that either group has a faster rate of target fixation. In the paroxytone condition, the L2 group might be slightly faster, but nearly half the HDI fell within the ROPE ($\beta = -0.001$, $\text{HDI} = [-0.003, 0.000]$, $\text{ROPE} = 0.514$, $\text{PD} = 0.922$). In the oxytone condition the opposite is true. That is, the L2 group may have been slightly slower, but, again, a large portion of the HDI fell within the ROPE ($\beta = 0.001$, $\text{HDI} = [0.000, 0.002]$, $\text{ROPE} = 0.697$, $\text{PD} = 0.957$). Taken together, we do not believe there is compelling evidence that the rate of target fixation differs between the HS and L2 groups. Additional plots and a table summary are provided in the [Supplementary material](#).

5.3. Proficiency and use

To assess the effects of language proficiency and use, we took the subset of the HS and L2 data from the time bin that corresponded with 200 ms after the offset of the initial syllable in the target items. We calculated the proportion of target fixations for each participant, in each condition and submitted these proportions to a zero-inflated beta regression model.⁴ The

⁴ More information regarding zero-inflated beta regression is available in the [Supplementary material](#).

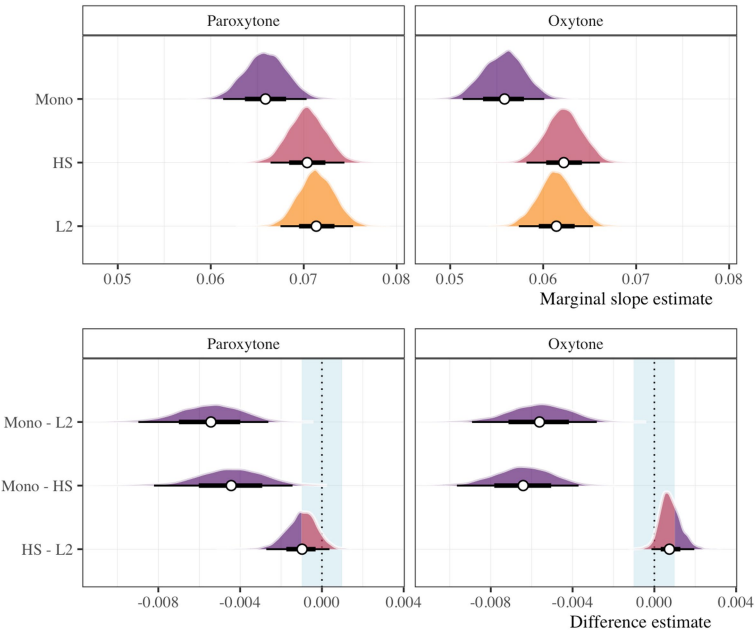


FIGURE 5 Marginal slope estimates (top) and pairwise difference estimates (bottom) for monolingual, HS, and L2 groups 200ms after the target syllable offset in paroxytone (CANto) and oxytone (canTÖ) conditions. Points represent posterior means along with the 66 and 95% HDI. In the bottom panels, the vertical blue rectangle illustrates a ROPE of ± 0.001 . Posterior mass falling within the ROPE is depicted in red and values outside the ROPE are in purple.

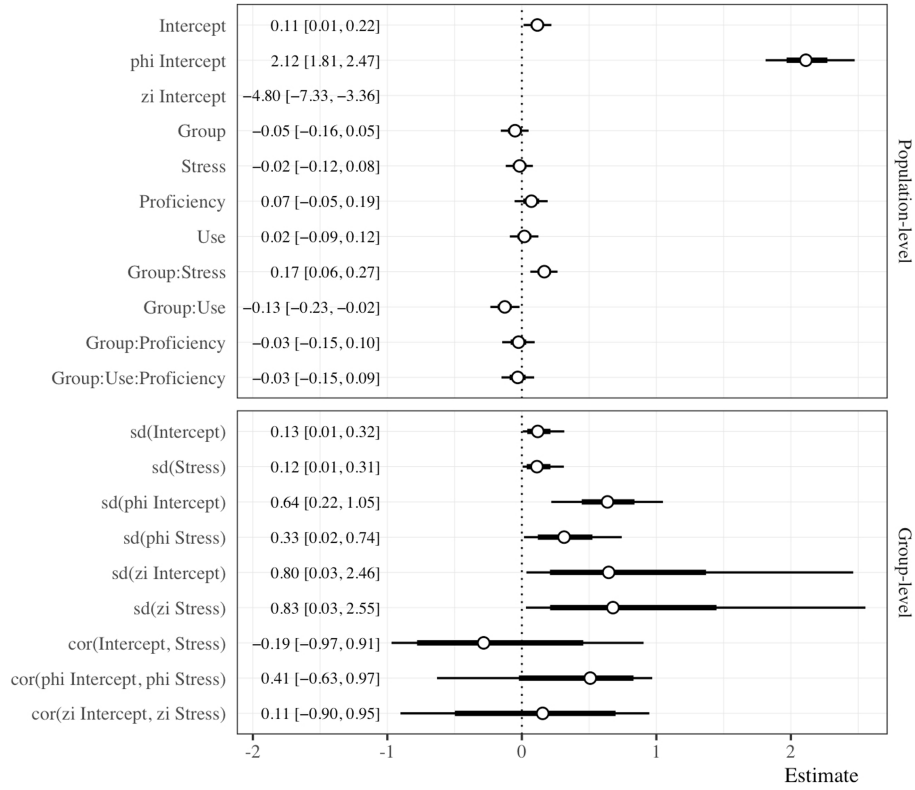


FIGURE 6 Forest plot of the zero-inflated beta regression. The horizontal axis represents the models estimates in log-odds. The vertical axis lists the terms estimated in the model. The points illustrate the posterior mean along with the 66% and 95% HDI. The vertical faceting separates the estimates into population-level and group-level effects.

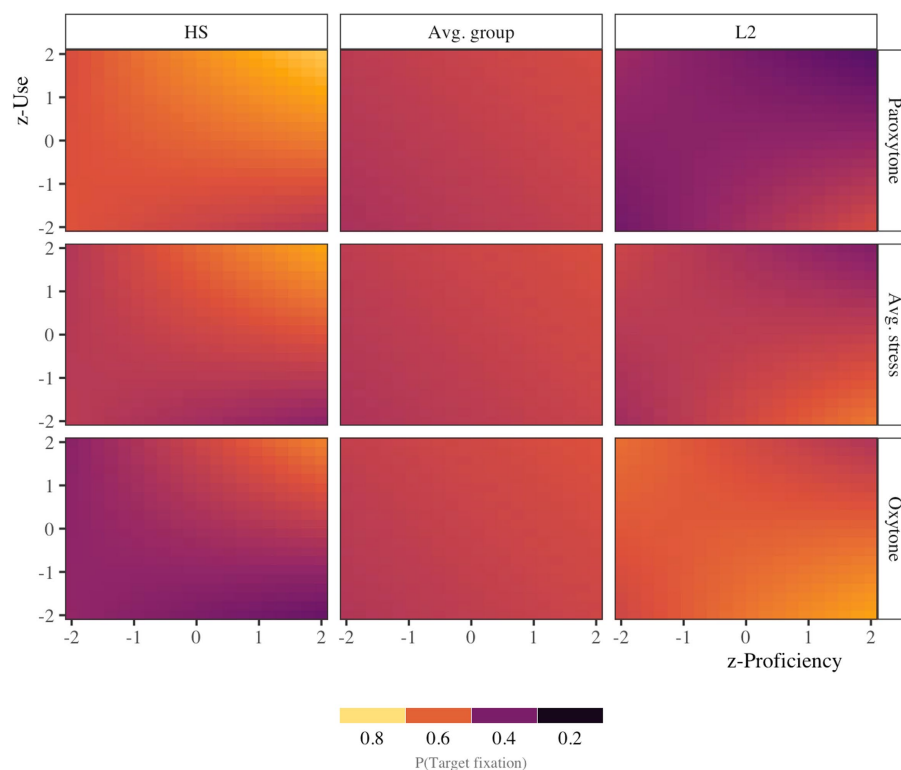


FIGURE 7

Heatmap of target fixations 200ms after target syllable offset. The heatmap illustrates the marginal effects of normalized proficiency and use scores. The top rows illustrate model estimates for the paroxytone condition, the middle rows marginalize over stress conditions, and the bottom rows represent model estimates for the oxytone condition. Moving from left to right, the first column provides estimates for the HS group, the middle column marginalizes over groups, and the rightmost column presents estimates for the L2 group. The vertical and horizontal axis display standardized language use and language proficiency (± 2 SD), respectively.

outcome was modeled as a function of group (HS, L2), stress (paroxytone, oxytone), use, and proficiency. Group and stress predictors were sum coded ($-1, 1$) and the continuous predictors were standardized (i.e., converted to z-scores), thus the model intercept provided an estimate of target fixation marginalizing over *group* and *stress*, with *use* and *proficiency* equal to 0 (i.e., at the unstandardized mean). The model included all two-way interactions as well as the group by use by proficiency three-way interaction. We included a group-level effect for participants with a varying slope for stress. A full description of the model specification and priors is available in the [Supplementary material](#).

The overall probability of fixating on a target was approximately 0.53 (Intercept: $\beta = 0.11$, HDI = [0.01, 0.22], ROPE = 0.39, PD = 0.99). There was no main effect for group ($\beta = -0.05$, HDI = [-0.16 , 0.05], ROPE = 0.83, PD = 0.83), nor stress ($\beta = -0.02$, HDI = [-0.12 , 0.08], ROPE = 0.97, PD = 0.63), though the two predictors did interact ($\beta = 0.17$, HDI = [0.06, 0.27], ROPE = 0.08, PD = 1). Holding proficiency and use constant at their mean, the HS group fixated on targets at a higher rate in the paroxytone condition ($\beta = 0.57$, HDI = [0.45, 0.69]) than in the oxytone condition ($\beta = 0.48$, HDI = [0.38, 0.58]). The opposite was true for the L2 group (paroxytone: $\beta = 0.54$, HDI = [0.42, 0.64]; oxytone: $\beta = 0.47$, HDI = [0.35, 0.61]). The forest plot provided in [Figure 6](#) summarizes the model. A model summary table is available in the beta regression subsection of the [Supplementary material](#).

There was also evidence of a group \times use interaction ($\beta = -0.13$, HDI = [-0.23 , -0.02], ROPE = 0.3, PD = 0.99). Although approximately 30% of the HDI fell within the ROPE, the model, and our prior assumptions, we are 99% certain that the interaction effect is negative. [Figure 7](#) provides a heat map that illustrates the relationship between proficiency, use, and stress in the bilingual groups. For the HS group, one observes higher target fixations (lighter colors) in the upper right-hand corners of each panel. That is to say, HSs fixated more on targets higher levels of use and proficiency, particularly in the oxytone condition. Target fixation was higher, nearly across the board, in the paroxytone condition. For the L2 group, on the other hand, one observes a higher propensity to fixate more on targets in the lower right-hand corners of each panel (lighter colors), when proficiency is higher, but not necessarily language use. Unlike the HS group, the L2 group seldom predicted in the paroxytone condition (upper right panel).

6. Discussion

We investigated whether AoO, language proficiency, and language use influenced how Spanish HSs and L2 learners form stress-tense suffix associations in Spanish disyllabic verbs, using an auditory eye-tracking task. Regarding the effects of stress and AoO, holding proficiency and use means constant, all groups fixated on target verbs above chance before hearing the syllable with the suffix

in all conditions. The only exception occurred with the HS group in the oxytone condition. Furthermore, HSs predicted less with oxytones, whereas L2 learners predicted less with paroxytones. Monolinguals fixated on targets more and earlier, but at a slower rate, than bilinguals in all conditions, and HSs more and earlier than L2 learners in paroxytones. However, HSs predicted later than L2 learners with oxytones. With respect to proficiency and use, HSs with higher proficiency and greater language use fixated on target oxytones more. Yet, while greater use was more important than higher proficiency for HSs, L2 learners with higher proficiency fixated on target paroxytones more, and the amount of use did not matter. These results show that L2 learners can acquire stress-suffix associations absent in their L1 after puberty, and that their ability depends on their L2 proficiency level rather than their AoO or L2 use. Next, we discuss the relevance of our findings with respect to stress type (oxytone, paroxytone), AoO (before, after puberty), language proficiency, and language use.

6.1. The effects of stress

Paroxytone words have the stress on the penultimate syllable (e.g., *SALta* “she/he jumps”), whereas oxytone words have it on the last syllable (e.g., *salTÓ* “she/he jumped”). The majority of Spanish words (Morales-Front, 2014) and English words (Kelly and Bock, 1988) are paroxytone. However, in English, oxytones and paroxytones are equally frequent in disyllabic uninflected words (Clopper, 2002), and oxytones are more frequent than paroxytones in disyllabic verbs (Chomsky and Halle, 1968). Relevant to our study, in Spanish, third person singular regular verbs are more frequent in present tense (paroxytone; 30,667/1,000,000) than preterit tense (oxytone; 12,030/1,000,000; CORPES, Real Academia Española). Furthermore, Spanish and English have contrastive stress, but suprasegmental cues have a greater functional load in Spanish than in English. Considering these data, we hypothesized that the monolinguals would predict regardless of stress type, that the HSs would predict more than the L2 learners, and that the HSs and L2 learners would predict more with oxytones than paroxytones because oxytones are more common in disyllabic verbs in their dominant language, English. The results of the tasks confirmed our hypothesis with the monolinguals. This group used stress to predict suffixes before hearing them above chance with both paroxytones and oxytones. These results are in line with studies showing that Swedish speakers use tone to predict number (singular vs. plural; Roll et al., 2010, 2013; Söderström et al., 2016) and tense (present vs. past; Roll, 2015; Söderström et al., 2016), and that Spanish speakers use stress to predict tense (present vs. past) with both paroxytones and oxytones (Sagarra and Casillas, 2018).

HS and L2 data did not support our hypothesis. Holding proficiency and use means constant, the L2 learners predicted above chance with both paroxytones and oxytones, in line with Sagarra and Casillas (2018). However, the HSs only predicted above chance with paroxytones, the L2 learners predicted more and earlier than the HSs with oxytones, and the HSs predicted more and earlier than the L2 learners with paroxytones. The differences between the HSs and the L2 learners can be explained by HSs' early AoO, more years of exposure to Spanish, or higher number of lexical competitors. First, the difference between HSs and L2 learners cannot be linked to AoO

because L2 learners predicted above chance in both the English-preferred condition (oxytones) and the Spanish-preferred condition (paroxytones). Second, the difference is likely not due to HSs' longer accumulated exposure to Spanish paroxytones (lexical frequency), the preferred condition in Spanish, because the HSs had trouble with oxytones, the preferred pattern in their dominant language, and because Sagarra et al. (under review) found no differences between English and Mandarin learners of Spanish, although lexical tone T4—which resembles paroxytones—is more frequent in Mandarin, and suprasegmentals have a higher functional load in Mandarin and Spanish than in English. Third, we attribute the differences between HSs and L2 learners to their current use of Spanish, in particular to lexical competition due to co-activation. The learners have a native lexicon and later-acquired, more fragmented L2 lexicon that makes the L1 lexicon dominate. This explains the learners' stronger and faster activation of oxytones, the preferred pattern in English. In contrast, the HSs have two L1 lexica that rapidly activate when hearing words. This produces higher competition in oxytones because English is their dominant language and English has more oxytone than paroxytone candidates, making it harder to use oxytone predictors (predictions are stronger for word beginnings that evoke few lexical competitors, e.g., Söderström et al., 2016). Our findings support theoretical models explaining HSs' variability and divergence from monolinguals in terms of lexical competition due to co-activation (e.g., Hatzidaki et al., 2011; Giezen and Emmorey, 2016). Our results also show that such competition exerts a greater influence on how HSs access words than lexical frequency, contrary to theoretical accounts proposing that lexical frequency offsets lexical competition (e.g., Hur et al., 2020; Perez-Cortes, 2020). Importantly, these studies employed offline tasks and examined morphosyntax (grammatical gender agreement) and syntax (mood). Finally, our results are in line with recent neurocognitive studies showing that higher language use increases functional brain connectivity and inhibitory control (see Pereira Soares, 2022, for a review).

One may argue that, because the participants saw a specific present-preterit verb pair before listening to each sentence, they focused on calibrating the frequency of the verbs on the screen to prioritize the most frequent verb pair and they ignored other lexical competitors. To explore this possibility, we calculated the lexical frequencies of the experimental verbs with the *LEXESP* dictionary of frequencies (Sebastián-Gallés, 2000). The experimental paroxytone verbs have a higher lexical frequency than their oxytone counterparts: 131.44 for paroxytones vs. only 44.94 for oxytones. If lexical frequency of the experimental verbs drives prediction, we would expect both bilingual groups to predict more and earlier with paroxytones than oxytones. However, (a) the L2 learners predicted equally with paroxytones and oxytones, (b) the L2 learners predicted oxytones more and earlier than the HSs, and (c) the HSs predicted paroxytones more and earlier than the L2 learners. These three findings demonstrate that the groups did not rely on the experimental verbs' lexical frequency. Although the task reduced the lexical cohort to two members, the experiment tapped into more automatic processes of lexical access, making participants considered additional lexical competitors.

Lexical frequency is relevant once a prediction has been made (token frequency), but phonotactic frequency determines what competitors are considered as prediction unfolds (type frequency).

Sagarra and Casillas (in progress) investigated the role of both phonotactic and lexical frequency on suprasegmentals (oxytone, paroxytone stress) and segmentals (CVC, CV syllabic structure) in advanced HSs and L2 learners. Eye-tracking data showed that higher phonotactic frequency increased fixations on targets in the HSs in all conditions, but not the L2 learners; also, lexical frequency did not affect HSs or L2 predictions. HS, but not L2 learners, consider phonotactic frequency when predicting, due to HSs' longer experience with the target language.

Taken together, the findings of Sagarra and Casillas (in progress) and of the current study indicate that HSs' lexical access depends on: (1) number of lexical competitors (HSs consider more competitors than monolinguals and L2 learners because HSs activate two L1 lexica); and (b) type frequency (higher phonotactic frequency affects HSs, but not L2 learners). Token frequency (lexical frequency) may also influence prediction (more frequent words tend to be more strongly activated, e.g., Roll et al., 2013), but cohort size seems to be the most important factor for HSs. Our data also demonstrate that we store suprasegmental information as we access words during comprehension and production, and we prioritize this information over semantic information when we start listening to a word to anticipate the ending.

6.2. The effects of AoO

To determine whether adults are able to make L2 stress-suffix associations absent in their L1 if they begin learning the L2 after puberty, we compared adult HSs and adult L2 learners with the same Spanish proficiency level and current use of Spanish. The results support our hypothesis that monolinguals, HSs, and L2 learners would predict above chance. Our L2 data are in consonance with studies indicating that non-beginning L2 learners predict tense suffixes using tone (Schremm et al., 2016) and stress information (Sagarra and Casillas, 2018), and number suffixes using tone information (Gosselke Berthelsen et al., 2018). Although there are no studies on lexical prediction with HSs, our HS data are consistent with studies suggesting that HSs make both syntactic predictions (Sekerina and Trueswell, 2011; Jegerski and Sekerina, 2020) and morphosyntactic predictions (Fuchs, 2021; Sagarra et al., 2021).

Without the HSs, our results could be erroneously interpreted as monolinguals predicting more and earlier than L2 learners due to the learners' late AoO. The presence of a HS group was necessary to discard this supposition in four ways. First, the monolinguals predicted more and earlier than the HSs, even though both groups began acquiring Spanish at birth. Second, the L2 learners predicted more and earlier with oxytones than the HSs (and the opposite pattern applies to paroxytones), even though the learners began learning Spanish years later than the HSs. Third, although AoO is later in the learners than the HSs and similar in the monolinguals and the HSs, only the monolinguals and the L2 learners predicted above chance with oxytones. Finally, the bilingual groups predicted faster than the monolinguals, and all the bilingual groups predicted at equal speed, regardless of the AoO differences between both bilingual groups. Collectively, these findings suggest that prediction differences between monolinguals and L2 learners and between HSs and L2 learners may

not be due to AoO but to differences in the amount and type of experience each group has had with Spanish.

The absence of AoO effects is on par with studies showing no differences between HSs and L2 learners using self-paced reading methodology (Foote, 2011; Rodríguez and Reglero, 2015), eye-tracking technique (written: Sagarra and Rodríguez, 2022; auditory: Sagarra et al., 2021), and ERPs (Wartenburger et al., 2003; Martohardjono et al., 2017). Singleton and Leśniewska (2021) argued that the critical period hypothesis is unfalsifiable and therefore irrelevant, because separating early and late bilinguals is fictional, considering the enormous degree of variability of individual language experiences in each of these two groups. These proposals are timely, given recent neurocognitive studies demonstrating how white matter microstructure changes with later AoO (Nichols and Joanisse, 2016; DeLuca et al., 2019), as well as with greater L2 use (Del Maschio et al., 2020; but see studies showing similarity in the brains of early (0–3 years) bilinguals and monolinguals, but increased cortical thickness in L2 learners). Finally, both AoO and L2 use influence brain areas related to cognitive control, but only L2 use affects areas normally activated during overall language comprehension and production (Fedeli et al., 2021). Our findings suggest that a person's ability to use suprasegmental information with acoustic realization different from the L1 is intact after puberty. Ultimately, the determining factor in successful learning is the amount of experience with the target language.

6.3. The effects of language proficiency

We measured proficiency with an adapted version of the DELE test, which assessed grammatical and vocabulary knowledge of Spanish. Our hypothesis that higher proficiency in Spanish would yield more fixations on targets was partially supported. The data revealed that proficiency interacted with group and stress: higher proficiency increased fixations on targets in HSs with paroxytones and in L2 learners with oxytones. Proficiency did not affect L2 learners' fixations on targets in paroxytones or HSs' fixation on oxytones. This makes sense because oxytones are the preferred condition for the L2 learners, whereas paroxytones are the preferred condition for the HSs. The beneficial effects of higher proficiency on L2 learners are consistent with studies with non-beginning learners forming L2 morphophonological associations (tone-suffix: Schremm et al., 2016; Gosselke Berthelsen et al., 2018; stress-suffix: Sagarra and Casillas, 2018). In contrast with our findings, Sagarra and Casillas also observed proficiency effects with oxytones. We speculate that differences in statistical analyses (GCAs vs. Bayesian) can explain the difference.

Overall, our results align with L2 and HS online studies that show positive outcomes stemming from higher language proficiency. Behavioral L2 studies revealed that higher L2 proficiency facilitated L2 prediction based on morphosyntactic associations (e.g., Lew-Williams and Fernald, 2010; Dussias et al., 2013; Sagarra et al., 2021; Henry et al., 2022; see Ito and Pickering, 2021, for a review; and see Mitsugi, 2020, for lack of proficiency effects), phonosemantic associations (Perdomo and Kaan, 2021, for bin 5), and morphophonological associations (Sagarra and Casillas, 2018). Higher L2 proficiency also benefited L2 morphosyntactic processing (see

Kirova and Camacho, 2021, for a review), as well as L2 morphological processing (Kimppa et al., 2019), L2 word activation (Berghoff et al., 2021), and L2 phonological processing (Jun and Oh, 2000; White et al., 2015; Konishi et al., 2018; Maddah and Reiterer, 2018). Neurocognitive L2 studies indicated that higher L2 proficiency facilitated L2 morphosyntactic processing (see Alemán Bañón et al., 2018, for a review) and shaped the brain (Pliatsikas et al., 2020), allowing learners to activate the same brain areas as monolinguals (Vingerhoets et al., 2003).

HS studies produced mixed findings. Behavioral HS studies examining L2 morphosyntactic prediction indicated that higher proficiency in the heritage language yielded more and faster fixations on targets (Sagarra et al., 2021). On the other hand, others showed no proficiency or AoO effects (Sagarra and Rodríguez, 2022). This difference may be attributed to Sagarra and Rodríguez's employment of a written task (Sagarra and Varela used an auditory task) and the type of grammatical structure (adjacent subject-verb number agreement, acquired early, vs. grammatical gender agreement, acquired late). Additionally, while several neurocognitive HS studies revealed beneficial proficiency effects on grammatical processing (Bice and Kroll, 2021), others did not demonstrate any proficiency effects on grammatical processing (Wartenburger et al., 2003, found that AoO, but not proficiency, was related to grammatical processing). Certain studies also merged proficiency and AoO effects in a "multilingual experience" composite score and were therefore unable to disentangle the effects of each (Hervais-Adelman et al., 2018).

6.4. The effects of language use

In our study, language use refers to the percentage of time actively using Spanish, and includes input, output, and interaction. Our findings partially supported our hypothesis that greater language use would increase fixations on targets in HSs and L2 learners. Higher language use increased fixations on oxytone targets in HSs, but language use did not influence any other condition or group. The benefits of greater language use on HSs concur with studies that demonstrated how greater use of the heritage language produced more and earlier fixations on targets in morphosyntactic predictions (Sagarra et al., 2021) and increased sensitivity to morphosyntactic violations (Caffarra et al., 2017). Our findings also align with HS studies showing: comparable sensitivity to morphosyntactic violations in HSs and bilingual native speakers raised abroad (Foote, 2011); comparable sensitivity to syntactic violations in HSs and L2 learners raised abroad (Jegerski and Sekerina, 2020); greater sensitivity to morphosyntactic violations in sequential than simultaneous bilinguals (the former use their heritage language more; Keating, 2022); slower syntactic processing in HSs than monolinguals (Sekerina and Trueswell, 2011); and greater sensitivity to morphosyntactic violations with higher literacy experience (Parshina et al., 2022).

So, why did the L2 learners not benefit from using Spanish more daily? Possible explanations are: the L2 learners had less lexical competitors because their Spanish is fragmented; the HSs had been exposed to Spanish longer than the L2 learners; the HSs had mostly acquired Spanish by actively using it, whereas the L2 learners had mostly acquired Spanish in classroom settings. Future online studies

comparing HSs varying in their degree of exposure and current use of their heritage language will shed light on this question (in progress).

6.5. The relationship between language proficiency and use

The inclusion of language proficiency and use measures within the same sample pool provided us with the unique opportunity to examine how much weight each of these variables exerts on both early and late bilinguals' processing. Our hypothesis that language use would have a stronger influence than language proficiency was supported for the HSs and rejected for the L2 learners. In effect, language use accounted for prediction in the HS group more than proficiency. With oxytones, the HSs showed maximum prediction with [+proficiency, +use], medium prediction with [−proficiency, +use], low prediction with [+proficiency, −use], and minimum prediction with [−proficiency, −use]. As previously mentioned, neither proficiency nor use affected HSs' prediction with paroxytones, their preferred condition. On the contrary, higher proficiency facilitated L2 learners' predictions with paroxytones, regardless of amount of L2 use. Neither proficiency nor use affected L2 predictions with oxytones, their preferred condition. As stated earlier, we attribute the absence of language use effects in the L2 learners to less lexical competitors and to more years of learning confined to the classroom. In a classroom context, learners normally learn *about* Spanish (grammar, vocabulary) and devote a less-than-ideal amount of time to actively *using* Spanish. Teachers typically do not cover stress-tense suffix associations in class, so learners need to learn these associations implicitly. Because it is not a matter of later AoO, "the earlier the better" approach that drives language learning curricula in many countries is not the answer unless students can interact in the target language for extensive amounts of time. Language practitioners and coordinators could incorporate curricular changes to replace "learning-about-language" time with "using-language" time. Considering studies reporting language use effects on L2 learners, adopting a language teaching methodology that focuses on communication, encouraging learners to live abroad, and administering tests that assess language use rather than proficiency could help. For example, Beatty-Martínez et al. (2020) observed that L2 learners of the same proficiency demonstrated differences in their sensitivity to code-switching rules (those code-switching more often were more sensitive to code-switching rules). This study suggests that L2 learners are able to take advantage of extensive L2 use.

Altogether, our results demonstrated that language proficiency and use are different constructs that have distinct consequences on bilingual language processing. This proposal is consistent with recent neurocognitive evidence showing differences between proficiency and use. For instance, Del Maschio et al. (2020) found that white matter microstructure increased with greater language use rather than AoO or proficiency. Similarly, other scholars reported that *later* AoO, a possible sign of greater language use, increased white matter microstructure (Nichols and Joanisse, 2016; DeLuca et al., 2019). Language use has also been associated with subcortical brain structures related to language management processes (DeLuca et al., 2019). Furthermore, proficiency and use seem to influence distinct brain areas: language use modulates

areas linked to cognitive control and general comprehension and production, whereas language proficiency affects areas related to word learning and language selection (Fedeli et al., 2021). Turning our attention to the increased importance of language use over language proficiency in HSs, the few studies investigating the effects of language proficiency and use on HSs' processing are consistent with our findings. Sagarra and Rodríguez (2022) reported no proficiency effects on HSs' (or L2 learners') morphosyntactic predictions, and Wartenburger et al. (2003) found no proficiency effects on HSs' grammatical judgments. Although Hervais-Adelman et al. (2018) equated higher proficiency with greater volume of brain areas associated with language control in HSs, their results were based on a *sui generis* variable mixing proficiency and AoO, and brain measures of gray matter "volume" involving voxel-based morphometry are difficult to interpret. Instead of looking at gray matter volume, scholars can examine cortical thickness (more experience-related) and surface area (more innate) independently. Concerning the different role of language proficiency and use on HSs' grammatical processing, Sagarra et al. (2021) observed that, while proficiency and use increased morphosyntactic predictions, language proficiency yielded faster predictions and more attention to gender suffixes (e.g., knowing that *-a* denotes feminine gender), and language use produced earlier predictions and more attention to inherent gender information in nouns lacking transparent gender suffixes (e.g., knowing that *pared* "wall" is feminine in Spanish). Finally, regarding language use affecting language processing in HSs but not L2 learners, Sagarra and Casillas (in progress) collected phonotactic frequency data with advanced HSs and L2 learners completing the same eye-tracking task with the same stimuli as the present study. They found that higher phonotactic frequency increased fixations on targets in HSs, but not in L2 learners. In light of the essential role that language use played on HS processing along with the distinct consequences of language proficiency and use on L2 and HS processing and prediction, future HS and L2 studies and placement tests should incorporate measures of language proficiency and language use.

7. Conclusion and theoretical implications

This study examined the role of AoO, language proficiency, and language use on stress-tense suffix associations involving a stressed syllable cuing a present suffix and an unstressed syllable cuing a preterit suffix in Spanish regular verbs by adult Spanish-English HSs, English-Spanish L2 learners, and Spanish monolinguals. Participants saw a paroxytone verb (*salta* "s/he jumps") and an oxytone verb (*saltó* "s/he jumped") side by side, heard a sentence containing one of the verbs, and selected the verb they had heard. In English disyllabic verbs, oxytones are more common, whereas in Spanish words, paroxytones are more typical. The two bilingual groups were uniform in their Spanish proficiency and use. Eye-tracking data indicated that all groups fixated on target verbs above chance before hearing the second syllable that contained the suffix, except the HSs in oxytones. Monolinguals fixated on targets more and earlier, but at a slower rate than HSs and L2 learners. In

turn, HSs fixated on targets more and earlier than L2 learners, except in oxytones where HSs fixated on targets less and later than L2 learners. This was due to HSs' high number of lexical competitors due to their double L1 lexica, rather than lexical frequency or AoO. Language proficiency accounted for prediction in HSs and L2 learners and interacted with language exposure: higher proficiency increased predictions of oxytones in HSs (HSs' unpreferred condition) but of paroxytones in L2 learners (L2 learners' unfavored condition). In contrast, language use only accounted for prediction in HSs: greater use increased their predictions of oxytones. We conclude that HSs' lexical access depends more on the number of lexical competitors (co-activation of two L1 lexica) and type (phonotactic) frequency than on token (lexical) frequency or AoO. Finally, language use accounted for HS predictions more than proficiency.

Our findings inform theoretical models in phonology, lexical access, language processing, language prediction, and neurocognition. First, our data align with phonology models positing that adult L2 learners can acquire suprasegmental information different from their L1 (e.g., Van Leussen and Escudero, 2015; Flege and Bohn, 2021), and lexical access models determining that prosody influences how we activate and store words in our brain (e.g., Roll, 2015). Our results are also consistent with L2 processing models arguing that higher proficiency facilitates L2 morphological activation and allows learners to move from decompositional to full-storage lexical access (e.g., Bybee, 1985; Gonnerman et al., 2007). Moreover, our analyses indicate that HS lexical access depends on co-activation cognitive demands resulted from activating a large number of lexical competitors in their two L1 lexica. Our data do not provide evidence that HS' unique processing patterns are due to reduced exposure to input (Montrul, 2008; Pires and Rothman, 2009; Polinsky, 2011) or to reduced current activation of their heritage language (Hulsen, 2000; Putnam and Sánchez, 2013). Our findings also fall in line with L1 (Kuperberg and Jaeger, 2016) and L2 (Kaan and Grüter, 2021) models claiming that prediction variability is partially caused by individual differences in "utility" and expand these models to HS populations. Utility refers to adopting a fight-or-flee approach to prediction. That is, listeners weigh the benefits (e.g., faster processing) of engaging in prediction against its cost (e.g., risking it to make incorrect predictions); if it is worth the risk, they predict; otherwise, they do not. Lastly, our conclusions are consonant with usage-based cognitive models advocating that native early and late bilingual listeners process and predict language probabilistically based on their individual language experiences, and that language proficiency and use are separate constructs that exert distinct effects on brain adaptations (DeLuca et al., 2020). To shed light on the causes of variability of bilingual language processing, future studies should include early and late bilinguals, online auditory implicit tasks, continuous (rather than categorical) measures of AoO, proficiency, use and exposure, and type and token frequency assessments. With a goal of increasing L2 learning in mind, language practitioners can provide learners with numerous opportunities to interact in the target language and L2 learners can live abroad to maximize actively using the L2. To conclude, the underpinning of bilingual language processing variability is built upon a simple yet tremendously fluid and powerful tenet: use it or lose it.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, and the following link: <https://osf.io/2p46d/>. Further inquiries can be directed to the corresponding author.

Ethics statement

This study was reviewed and approved by Institutional Review Board at Rutgers, The State University of New Jersey, New Brunswick. The participants provided their written informed consent to participate in this study.

Author contributions

NS conceived and designed the study, collected the data, wrote the Introduction, Background, Methods, Discussion, and Conclusion sections of the article. JC analyzed the data, wrote the Lexical stress, and Results sections of the article, and created the tables and figures. All authors contributed to the article and approved the submitted version.

References

- Abu-Rabia, S., and Kehat, S. (2004). The critical period for second language pronunciation: is there such a thing? *Educ. Psychol.* 24, 77–97. doi: 10.1080/0144341032000146467
- Alemán Bañón, J., Fiorentino, R., and Gabriele, A. (2018). Using event-related potentials to track morphosyntactic development in second language learners: the processing of number and gender agreement in Spanish. *PLoS One* 13:e0200791. doi: 10.1371/journal.pone.0200791
- Barr, D. (2008). Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *J. Mem. Lang.* 59, 457–474. doi: 10.1016/j.jml.2007.09.002
- Beatty-Martínez, A. L., Navarro-Torres, C. A., Dussias, P. E., Bajo, M. T., Guzzardo Tamargo, R. E., and Kroll, J. F. (2020). Interactional context mediates the consequences of bilingualism for language and cognition. *J. Exp. Psychol. Learn. Mem. Cogn.* 46, 1022–1047. doi: 10.1037/xlm0000770
- Berghoff, R., McLoughlin, J., and Bylund, E. (2021). L1 activation during L2 processing is modulated by both age of acquisition and proficiency. *J. Neurolinguistics* 58:100979. doi: 10.1016/j.jneuroling.2020.100979
- Bice, K., and Kroll, J. (2021). Grammatical processing in two languages: how individual differences in language experience and cognitive abilities shape comprehension on heritage bilinguals. *J. Neurolinguistics* 58:100963. doi: 10.1016/j.jneuroling.2020.100963
- Black, M., Joannis, M. E., and Rafat, Y. (2020). Language dominance modulates the perception of Spanish approximants among late bilinguals. *Languages* 5:7. doi: 10.3390/languages5010007
- Boersma, P., and Weenink, D. (2021). Praat: Doing phonetics by computer [computer program]. Version 6.1.44, Available at: <https://www.fon.hum.uva.nl/ptraat/> (Accessed 17 May 2021).
- Bowles, M. (2011). Measuring implicit and explicit linguistic knowledge: what can heritage language learners contribute? *Stud. Second. Lang. Acquis.* 33, 247–271. doi: 10.1017/S0272263110000756
- Bullock, B., and Lord, G. (2003). “Analogy as a learning tool in second language acquisition” in *Romance linguistics: Theory and acquisition*. eds. A. T. Pérez-Leroux and Y. Roberge, vol. 244 (John Benjamins Publishing), 281–297.
- Bürkner, P.-C. (2017). Brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *R J.* 10, 395–411. doi: 10.32614/RJ-2018-017
- Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*, vol. 9 (John Benjamins Publishing).
- Caffarra, S., Barber, H., Molinaro, N., and Carreiras, M. (2017). When the end matters: influence of gender cues during agreement computation in bilinguals. *Lang Cogn Neurosci* 32, 1069–1085. doi: 10.1080/23273798.2017.1283426
- Chomsky, N., and Halle, M. (1968). *The sound pattern of English*. Cambridge, MA: The MIT Press.
- Chrabaszcz, A., Winn, M., Lin, C. Y., and Idsardi, W. J. (2014). Acoustic cues to perception of word stress by english, mandarin, and russian speakers. *J. Speech Lang. Hear. Res.* 57, 1468–1479. doi: 10.1044/2014_JSLHR-L-13-0279
- Clopper, C. G. (2002). Frequency of stress patterns in English: a computational analysis. IULC Working Papers Online, 2, 1–9.
- Cooper, N., Cutler, A., and Wales, R. (2002). Constraints of lexical stress on lexical access in English: evidence from native and non-native listeners. *Lang. Speech* 45, 207–228. doi: 10.1177/00238309020450030101
- Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words* MIT Press.
- De Bruin, A. (2019). Not all bilinguals are the same: a call for more detailed assessments and descriptions of bilingual experiences. *Behav. Sci.* 9:33. doi: 10.3390/bs9030033
- Del Maschio, N., Sulpizio, S., Toti, M., Caprioglio, C., Del Mauro, G., Fedeli, D., et al. (2020). Second language use rather than second language knowledge relates to changes in white matter microstructure. *J. Cult. Cogn. Sci.* 4, 165–175. doi: 10.1007/s41809-019-00039-z
- DeLuca, V., Rothman, J., and Platsikas, C. (2019). Linguistic immersion and structural effects on the bilingual brain: a longitudinal study. *Biling. Lang. Cogn.* 22, 1160–1175. doi: 10.1017/S1366728918000083
- DeLuca, V., Segal, K., Mazaheri, A., and Krott, A. (2020). Understanding bilingual brain function and structure changes? U bet! A unified bilingual experience trajectory model. *J. Neurolinguistics* 56:100930. doi: 10.1016/j.jneuroling.2020.100930
- Di Pisa, G., and Marinis, T. (2022). Gender assignment and agreement in the oral production of heritage speakers of Italian living in Germany. *Lingue e Linguaggio, Rivista semestrale* 1, 99–120. doi: 10.1418/104451
- Ditchburn, R. (1980). The function of small saccades. *Vision Res.* 20, 271–272. doi: 10.1016/0042-6989(80)90112-1
- Dussias, P., Valdés Kroff, J., Guzzardo Tamargo, R., and Gerfen, C. (2013). When gender and looking go hand in hand: grammatical gender processing in L2 Spanish. *Stud. Second. Lang. Acquis.* 35, 353–387. doi: 10.1017/S0272263112000915
- Face, T. L. (2000). “The role of syllable weight in the perception of spanish stress” in *Hispanic linguistics at the turn of the millennium*. eds. H. Campos and E. Herburger (Sommerville, MA: Cascadia Proceedings Project), 1–13.
- Face, T. L. (2005). Syllable weight and the perception of spanish stress placement by second language learners. *J. Lang. Learn.* 3, 90–103.
- Face, T. L. (2006). Cognitive factors in the perception of Spanish stress placement: Implications for a model of speech perception. *Linguistics* 44, 1237–1267. doi: 10.1515/LING.2006.040
- Faretta-Stutenberg, M., and Morgan-Short, K. (2018). The interplay of individual differences and context of learning in behavioral and neurocognitive second language development. *Second. Lang. Res.* 34, 67–101. doi: 10.1177/0267658316684903

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1141174/full#supplementary-material>

- Fedeli, D., Del Maschio, N., Sulpizio, S., Rothman, J., and Abutalebi, J. (2021). The bilingual structure connectome: Dual-language experiential factors modulate distinct cerebral networks. *Brain Lang.* 220:104978. doi: 10.1016/j.bandl.2021.104978
- Fischer, B. (1992). "Saccadic reaction time: implications for reading, dyslexia, and visual cognition" in *Eye movements and visual cognition*. ed. K. Rayner, Springer Series in Neuropsychology (New York, NY: Springer)
- Flège, J. E., and Bohn, O. S. (2021). "The revised speech learning model (SLM-r)" in *Second language speech learning: Theoretical and empirical progress*. eds. R. Wayland (Cambridge, UK: Cambridge University Press), 3–83.
- Flège, J. E., and MacKay, I. R. (2004). Perceiving vowels in a second language. *Stud. Second. Lang. Acquis.* 26, 1–34. doi: 10.1017/S0272263104026117
- Foote, R. (2011). Integrated knowledge of agreement in early and late English–Spanish bilinguals. *Appl. Psycholinguist.* 32, 187–220. doi: 10.1017/S0142716410000342
- Fuchs, Z. (2021). Facilitative use of grammatical gender in heritage Spanish. *Linguist Approach Bilingual* 12, 845–871. doi: 10.1075/lab.20024.fuc
- Fuchs, Z. (2022). Eyetracking evidence for heritage speakers' access to abstract syntactic agreement features in real-time processing. *Front. Psychol.* 13:960376. doi: 10.3389/fpsyg.2022.960376
- Gelman, A., Simpson, D., and Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy* 19, 1–13. doi: 10.3390/e19100555
- Giezen, M. R., and Emmorey, K. (2016). Language co-activation and lexical selection in bimodal bilinguals: Evidenced from picture-word interference. *Biling. Lang. Cogn.* 19, 264–276. doi: 10.1017/S1366728915000097
- Gonnerman, L. M., Seidenberg, M. S., and Andersen, E. S. (2007). Graded semantic and phonological similarity effects in priming: evidence for a distributed connectionist approach to morphology. *J. Exp. Psychol. Gen.* 136, 323–345. doi: 10.1037/0096-3445.136.2.323
- Gordon, M., and Roettger, T. (2017). Acoustic correlates of word stress: a cross-linguistic survey. *Linguistics Vanguard* 3, 1–11. doi: 10.1515/lingvan-2017-0007
- Gosselke Berthelsen, S., Horne, M., Brännström, K. J., Shtyrov, Y., and Roll, M. (2018). Neural processing of morphosyntactic tonal cues in second-language learners. *J. Neurolinguistics* 45, 60–78. doi: 10.1016/j.jneuroling.2017.09.001
- Gosselke Berthelsen, S., Horne, M., Shtyrov, Y., and Roll, M. (2020). Different neural mechanisms for rapid acquisition of words with grammatical tone in learners from tonal and non-tonal backgrounds: ERP evidence. *Brain Res.* 1729:146614. doi: 10.1016/j.brainres.2019.146614
- Gosselke Berthelsen, S., Horne, M., Shtyrov, Y., and Roll, M. (2021). Phonological transfer effects in novice learners: a learner's brain detects grammar errors only if the language sounds familiar. *Biling. Lang. Cogn.* 24, 656–669. doi: 10.1017/S1366728921000134
- Guion, S., Flège, J., Liu, S., and Yeni-Komshian, G. (2000). Age of learning effects on the duration of sentences produced in a second language. *Appl. Psycholinguist.* 21, 205–228. doi: 10.1017/S014271640000203
- Hartshorne, J. K., Tenenbaum, J. B., and Pinker, S. (2018). A critical period for second language acquisition: evidence for 2/3 million English speakers. *Cognition* 177, 263–277. doi: 10.1016/j.cognition.2018.04.007
- Hatzidaki, A., Branigan, H. P., and Pickering, M. J. (2011). Co-activation of syntax in bilingual language production. *Cogn. Psychol.* 62, 123–150. doi: 10.1016/j.cogpsych.2010.10.002
- Henry, N., Jackson, C. M., and Hopp, H. (2022). Cue coalitions and additivity in predictive processing: the interaction between case and prosody in German. *Second. Lang. Res.* 38, 397–422. doi: 10.1177/0267658320963151
- Hervais-Adelman, A., Egorova, N., and Golestani, N. (2018). Beyond bilingualism: multilingual experience correlates with caudate volume. *Brain Struct. Funct.* 223, 3495–3502. doi: 10.1007/s00429-018-1695-0
- Holt, L. L., and Lotto, A. J. (2006). Cue weighting in auditory categorization: implications for first and second language acquisition. *J. Acoust. Soc. Am.* 119, 3059–3071. doi: 10.1121/1.2188377
- Huetig, F., and McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *J. Mem. Lang.* 57, 460–482. doi: 10.1016/j.jml.2007.02.001
- Hulsen, A. (2000). Language loss and language processing: Three generations of Dutch migrants in New Zealand. Unpublished Dissertation. Universiteit Nijmegen, Netherlands.
- Hur, E., López Otero, J., and Sanchez, L. (2020). Gender agreement and assignment in Spanish heritage speakers: does frequency matter? *Languages* 5:48. doi: 10.3390/languages5040048
- Ingalvson, E. M., Holt, L. L., and McClelland, J. L. (2012). Can native Japanese listeners learn to differentiate /r-l/ on the basis of F3 onset frequency? *Biling. Lang. Cogn.* 15, 255–274. doi: 10.1017/S1366728911000447
- Isabelli-García, C., and Lacorte, M. (2016). Language learners' characteristics, target language use, and linguistic development in a domestic immersion context. *Foreign Lang. Ann.* 49, 544–556. doi: 10.1111/flan.12215
- Ito, A., Dunn, M. S., and Pickering, M. (2017). Effects of language production on prediction: word vs. picture visual world study. Poster presented at Architectures and mechanisms of language processing.
- Ito, A., and Pickering, M. J. (2021). "Automaticity and prediction in non-native language comprehension" in *Prediction in second-language processing and learning*. eds. T. Grüter and E. Kaan (Philadelphia: John Benjamins B. V.).
- Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., Siebert, C., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition* 87, B47–B57. doi: 10.1016/S0010-0277(02)00198-1
- Jegerski, J., and Sekerina, I. (2020). The processing of input with differential object marking by heritage Spanish speakers. *Biling. Lang. Cogn.* 23, 274–282. doi: 10.1017/S1366728919000087
- Jun, S. A., and Oh, M. (2000). Acquisition of second language intonation. In: Sixth International Conference on Spoken Language Processing.
- Kaan, E., and Grüter, T. (2021). *Prediction in second language processing and learning* John Benjamins Publishing Company.
- Keating, G. D. (2022). The effect of age of onset of bilingualism on gender agreement processing in Spanish as a heritage language. *Lang. Learn.* 72, 1170–1208. doi: 10.1111/lang.12510
- Kelly, M. H., and Bock, J. K. (1988). Stress in time. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 389–403. doi: 10.1037/0096-1523.14.3.389
- Kim, Y. (2020). Discrepancy between heritage speakers' use of suprasegmental cues in the perception and production of Spanish lexical stress. *Biling. Lang. Cogn.* 23, 233–250. doi: 10.1017/S1366728918001220
- Kimppa, L., Shtyrov, Y., Hut, S. C., Hedlund, L., Leminen, M., and Leminen, A. (2019). Acquisition of L2 morphology by adult language learners. *Cortex* 116, 74–90. doi: 10.1016/j.cortex.2019.01.012
- Kirova, A., and Camacho, J. (2021). Failed gender agreement in L1 English L2 Spanish: syntactic or lexical problem? *Glossa* 6, 1–35. doi: 10.16995/glossa.5761
- Klein, D., Mok, K., Chen, J.-K., and Watkins, K. E. (2014). Age of language learning shapes brain structure: A cortical thickness study of bilingual and monolingual individuals. *Brain Lang.* 131, 20–24. doi: 10.1016/j.bandl.2013.04.014
- Konishi, T., Yun, J., and Kondo, M. (2018). Acoustic correlates of L2 English stress — comparison of Japanese English and Korean English. *Phonetics Speech Sci.* 10, 9–14. doi: 10.13064/KSSS.2018.10.1.009
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Adv. Methods Pract. Psychol. Sci.* 1, 270–280. doi: 10.1177/2515245918771304
- Kuperberg, G. R., and Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31, 32–59. doi: 10.1080/23273798.2015.1102299
- Lemmerth, N., and Hopp, H. (2019). Gender processing in simultaneous and successive bilingual children: cross-linguistic lexical and syntactic influences. *Lang. Acquis.* 26, 21–45. doi: 10.1080/10489223.2017.1391815
- Lew-Williams, C., and Fernald, A. (2010). Real-time processing of gender-marked articles by native and non-native Spanish speakers. *J. Mem. Lang.* 63, 447–464. doi: 10.1016/j.jml.2010.07.003
- Lloyd-Smith, A., Einfeldt, M., and Kupisch, T. (2019). Italian-German bilinguals: the effects of heritage language use on accent in early-acquired languages. *Int. J. Biling.* 24, 289–304. doi: 10.1177/1367006919826867
- Lord, G. (2007). The role of the lexicon in learning second language stress patterns. *Appl. Lang. Learn.* 17, 1–14.
- Lozano-Argüelles, S., Sagarra, N., and Casillas, J. V. (2022). Interpreting experience and working memory effects on L1 and L2 morphological prediction. *Front. Lang. Sci.* 1:1065014. doi: 10.3389/flang.2022.1065014
- Luk, G., and Pliatsikas, C. (2015). Converging diversity to unity: commentary on the neuroanatomy of bilingualism. *Lang. Cogn. Neurosci.* 31, 349–352. doi: 10.1080/23273798.2015.1119289
- Maddah, Z. G., and Reiterer, S. M. (2018). "Language transfer vs. language talent? Individual differences and aptitude in L2 phonology of Persian-speaking learners of English" in *Exploring language aptitude: Views from psychology, the language sciences, and cognitive neuroscience. English language education*. ed. S. Reiterer, vol. 16 (Cham: Springer), 363–388.
- Martohardjono, G., Phillips, I., Madsen, C. N., and Schwartz, R. (2017). "Cross-linguistic influence in bilingual processing: an ERP study." In Proceedings of the 41st Boston University Conference on Language Development, edited by M. LaMendola and J. Scott. Somerville, MA: Cascadia Press. (pp. 452–465).
- Mayberry, R. I., and Kluender, R. (2017). Rethinking the critical period for language: new insights into an old question from American sign language. *Biling. Lang. Cogn.* 21, 886–905. doi: 10.1017/S1366728917000724
- Mitsugi, S. (2020). Generating predictions based on semantic categories in a second language: a case of numeral classifiers in Japanese. *Int. Rev. Appl. Linguist. Lang. Teach.* 58, 323–349. doi: 10.1515/iral-2017-0118
- Montrul, S. (2008). *Incomplete acquisition in bilingualism: Re-examining the age factor* John Benjamins.

- Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge: Cambridge University Press.
- Montrul, S., Davidson, J., De la Fuente, I., and Foote, R. (2013). Early language experience facilitates the processing of agreement in Spanish heritage speakers. *Biling. Lang. Cogn.* 17, 118–138. doi: 10.1017/S1366728913000114
- Montrul, S., Davidson, J., De La Fuente, I., and Foote, R. (2014). Early language experience facilitates gender agreement processing in Spanish heritage speakers. *Bilingualism* 17, 118–138. doi: 10.1017/S1366728913000114
- Morales-Front, A. (2014). “El acento [the accent]” in *Fonología generativa contemporánea de la Lengua española [contemporary generative phonology of the Spanish language]*, eds R. A. Núñez, S. Colina and T. G. Bradley. 2nd ed (Washington, DC: Georgetown University Press), 235–265.
- Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Appl. Linguis.* 35, 463–482. doi: 10.1093/applin/amu024
- Nichols, E. S., and Joanisse, M. F. (2016). Functional activity and white matter microstructure reveal the independent effects of age of acquisition and proficiency on second-language learning. *Neuroimage* 143, 15–25. doi: 10.1016/j.neuroimage.2016.08.053
- Ortega-Llebaria, M., Gu, H., and Fan, J. (2013). English speakers’ perception of Spanish lexical stress: context-driven L2 stress perception. *J. Phon.* 41, 186–197. doi: 10.1016/j.wocn.2013.01.006
- Parshina, O., Lopukhina, A., and Sekerina, I. A. (2022). Can heritage speakers predict lexical and morphosyntactic information in reading? *Languages* 7:60. doi: 10.3390/languages7010060
- Perdomo, M., and Kaan, E. (2021). Prosodic cues in second-language speech processing: a visual world eye-tracking study. *Second. Lang. Res.* 37, 349–375. doi: 10.1177/0267658319879196
- Pereira Soares, S. M. (2022). Examining effects of early (heritage) bilingualism for later multilingual acquisition and neurocognition. Doctoral thesis. University of Konstanz.
- Perez-Cortes, S. (2020). Lexical frequency and morphological regularity as sources of heritage speaker variability in the acquisition of mood. *Second. Lang. Res.* 38, 149–171. doi: 10.1177/0267658320918620
- Pike, K. L. (1945). *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.
- Pires, A., and Rothman, J. (2009). Disentangling sources of incomplete acquisition: an explanation for competence divergence across heritage grammars. *Int J Bilingual* 13, 211–238. doi: 10.1177/1367006909339806
- Platsikas, C., DeLuca, V., and Voits, T. (2020). The many shades of bilingualism: language experiences modulate adaptations in brain structure. *Lang. Learn.* 70, 133–149. doi: 10.1111/lang.12386
- Polinsky, M. (2011). Reanalysis in adult heritage language. *Stud. Second. Lang. Acquis.* 33, 305–328. doi: 10.1017/S027226311000077X
- Putnam, M. T., and Sánchez, L. (2013). What’s so incomplete about incomplete acquisition? A prolegomenon to modeling heritage language grammars. *Linguist Approach Bilingual* 3, 478–508. doi: 10.1075/lab.3.4.04put
- Ranta, L., and Meckelborg, A. (2013). How much exposure to English do international graduate students really get? Measuring language use in a naturalistic setting. *Can Modern Lang Rev* 69, 1–33. doi: 10.3138/cmlr.987
- Rodríguez, E., and Reglero, L. (2015). Heritage and L2 processing of person and number features: evidence from Spanish subject-verb agreement. *EuroAmerican J Appl Linguist Lang* 2, 11–30. doi: 10.21283/2376905X.3.46
- Roll, M. (2015). A neurolinguistic study of south Swedish word accents: electrical brain potentials in nouns and verbs. *Nordic J Linguist* 38, 149–162. doi: 10.1017/S0332586515000189
- Roll, M., Horne, M., and Lindgren, M. (2010). Word accents and morphology—ERPs of Swedish word processing. *Brain Res.* 1330, 114–123. doi: 10.1016/j.brainres.2010.03.020
- Roll, M., Söderström, P., and Horne, M. (2013). Word-stem tones cue suffixes in the brain. *Brain Res.* 1520, 116–120. doi: 10.1016/j.brainres.2013.05.013
- Romberg, A. R., and Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 906–914. doi: 10.1002/wcs.78
- Saalfeld, A. K. (2012). Teaching L2 Spanish stress. *Foreign Lang. Ann.* 45, 283–303. doi: 10.1111/j.1944-9720.2012.01191.x
- Sagarra, N., and Casillas, J. V. (2018). The use of stress in L1 and L2 lexical access. Special issue on second language processing (Ed. Jiang). *J Sec Lang Stud* 1, 31–59. doi: 10.1075/jsls.17026.sag
- Sagarra, N., Fernández-Arroyo, L., Lozano-Argüelles, C., and Casillas, J. V. (under review). Similar but different: Unraveling the Complexities of L2 Lexical Stress: L1 Transfer, L2 Proficiency, and L2 Exposure.
- Sagarra, N., and Herschensohn, J. (2010). The role of proficiency and working memory in gender and number agreement processing in L1 and L2 Spanish. *Lingua* 120, 2022–2039. doi: 10.1016/j.lingua.2010.02.004
- Sagarra, N., and Rodríguez, N. (2022). Subject-verb number agreement in bilingual processing: age of acquisition and proficiency. *Languages* 7:15. doi: 10.3390/languages7010015
- Sagarra, N., and Casillas, J. V. (in progress). Phonotactic frequency effects on heritage speakers but not L2 learners.
- Sagarra, N., Varela, J., and Fernández-Arroyo, L. (2021). “Early age of onset and early language experience facilitate grammatical gender processing.” in *Hispanic Linguistics Symposium*. Wake Forest University, North Carolina.
- Schmidt, S. (2022). Why-interrogatives in Italian heritage speakers: at the interface between syntax and discourse. *Lingue e Linguaggio, Rivista Semestrale* 1, 121–141. doi: 10.1418/104452
- Schremm, A., Söderström, P., Horne, M., and Roll, M. (2016). Implicit acquisition of tone-suffix connections in L2 learners of Swedish. *Mental Lexicon* 11, 55–75. doi: 10.1075/ml.11.1.03sch
- Sebastián-Gallés, N. (2000). *LEXESP: Léxico informatizado del español* Edicions Universitat Barcelona.
- Sekerina, I. A., and Trueswell, J. C. (2011). Processing of contrastiveness by heritage Russian bilinguals. *Biling. Lang. Cogn.* 14, 280–300. doi: 10.1017/S1366728910000337
- Singleton, D., and Leśniewska, J. (2021). The critical period hypothesis for L2 acquisition: an unfalsifiable embarrassment? *Languages* 6:149. doi: 10.3390/languages6030149
- Söderström, P., Horne, M., Frid, J., and Roll, M. (2016). Pre-activation negativity (PrAN) in brain potentials to unfolding words. *Front. Hum. Neurosci.* 10:512. doi: 10.3389/fnhum.2016.00512
- Sóskuthy, M. (2017). Generalized additive mixed models for dynamic analysis in linguistics: a practical introduction. arXiv [Preprint] arXiv:1703.05339.
- Soto-Faraco, S., Sebastián-Gallés, N., and Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. *J. Mem. Lang.* 45, 412–432. doi: 10.1006/jmla.2000.2783
- Surraín, S., and Luk, G. (2019). Describing bilinguals: a systematic review of labels and descriptions used in the literature between 2005–2015. *Biling. Lang. Cogn.* 22, 401–415. doi: 10.1017/S1366728917000682
- Tremblay, A., Broersma, M., and Coughlin, C. E. (2018). The functional weight of a prosodic cue in the native language predicts the learning of speech segmentation in a second language. *Biling. Lang. Cogn.* 21, 640–652. doi: 10.1017/S136672891700030X
- Van Leussen, J.-W., and Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Front. Psychol.* 6:e01000. doi: 10.3389/fpsyg.2015.01000
- Verissimo, J., Heyer, V., Jacob, G., and Clahsen, H. (2018). Selective effects of age of acquisition on morphological priming: evidence for a sensitive period. *Lang. Acquis.* 25, 315–326. doi: 10.3390/languages6030149
- Vingerhoets, G., Van Borsel, J., Tesink, C., Van den Noort, M., DEHlaere, K., Seurinck, R., et al. (2003). Multilingualism: an fMRI study. *Neuroimage* 20, 2181–2196. doi: 10.1016/j.neuroimage.2003.07.029
- Wartenburger, I., Heekeren, H. R., Abutalehbi, J., Cappa, S. F., Villringer, A., and Perani, D. (2003). Early setting of grammatical processing in the bilingual brain. *Neuron* 37, 159–170. doi: 10.1016/S0896-6273(02)01150-9
- White, E. J., Titone, D., Genesee, F., and Steinhauer, K. (2015). Phonological processing in late second language learners: the effects of proficiency and task. *Bilingualism* 20, 162–183. doi: 10.1017/S1366728915000620
- Winter, B., and Wieling, M. (2016). How to analyze linguistic change using mixed models, growth curve analysis and generalized additive modeling. *J Lang Evolut* 1, 7–18. doi: 10.1093/jole/lzv003
- Yeni-Komshian, G. H., Flege, J. E., and Siu, S. (2000). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. *Biling.: Lang. Cogn.* 3, 131–149. doi: 10.1017/S1366728900000225
- Zeller, J., Bylund, E., and Lewis, A. G. (2022). The parser consults the lexicon in spite of transparent gender marking: EEG evidence from noun class agreement processing in Zulu. *Cognition* 226:105148. doi: 10.1016/j.cognition.2022.105148



OPEN ACCESS

EDITED BY

Montserrat Comesaña,
University of Minho, Portugal

REVIEWED BY

Ethan Kutlu,
The University of Iowa, United States
Ana Rita Sá-Leite,
Santiago de Compostela, Spain
Zuzanna Fuchs,
University of Southern California, United States

*CORRESPONDENCE

Jason Rothman
✉ jason.rothman@uit.no

RECEIVED 02 December 2022

ACCEPTED 03 April 2023

PUBLISHED 12 June 2023

CITATION

Luque A, Rossi E, Kubota M, Nakamura M,
Rosales C, López-Rojas C, Rodina Y and
Rothman J (2023) Morphological transparency
and markedness matter in heritage speaker
gender processing: an EEG study.
Front. Psychol. 14:1114464.
doi: 10.3389/fpsyg.2023.1114464

COPYRIGHT

© 2023 Luque, Rossi, Kubota, Nakamura,
Rosales, López-Rojas, Rodina and Rothman.
This is an open-access article distributed under
the terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Morphological transparency and markedness matter in heritage speaker gender processing: an EEG study

Alicia Luque^{1,2}, Eleonora Rossi³, Maki Kubota⁴,
Megan Nakamura³, César Rosales³, Cristina López-Rojas⁵,
Yulia Rodina⁴ and Jason Rothman^{2,4*}

¹Department of Applied Language Studies, Nebrija University, Madrid, Spain, ²Nebrija Research Center in Cognition, Nebrija University, Madrid, Spain, ³Department of Linguistics, University of Florida, Gainesville, FL, United States, ⁴Department of Language and Culture, UiT the Arctic University of Norway, Tromsø, Norway, ⁵Mind, Brain and Behavior Research Center, University of Granada, Granada, Spain

The present study investigated the qualitative nature of grammatical gender knowledge and processing in heritage speakers (HSs) of Spanish living in the United States. Forty-four adult Spanish HS bilinguals participated, completing a behavioral grammatical gender assignment task and a grammaticality judgment task (GJT) while their brain activity was recorded using electroencephalography (EEG). The EEG GJT task included grammatical and ungrammatical sentences with grammatical gender violations on inanimate nouns, where transparency of the morpho(phono)logical cue and markedness were manipulated. The results of this study revealed that grammatical gender violations elicited the typical P600 effect across all relevant conditions, indicating that the grammatical representations and processing of grammatical gender in HSs are qualitatively similar to those in Spanish-dominant native speakers. Given the experimental manipulation in this study, these findings also suggest that both morphological transparency and markedness play significant roles in how grammatical gender is processed. However, the results of this study differ from those reported in previous studies with Spanish-dominant native speakers, as the P600 effect found was accompanied by a biphasic N400 effect. This pattern of results is interpreted as further evidence that the bilingual experience of HSs modulates certain aspects of morphosyntactic processing, particularly conferring a greater reliance on morphology. Additionally, the results of this study highlight the importance of incorporating neurolinguistic online processing methods to better understand what underlies HS bilingual competence and processing outcomes.

KEYWORDS

heritage bilingualism, Spanish as a heritage language, grammatical processing, gender agreement, morphological transparency, morphological markedness, event-related potentials

1. Introduction

Heritage speaker bilinguals (HSs) are native, early bilinguals of a heritage language (HL). A language qualifies as a HL when it is spoken at home—often a minoritized language—yet is distinct from the majority language(s) spoken within the larger societal context (e.g., Rothman, 2009; Montrul, 2011; Polinsky, 2018). Like homeland native speakers, HSs acquire their HL as a first language (L1), early and naturalistically. Yet, HSs often acquire the heritage L1 in a context

of significantly reduced input and/or opportunities over the lifespan to use and/or be trained in it. Thus, it is unsurprising that a substantial amount of research has documented significant differences between HSs and homeland native speakers (Montrul, 2016; Polinsky, 2018; Polinsky and Scontras, 2020) across a wide range of grammatical domains. Among these, a widely studied domain—its acquisition and processing—is grammatical gender. Relevant studies report varied results, ranging from HS performance similar to what would be expected of homeland natives to data suggesting qualitatively different gender systems in HLs (e.g., Gathercole, 2002; Gathercole and Thomas, 2005; Polinsky, 2008; Kupisch et al., 2013; Unsworth et al., 2014; Fuchs et al., 2015; Montrul, 2016; Rodina and Westergaard, 2017; Scontras et al., 2018; Di Pisa et al., 2022). Innovations with gender in HS comprehension, production and processing are perhaps surprising considering that, at least when transparency of the grammatical gender system is high as in Spanish, both (lexical) assignment and (syntactic) agreement have been shown to be acquired early by homeland native children. In fact, a series of studies have shown that mastery of gender marking on articles and adjectives in homeland native children reach target-like levels (at around 90%) around age 4 (Pérez-Pereira, 1991; López Ornat et al., 1994; Lew-Williams and Fernald, 2007; Arias-Trejo and Alva, 2013). That said, Spanish-speaking homeland native children are sensitive to the morphophonological form of the nouns and acquire the gender of transparent nouns somewhat earlier than that of opaque ones (cf. Sadek, 1975; Montrul, 2004; Gathercole et al., 2022).

Given its early acquired status, its robust frequency and its obligatory and salient nature—e.g., in Spanish, a plurality of nouns have reliable, transparent morphological exponents, matched in agreement across all elements in the determiner phrase pre- (articles/determiners) and post-nominally (adjectives)—it is not clear why grammatical gender should be an *a priori* vulnerable domain.

With few exceptions (e.g., Fuchs, 2021, 2022; Di Pisa et al., 2022), experimental evidence for grammatical gender development in heritage languages largely comes from studies using offline behavioral methods, such as spontaneous and elicited oral production and comprehension tasks. While these results demonstrate differences in HS performances in gender agreement from homeland natives, online research methods, although scarce by comparison, question any generalization regarding the vulnerability of gender in HL grammars, i.e., beyond lexical assignment. Studies employing neuroimaging techniques such as electroencephalography (EEG) with HSs are scarce, despite compelling reasons to promote their use (cf. Bayram et al., 2021).

EEG measures the summation of post-synaptic potentials generated from groups of neurons firing at the same time. This activity, although spontaneous and naturally occurring, also changes as a result of different cognitive, perceptual, or sensory demands. This makes it an excellent tool for understanding the neuronal basis of higher-order cognitive processes, such as, but not limited to, language processing. Although there are several types of analyses one can do with EEG to study bilingual language processing and related neurocognition (see Rossi et al., 2022), the most common in psycholinguistic research is to analyze the EEG signal in the time domain as Event-Related Potentials (ERPs) in order to extract neural responses to a specific event (stimuli) by averaging the time-locked signal over multiple experimental trials (Luck and Kappenman, 2011). ERPs are thus represented as waveform components of the signal at a precise time in

response to a given stimulus. In the case of language, although not specific to linguistic processing *per se*, components like the N400 or P600 reliably emerge, corresponding to matched language stimuli that do and do not contain specific types of anomalies (e.g., grammatical error, infelicity). Online methods are, in principle, less subject to issues of metalinguistic and literacy effects that have been argued to disproportionately affect HSs' performances (Kupisch and Rothman, 2016; Polinsky, 2018). As such, examining how ERP components manifest while HSs process gender errors can offer unique insights into how their grammars are underlyingly represented and how such knowledge is deployed for processing beyond what can be understood from behavioral methods alone. Recent work using eye-tracking and self-paced reading for grammatical gender already suggests that HS processing is much less distinct from homeland natives than one might have expected from previous behavioral studies (e.g., Fuchs, 2021, 2022; Di Pisa et al., 2022). Thus, EEG promises to at least complement, if not go beyond, such evidence, allowing a look into how HSs' brains process gender in real-time. With this in mind, the present study aims to fill several gaps simultaneously.

It is important to note that very few previous studies have used EEG to investigate HS linguistic processing (e.g., Van Rijswijk, 2016; Martohardjono et al., 2017). Given this, in the present study, we chose to venture into EEG with HSs within an otherwise well-studied domain of grammar in HSs, namely grammatical gender, using behavioral methods. Crucially, we do so against a backdrop of well-established use of EEG to examine gender processing for other relevant populations, namely functional monolinguals (i.e., in our terminology, homeland natives) and sequential second language (L2) bilinguals of Spanish. Using EEG with HSs, then, responsibly adds a new and crucial type of data to discussions that have emerged based on inconsistency in the HS behavioral literature of grammatical gender. Given that EEG can be an asset for adjudicating between previous ambiguous or contradictory data due to its high temporal resolution for capturing language processing in real-time, the relationship we assume between grammatical representations and real-time processing (Phillips and Ehrenhofer, 2015), and the fact that automatic brain responses are unlikely to be (less) subject to meta-/extralinguistic processes that could complicate (interpretations of) HSs' empirical performances. At the same time, data from the present study can provide a test case on the efficacy of a largely absent source of evidence for HS processing more generally (cf. Bayram et al., 2021).

1.1. Grammatical gender system in Spanish

Grammatical gender (henceforth, gender) is an inherent property of nouns. Cross-linguistically, languages differ in terms of whether they have gender, and for those that do, the specificity of their particular system sits across at least two axioms: quantity and transparency. Whereas some languages have two genders, like Spanish, others have three or more (e.g., German). Yet not all so-called simple systems are equal, for example, while Spanish and Dutch each have only two gender values (masculine/feminine and common/neuter, respectively), there are important differences between the two. For instance, Spanish features a relatively transparent gender system, characterized by highly reliable morphophonological cues that indicate gender assignment. In contrast, Dutch presents a more opaque system. Nevertheless, regardless of the system's relative

transparency, certain patterns can be observed in how gender is generally assigned to nouns. These patterns include natural gender correspondence, as well as cues based on a word's semantics or phonology. The latter becomes particularly apparent when examining inanimate nouns. However, gender assignment is generally arbitrary, with gender being reflected through syntactic agreement with other accompanying elements at the sentence level (Corbett, 1991).

Most Spanish dialects have a two-way gender system in which nouns are assigned either masculine or feminine values. Spanish nouns are marked for lexical gender using both transparent and opaque morphology. Transparent gender—where the final vowel reliably provides a cue to gender assignment—is signaled by the endings *-o* (*verano_{masc}* “summer”) or *-a* (*casa_{fem}* “house”) and is present in two-thirds of the Spanish lexicon (Harris, 1991). Indeed, approximately 99.5% of Spanish nouns ending in *-o* are masculine and around 96.3% of nouns ending in *-a* are feminine (Teschner and Russell, 1984). The remaining one-third of nouns in the Spanish lexicon do not offer strong distributional cues favoring one or the other gender assignment, except for those that offer other reliable gender cue patterns such as for the endings *-ción* and *-idad* in the case of feminine nouns. There are additional opaque gender cue patterns and tendencies which are also the focus of the current study, such as nouns ending in either a consonant (*pan_{masc}* “bread”; *amistad_{fem}* “friendship”) or the vowel *-e* (*coche_{masc}* “car”; *calle_{fem}* “street”), which can be either masculine and feminine to similar degrees.

Additionally, current trends in morphological theory posit that Spanish masculine and feminine agreement features are asymmetrically represented (cf. Battistella, 1990; Harris, 1991; Cowper, 2005). Specifically, masculine is argued to be the default and thus unmarked relative to feminine. Under some approaches (e.g., Harris, 1991), masculine is actually the absence of a gender specification whereas feminine is the specific form that carries gender features. This approach would account for the fact that masculine is generally more frequent (new lexical entries to Spanish almost invariably take masculine), less error-prone in gender assignment, and the processing of agreement errors is often less costly than for feminine ones. Empirical work supports this position. For example, a corpus study by De la Cruz Cabanillas et al. (2007) revealed that 81.84% of English loanwords in Spanish were assigned masculine gender. Antón-Méndez et al. (2002) investigated noun-adjective gender agreement relations in homeland Spanish natives, finding that agreement errors were more frequent when the head noun was feminine (i.e., marked). Alemán Bañón and Rothman (2016) used EEG to show that agreement violations on marked elements are detected more easily. These findings are consistent with the claim that marked features are more disruptive to process.

1.2. Grammatical gender acquisition/processing

Regarding the acquisition of gender, research has shown that child HSs of Spanish achieve target-like mastery of gender at an early age (Pérez-Pereira, 1991; López-Ornat, 1997; Mariscal, 2009), not differing from what would be expected of milestones in homeland natives. By contrast, some longitudinal data from HS preschoolers acquiring Spanish in the US show that gender marking on articles and adjectives does not always reach ceiling accuracy by age 4 (Anderson, 1999).

In fact, in some cases, Anderson's study showed that gender errors persist and actually increase over time due to more exposure to the majority language (in this case English, a non-gender language). Errors are mainly attributed to the overuse of masculine with feminine nouns, an error pattern also reported for adult Spanish HSs (Lipski, 1993; Montrul et al., 2008; Hurr et al., 2020), more specifically, Montrul et al. (2008) showed that feminine gender was more “vulnerable,” especially with morphologically opaque nouns, as assessed by HSs' performance in an oral picture description task.

While the picture emerging from behavioral tasks would suggest differences in HS Spanish gender systems compared to homeland natives, it is relevant to note that the degree of divergence is modulated by the modality of the experimental task, with oral tasks eliciting fewer errors than written tasks (cf. Montrul et al., 2008; Alarcón, 2011). For example, participants in Montrul et al. (2008) ($n = 69$, mean age = 22.7) produced on average 11% errors in an oral picture description task but 15 and 17% errors in a written picture interpretation and a written gender recognition task, respectively. Modality differences like this are not surprising, given that unlike homeland natives (or non-sequential L2 bilinguals for that matter), for HSs oral communication is not only by far the primary locus of language use, but in some cases, it is the only form. In sum, in the aggregate, evidence from offline behavioral studies with adult HSs of Spanish suggests that grammatical gender may be vulnerable in Heritage Spanish with gender transparency on the head noun being particularly error-prone for morphologically opaque feminine nouns.

Gender retrieval and agreement processing have, in general, been studied online rather extensively via eye movements (eye-tracking) and EEG (see Molinaro et al., 2011; Kaan et al., 2021 for review). However, besides a handful of recent studies using either self-paced reading/listening and eye-tracking, there is comparatively little available for (Spanish) HSs, and none using EEG. An eye-tracking study by Fuchs (2021, 2022) compared the use of gender predictively in the visual world paradigm in adult Spanish HSs ($n = 21$, mean age 22.3) and a group of homeland Spanish natives. The results demonstrate that HSs make use of the definite articles *el_{masc}* and *la_{fem}* to predict the gender of an upcoming noun in a manner qualitatively similar to homeland natives. Not surprisingly, some differences between the two groups still occurred. After all, the groups are in many ways not comparable, given important differences in their experiences with Spanish (see Rothman et al., 2022). Although HSs fixated on target nouns faster in gender mismatch than in match conditions, they were slower than the homeland natives in both conditions overall. Notwithstanding, the differences Fuchs reports are quantitative in nature, suggesting both groups have qualitatively similar gender representations.

A similar picture emerges from a recent processing study examining the role of morphological markedness in HL gender processing using a combination of online and offline measures, such as a self-paced reading task and a GJT, by Di Pisa et al. (2022). Although the HL in this study is Italian, the results complement Fuchs' nicely and are of particular interest given what the present study examines. The Italian HSs showed clear evidence of a qualitatively similar underlying system of grammatical gender compared to homeland Italian natives. Moreover, the results from Di Pisa et al. (2022) also indicate a considerable modulatory role of gender transparency on the head noun as well as a markedness effect pertaining to the type of agreement error: feature clash errors were

more costly than default ones. This pattern, only shown by the HS group, lead the authors to argue for a heightened dependency on overt morphology in the case of HS processing.

As mentioned, ERP research on HS gender processing simply does not exist, however, there is a substantial body of research on homeland Spanish natives and L2 learners of Spanish that, given the context of the present study, is worth briefly reviewing. Those studies have mostly focused on grammatical gender processing under conditions of agreement violations with transparent nouns (those ending in *-o* or *-a*). In their aggregate, findings from Spanish functional monolinguals convincingly show that determiner-noun agreement violations elicit a greater posterior positivity around 600 milliseconds (ms) after stimulus onset (P600), as compared to conditions without violations (Barber and Carreiras, 2005; Caffarra and Barber, 2015). The P600 effect has been argued to reflect processes of syntactic integration, reanalysis and repair (Osterhout and Mobley, 1995; O'Rourke and Van Petten, 2011), or non-syntactic late integration (Brouwer et al., 2012), as well as costs associated with structure building, checking and reprocessing (Van de Meerendonk et al., 2009). The typical P600 effect found in this domain can also (but not always) be preceded by an increased left anterior negativity (LAN) between 300 ms and 500 ms poststimulus (e.g., Barber and Carreiras, 2005), attributed to processes of automatic detection of morphosyntactic violations (Friederici, 2002), difficulties integrating mismatching information (Gunter et al., 2000), or working memory costs (Coulson et al., 1998).

Few ERP studies have compared how gender agreement violations with morphologically transparent vs. opaque nouns are processed and even fewer where morphological markedness is jointly or independently considered. For Spanish functional monolinguals (Caffarra and Barber, 2015), the LAN-P600 pattern has been observed for gender violations with both transparent and opaque nouns. Transparent nouns, however, elicited a greater LAN than opaque ones around 400 ms after the nouns. Yet, no interaction was found between the biphasic pattern and noun transparency. These results were interpreted as suggesting that functional Spanish monolinguals are sensitive to the formal gender cues on the nouns, but this distributional information does not have a strong impact on agreement computation. In other words, gender cues may be redundant in recovering gender and computing agreement dependences, at least for homeland-dominant speakers. A further comparison with two groups of Spanish-Basque early bilinguals by Caffarra et al. (2017) is of relevance, especially for the present study. This study tested Basque-dominant bilinguals and Spanish-dominant bilinguals in the Basque country, a bilingual region in Northern Spain. This study tested Basque-dominant bilinguals and Spanish-dominant bilinguals in the Basque country, a bilingual region in Northern Spain. The ERP results showed that dominant Basque bilinguals elicited only a P600 effect for gender violations on opaque nouns, whereas the Spanish-dominant bilinguals showed a pattern similar to the Spanish functional monolinguals in Caffarra and Barber (2015) i.e., a biphasic LAN-P600 effect. The authors conclude that the processing of gender violations with opaque nouns in particular is affected by potentially unstable lexical representations arising on a continuum dependent on the individual's context of bilingualism and its ensuing reduction of experience with/use of Spanish on a daily basis. This is interesting in light of the behavioral evidence from the Basque-dominant bilinguals, those with the higher tendency to show the aforementioned effects, which showed high accuracy in online grammaticality judgment and an offline gender decision task. Such a result dovetails, in our view, nicely with the argumentation of Di Pisa et al. (2022), who interpreted their reaction

time results also showing a transparency effect to indicate a greater reliance/awareness of bilinguals to overt morphological exponents. The fact that this only appears to be supported in the behavioral results, however, does not entirely offer clarity on the matter but might have something to do with differences in bilingualism contexts given that the Caffarra et al. (2017) bilinguals are not HSs and live in a context where naturalistic exposure to Spanish is omnipresent in all aspects of a bilingual society.

Among the EEG studies that have examined the role of morphological markedness for Spanish gender processing during online sentence comprehension, Alemán Bañón and Rothman (2016) investigated homeland native speakers' processing and neural sensitivity to gender agreement violations in noun-adjective concord at a distance (with an intervening CP), where half of the nouns were masculine opaque (e.g., pastel_{masc} "cake") and the other half were feminine opaque (catedral_{fem} "cathedral"). Results from their study showed that homeland Spanish natives elicited a P600 effect, suggesting that they were sensitive to agreement violations. They also suggested that morphological markedness modulates the magnitude of the effect: there was a significant difference both in the timing and amplitude of the P600 response to feature-clash violations as compared to the default ones. In other words, homeland Spanish natives detected and revised mismatching noun-adjective gender violations for feminine adjectives more quickly than for masculine ones. These results are consistent with previous studies relating the time course of the P600 with the detection of structural anomalies during sentence processing (cf. Sassenhagen and Bornkessel-Schlesewsky, 2015). Alemán Bañón et al. (2017) conducted the same experiment with Spanish L2 learners. The L2 learners, similar to the homeland Spanish natives in the Alemán Bañón and Rothman's (2016) study, were sensitive to agreement violations as revealed by a P600 effect. This is especially noteworthy considering that the opaque morphological nature of the nouns in the experiment did not provide strong (morphophonological) distributional cues to gender. Additionally, the EEG data revealed that markedness also impacted online grammatical processing—a significantly earlier P600 effect emerged for feature-clash than default gender violation errors—although the effect was quantitatively smaller than for the homeland natives. On the behavioral side and potentially relevant for the context of HL processing, the results also indicated that the L2 participants made significantly more assignment than agreement errors, suggesting that L2 bilinguals had less difficulty with the syntactic aspects of gender than the lexical ones.

1.3. Research questions and hypotheses

With the contexts provided in this background review, we pose the following questions and hypotheses:

Question 1: What are the event-related potential (ERP) signatures of grammatical gender agreement processing in Spanish as a Heritage Language?

Based on previous research, we expect to find qualitatively similar effects in HSs for grammatical gender processing as has been reported in the literature for other native speakers of Spanish. In other words, we expect to at least see evidence of a P600 effect. In line with the results by Caffarra et al. (2017), we do not expect a LAN to accompany the P600 precisely because our HSs of Spanish are English-dominant speakers who are likely to have significantly less use of /exposure to

Spanish—at least at the aggregate level—than the Basque-dominant Spanish speakers, who did not show a LAN effect.

Question 2: Do we find evidence of neurophysiological signatures related to the processing of grammatical gender being modulated by various aspects of overt morphology (i.e., transparency and markedness)?

Following from what [Di Pisa et al. \(2022\)](#) argue, if it is the case that HSs are more reliant on overt morphology—even when in Caffarra and colleagues’ words it is redundant as is the case with gender agreement in Romance languages—we would expect our HSs to be highly sensitive to both transparency and markedness. Note, however, that the two sit at various levels of complexity. This could play out differentially for HSs even if the general proposal that they are more sensitive to morphology is on the right track: transparency sits at the level of the lexical representation of individual nouns whereas markedness characterizes the gender system itself. As such, all things being equal, we would expect markedness to robustly affect HS processing across the board. We expect this to be reflected via differences in the amplitude of the ERP signatures reflecting the relative cost of processing a default error over a feature-clash one (see [Alemán Bañón and Rothman, 2016](#)), due to HSs’ enhanced morphological sensitivity, potentially bootstrapped by a more generalized HS reliance on defaults overall ([Polinsky, 2018](#)). Alternatively, while we expect potential transparency effects, as they might be modulated by other individual factors distinguishing HSs from each other (e.g., HL proficiency or use/exposure) this effect is more likely to be washed out in an aggregated analysis.

2. Methods

2.1. Participants

Given the rich EEG literature on grammatical gender processing in Spanish for homeland natives and successive L2 bilinguals, from which we have established EEG signatures for the experimental stimuli we use, and following the argumentation of [Rothman et al. \(2022\)](#) that questions the need, utility and appropriateness of monolingual comparison groups under such circumstances, our population herein is solely comprised of HSs: 44 (32=females) English-dominant HSs of Spanish. At the time of testing, all participants were enrolled as undergraduate students at the University of Florida in the US. All our HS participants reported being native speakers of Spanish and having acquired English simultaneously or sequentially in childhood as an L2. Additionally, 4 participants reported being native speakers of (heritage) Portuguese.¹ The criteria

1 During peer-review, two reviewers expressed concerns about the inclusion of these four participants due to potential gender interference effects from Portuguese. To ensure transparency and reproducibility, we clarify our rationale for retaining them in our final dataset. Our analysis of their behavioral performance showed no significant differences compared to the remaining cohort, indicating no evidence of potential gender interference effects from Portuguese. Notably, these individuals were heritage speakers of both Spanish and Portuguese, underscoring the fact that some speakers have multiple home languages that differ from the majority societal language. Thus, we believed

TABLE 1 Participant characteristics.

	M (SD) [Range]
Sex	31 females
Age (years)	20.02 (1.49) [18–24]
Number of native languages ^a	2.06 (0.26) [2–3]
Number of additional languages ^b	0.42 (0.76) [0–3]
Spanish: Age of first exposure ^c (years)	0.98 (0.95) [0–5]
English: Age of first exposure ^c (years)	3.7 (2.54) [0–10]
Spanish: Percentage of daily social language use ^d	11.86 (13.51) [0–80]
English: Percentage of daily social language use ^d	70.31 (35.08) [13.33–100]
Spanish: Self-rated listening proficiency ^e	6.65 (0.65) [4–7]
Spanish: Self-rated speaking proficiency ^e	6.11 (0.85) [4–7]
Spanish: Self-rated writing proficiency ^e	5.32 (1.12) [3–7]
Spanish: Self-rated reading proficiency ^e	5.86 (1.01) [3–7]
English: Self-rated listening proficiency ^e	6.74 (1.09) [6–7]
English: Self-rated speaking proficiency ^e	6.69 (1.10) [6–7]
English: Self-rated writing proficiency ^e	6.46 (1.20) [4–7]
English: Self-rated reading proficiency ^e	6.67 (1.10) [6–7]
LexTALE-Span ^f	59.3 (6.60) [50–74]

^aIncluding Spanish and English. Additionally, 4 participants reported being also native speakers of (heritage) Portuguese.
^bIncluding Portuguese, French, German, Mandarin Chinese, Italian, Bengali, Russian, Korean, Japanese, and American Sign Language (ASL).
^cDue to the fact that all participants indicated having been exposed to both languages before the age of 5, these answers respond to the following question: “When did you start using language Spanish/English at home or at school (whichever came first)?”
^dBased off participants’ responses to how many hours a day they spent talking to non-family members (i.e., friends, co-workers, other). We took 15 h/day to represent 100% given that we are supposed to sleep 8 to 9 h on average, thus, if a participant reported spending a total of 6 h a day speaking English or Spanish to non-family members, we considered that to represent 40% of their percentage of daily social language use (e.g., (6×100)/15 = 40%).
^eSelf-rated proficiency on 1 (‘Very Poor’) to 7 (‘Excellent’) scale.
^fLexTALE-Span = Lexical Test for Advanced Learners of Spanish. The original version of the task consists of a total of 90 items ([Izura et al., 2014](#)). However, due to a technical issue during task administration, some of our participants were only presented with 87 items. To maintain consistency in our group results analysis, we adjusted the total number of items to 87 for all participants. The score reported here represents the averaged percentage of accurate responses in the task.

to participate in the study required individuals to indicate via a pre-screening questionnaire that they (a) had been exposed to Spanish either at home or in the community before age 5, (b) to have normal or corrected-to-normal vision and hearing, be right-handed, and (c) to have no history of diagnosis of neurological or learning disorders. See [Table 1](#) for demographic details characterizing our participants, including scores for key measures we detail in the following section.

2.2. General study design

The present study is part of a larger study; in this section, we only report the details regarding the tasks specifically related to examining grammatical gender agreement processing in Spanish as a HL. The

that excluding these participants without evidence of interference would overlook the diverse reality of Spanish heritage speakers, therefore, they remained part of our final dataset and data analyses.

present study was comprised of a pre-screening and one in-lab experimental session. During the pre-screening, participants provided informed consent and completed an online questionnaire, aimed at gathering detailed language and demographic history background information using the LHQ.3 (Li et al., 2020), as well as general health and handedness. Participants meeting the pre-screening criteria were invited to the in-lab experimental session. For this session, all in-task instructions were written in Spanish. First, participants completed a lexical decision task in Spanish (LexTALE-Span; Izura et al., 2014) as an objective proficiency measure and a Spanish gender assignment task—testing each participant's assigned gender value for the full set of nouns used in the EEG experiment. For both the lexical decision and the gender assignment task, response accuracy and RTs were collected. After these behavioral measures were complete, participants were fitted with an EEG and sat for the capping procedure for approximately 15–20 mins. Lastly, participants completed the main task, a Spanish grammaticality judgment task (GJT), while EEG was recorded. Upon completion of the study, participants were then debriefed and compensated with either course credit or a \$40 gift card.

2.3. Materials and procedure

2.3.1. Gender assignment task

In order to obtain each participant's own baseline for lexical gender assignment, the gender assignment task used the same set of nouns that would be presented in the sentences in the GJT. Three versions were created and counterbalanced across participants; a participant assigned to version 1 of the gender assignment task also completed version 1 of the GJT. Thus, each participant saw a total of 180 critical nouns, 90 masculine inanimate nouns (30 transparent and 60 opaque) and 90 inanimate feminine nouns (30 transparent and 60 opaque). A total of three blocks were created comprised of 30 items each. Words in each block were automatically randomized. Participants were seated in front of a 22-inch monitor. The task was presented in E-Prime 3.0 (Psychology Software Tools, 2016, Pittsburgh, PA) and completed on a computer using a keyboard. Participants were asked to indicate the grammatical gender of each word presented in the screen by selecting the appropriate gender-marked determiner from two options e_{masc} or la_{fem} appearing on the screen. During each trial, a fixation cross was presented for 500 ms. Then, each word appeared in the middle of the screen for 500 ms. After the word was presented, a prompt indicated that a response was required. The next trial began following their response. The task took approximately 7 mins to complete.

2.3.2. Grammaticality judgment task

The EEG GJT had a 2x2x2 design with grammaticality (grammatical vs. ungrammatical agreement), gender (masculine vs. feminine) and transparency (transparent vs. opaque) as factors. Each condition consisted of 60 sentences with grammatical agreement targeted at the adjective, resulting in a total of 240 grammatical sentences. Another set of 60 sentences for each condition type was created by manipulating ungrammatical gender agreement between the target noun and its corresponding adjective across the four experimental conditions, resulting in a total of 240 ungrammatical sentences, 60 per experimental condition. Each of the four experimental conditions included grammatical and ungrammatical items for each gender, half of the critical inanimate nouns had transparent endings (masculine -o and

feminine -a) while the other half had opaque endings (-e or consonant). Even though we tried to control for frequency as closely as possible, given the attested differences in frequency between masculine vs. feminine and transparent vs. opaque, in our study there was a significant difference in log frequency (based on the SUBTLEX-ESP corpus, Cueto et al., 2012) between masculine and feminine nouns ($t = -2.33$ $p = 0.02$) as well as between transparent and opaque nouns ($t = -2.97$, $p = 0.003$), as expected. To account for this, we included frequency as a control variable in the behavioral accuracy model (as described in 3.2.1). These 480 sentences were counterbalanced across three experimental lists, such that a given learner would see a total of 40 items per condition (20 grammatical and 20 ungrammatical) for each of the four experimental conditions (i.e., masculine transparent, masculine opaque, feminine transparent, feminine opaque). Importantly, no participant saw the same sentence twice. Markedness was also manipulated within the ungrammatical agreement conditions via directionality of the overt marking on the adjective concord: (a) default errors had a feminine noun with a masculine inflected adjective and (b) feature clash errors had a masculine noun with an adjective inflected as feminine. Importantly, we made sure that grammatical gender only appeared as a morphosyntactic feature without any semantic significance. In other words, all items that were included as part of the relevant gender conditions in the present study had grammatical gender, but no semantic or natural gender (assigned based on the semantic notion of biological sex; see Table 2).

TABLE 2 Example grammaticality judgment task stimuli by condition.

Condition	Grammatical control	Ungrammatical violation
Masculine Transparent	Mateo visitó un _{masc} pueblo _{masc} <i>pequeño</i> _{masc} con sus amigas. Mateo visited a _{masc} <i>small</i> _{masc} town _{masc} with his friends.	*Mateo visitó un _{masc} pueblo _{masc} <i>pequeña</i> _{fem} con sus amigas. *Mateo visited a _{masc} <i>small</i> _{fem} town _{masc} with his friends. (<i>feature-clash error</i>)
Masculine Opaque	Carla pidió un _{masc} postre _{masc} <i>dietético</i> _{masc} después del almuerzo. Carla ordered a _{masc} low-calorie _{masc} dessert _{masc} after lunch.	*Carla pidió un _{masc} postre _{masc} <i>dietética</i> _{fem} después del almuerzo. *Carla ordered a _{masc} low-calorie _{fem} dessert _{masc} after lunch. (<i>feature-clash error</i>)
Feminine Transparent	Leonor vio una _{fem} película _{fem} <i>romántica</i> _{fem} en el cine. Leonor watched a _{fem} <i>romantic</i> _{fem} movie _{fem} in the theater.	*Leonor vio una _{fem} película _{fem} <i>romántico</i> _{masc} en el cine. *Leonor watched a _{fem} <i>romantic</i> _{masc} movie _{fem} in the theater. (<i>default error</i>)
Feminine Opaque	María dio una _{fem} clase _{fem} <i>entretenida</i> _{fem} el lunes pasado. María taught an _{fem} <i>engaging</i> _{fem} class _{fem} last Monday.	*María dio una _{fem} clase _{fem} <i>entretenido</i> _{masc} el lunes pasado. *María taught an _{fem} <i>engaging</i> _{fem} class _{fem} last Monday. (<i>default error</i>)

Italics indicate the critical word in each sentence. Violation sentences are indicated by an asterisk. As shown above, transparency is assessed via the potential differences between the transparent and opaque conditions, and markedness is assessed via the potential differences between the masculine and feminine ungrammatical agreement conditions.

Since Spanish requires the determiner to be present before the noun in all sentences with adjectival modifiers—bare nominals are impossible—we added an additional set of 60 ungrammatical sentences that contained agreement violations between the determiner and its head noun, i.e., *Mariano fotografió una_{fem} tornado_{masc} peligroso_{masc} (*Mariano photographed a_{fem} dangerous_{masc} tornado_{masc}). This was done simply to avoid the pattern that all sentences in the experiment provided the correct gender assignment cue via the pre-nominal article. Additionally, the GJT included 240 sentences containing number agreement violations that are part of a different study. Finally, an additional set of 120 filler items were included. For all sentences, length ranged from 7 to 8 words. None of the critical words were repeated, and violations never occurred in initial or final sentence positions. In sum, the GJT was comprised of a total of 900 sentences, however, in this manuscript we report only findings for trials including the gender agreement conditions described.

Like the assignment task, all experimental items were distributed across three lists using a Latin square design such that participants only viewed one sentence from each sentence frame. In total, each list contained 420 sentences (240 experimental items/180 filler items). A total of six blocks were created comprised of 70 items. Sentences in each block were automatically randomized.

Experimental sentences were presented using E-Prime 3.0 software in a rapid serial visual presentation (RSVP) paradigm. Participants read sentences in Spanish one word at a time in the center of the screen EEG was recorded and were instructed to indicate grammaticality at the end of each sentence via a button-press using an external keyboard. Each trial started with a 500 ms fixation cross followed by a 150 ms

interstimulus interval (ISI). Then, each word appeared in the middle of the screen for 300 ms followed by a 150 ms ISI for all sentence items except for the last one. The next trial began following their responses. The task took approximately 50 mins to complete (Figure 1).

2.4. EEG recording and pre-processing

Continuous EEG data were acquired using an array of 32 Ag/AgCl scalp electrodes using BrainVision Products active electrodes (Brain Products GmbH, Gilching, Germany) organized in accordance with the 10–20 system. Additionally, vertical and horizontal eye movements were measured using two sets of bipolar electrooculogram (EOG) electrodes placed above and below the left eye (vertical) and on the right and left canthi (horizontal). An online reference electrode was placed on the right mastoid and another was placed on the left mastoid for later re-referencing. Impedances were maintained at <10 k Ω . The signal was amplified using a Brain Vision actiCHamp amplifier with a 24-bit analog to digital conversion and was continuously recorded at a 1,000 Hz sampling rate without online filters. All data were pre-processed offline using Brain Vision Analyzer version 2.2 (Brain Products GmbH, Gilching, Germany). All EEG data were re-referenced to the average of both mastoids and filtered using a 0.1–30 Hz IIR Butterworth filter with a 12 dB slope. An independent components analysis (ICA) was used to identify and remove vertical and horizontal eye movements. After ICA, the data were subjected to a final inspection. All final artifact rejection was done using a semi-automatic mode followed by visual confirmation. Participant data with artifact rejection

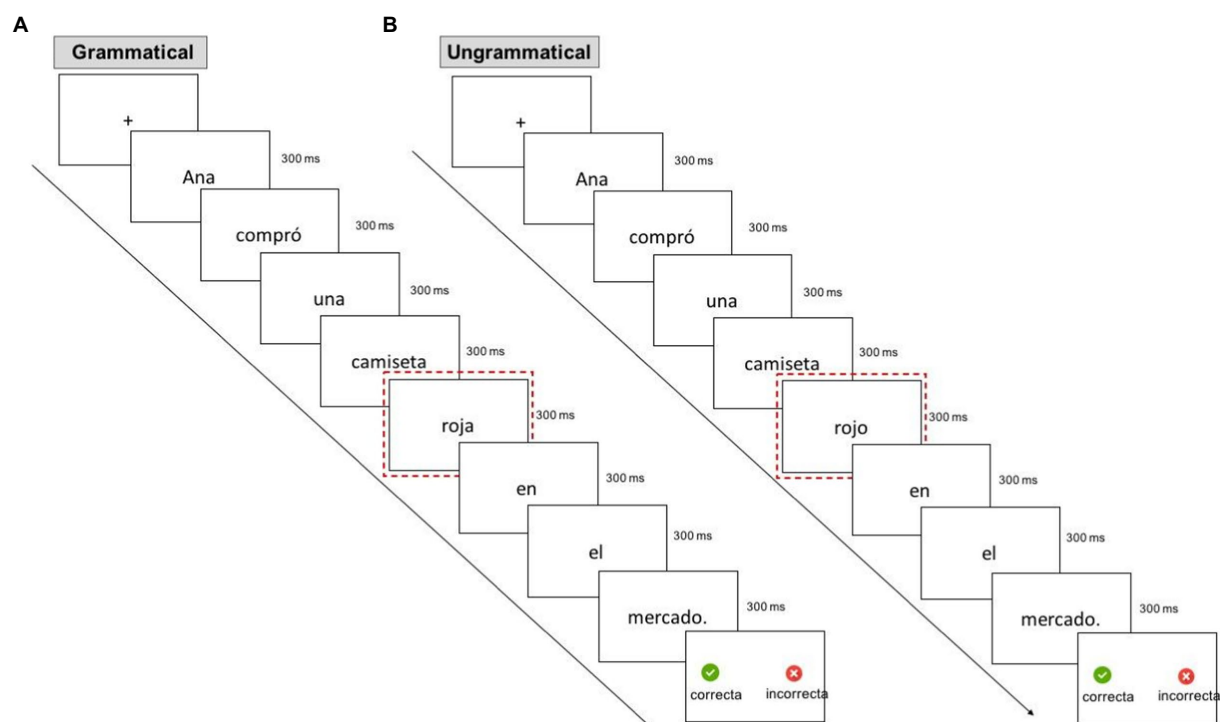


FIGURE 1

Schematic representation of example trial sequence from the grammaticality judgment task. (A) Illustrates the Grammatical condition and (B) the Ungrammatical condition. The dotted red element illustrates the target item.

rates greater than 25% were excluded from the analysis, resulting in the loss of 1 participant. Additionally, three more participants were excluded due to technical issues during EEG data acquisition. After excluding these participants, the overall mean rejection rate remained below 10%. The final analysis was conducted on correct responses only, with an average of included trials across participants of 30.21 (SD = 6.65), 27.53 (SD = 7.03), 28.97 (SD = 6.72), and 27.26 (SD = 7.20) in the masculine transparent, masculine opaque, feminine transparent, and feminine opaque conditions, respectively (out of a total of 40 trials each participant saw per experimental condition).

2.5. ERP analysis

Once the pre-processing steps were complete, epochs were extracted, and baseline corrected across all trials and across all conditions from -200 ms to 0 ms then averaged by condition. Mean amplitude ERP data were analyzed in 100 ms moving windows beginning from 0 ms prior to stimulus onset to 950 ms post-onset. A total of 10 windows were extracted. All 10 extracted time windows were included in our analysis. Analyses were conducted only for correct trials. Given the exploratory nature of this study, we did not necessarily expect that HSS would evidence the same ERP components observed in the functional monolingual literature (P600 and possibly the N400 and LAN), however, we were guided by them. Thus, we decided to focus our analyses on the full-time spectrum to be able to capture, if present, the early and later ERP components that have been consistently shown with different aspects of grammatical gender processing. All stimuli, data, and analyses scripts can be found on the following public OSF repository:²

² https://osf.io/57gac/?view_only=f08cc9da3a384e2ba1995f34980c0890

2.6. Statistical analyses

Performance data from the gender assignment task and the behavioral and EEG portions of the GJT were analyzed using generalized linear mixed effects models (Baayen et al., 2008) in R (R Core Team, 2016). Pairwise *post-hoc* comparisons with Tukey's contrasts were conducted using the *emmeans* package (Lenth, 2022). Additionally, likelihood ratio tests were conducted to analyze performance on the EEG portion of the GJT using the mixed function in the *afex* package (Singmann et al., 2022). All categorical variables were sum-coded and numerical variables were centered around the mean. The *ggplot2* package (Wickham, 2016) was used to generate Figures 2, 3, which illustrate performance (i.e., accuracy) on the gender assignment task and the behavioral portion of the EEG grammaticality judgment task. Additionally, *ggplot2* was used to create Figure 4, showcasing the time course of group-averaged brain signatures associated with the processing of the experimental conditions under investigation. Brain Vision Analyzer 2.2 was employed to generate Figure 5, which displays the topographical distribution of the ERP effects found across the different time-windows explored.

3. Results

3.1. Gender assignment task

Descriptive results show higher accuracy for masculine than feminine (Masculine: $M=0.91$, $SD=0.27$, Feminine: $M=0.8$, $SD=0.35$) and for transparent over opaque conditions (Transparent: $M=0.94$, $SD=0.23$, Opaque: $M=0.77$, $SD=0.39$), with feminine opaque being the lowest overall. Overall accuracy of the gender assignment task is presented in Figure 2. The results of the generalized linear mixed effects model (Marginal $R^2=0.19$; Conditional $R^2=0.38$) further demonstrate a significant main effect of gender ($\chi^2=74.29$ $p<0.001$), transparency

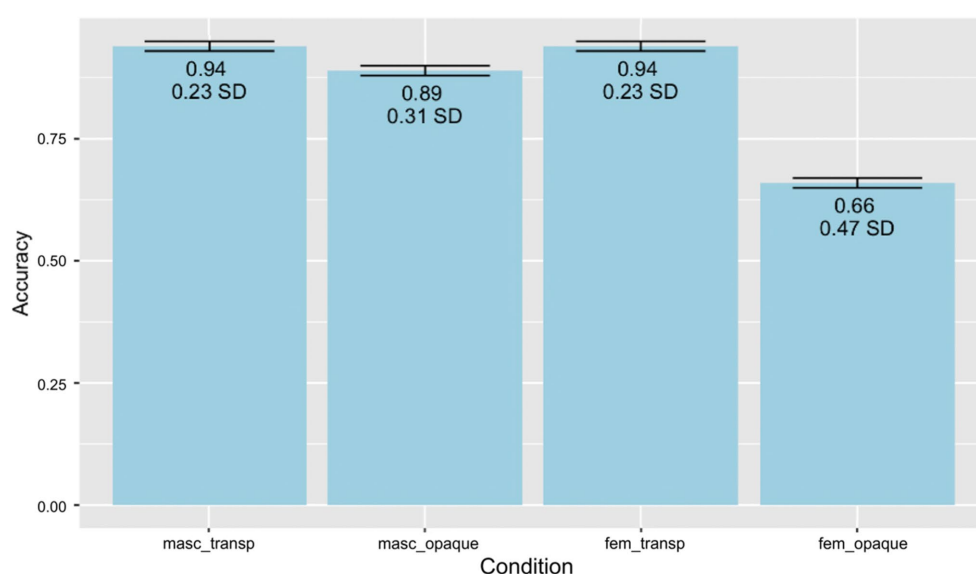


FIGURE 2

Gender assignment task: behavioral accuracy split by experimental condition. Error bars indicate standard error.

($\chi^2 = 360.05$, $p < 0.001$), as well as an interaction between gender and transparency ($\chi^2 = 73.46$, $p < 0.001$). This confirms that participants performed better on masculine than feminine ($E = -0.41$, $z = -8.79$) and on transparent than opaque ($E = -0.80$, $z = -16.92$) conditions. The only post-hoc comparison that was not significant was between feminine transparent and masculine transparent conditions ($E = -0.003$, $z = -0.01$, $p = 1.00$). In general, then, we can say that performance on the lexical gender assignment task for the nouns used in the EEG study indicates that participants performed at a rather target-like level. Not surprisingly, HSS' assignment diverges from the gender values traditionally ascribed to particular nouns and occurs when the morphology does not offer direct cues, that is, for opaque nouns. Focusing on opaque nouns, we already note what seems to be a markedness effect whereby feminine assignment is significantly degraded with respect to masculine. This pattern can either reflect a true and direct markedness effect or result from an indirect markedness effect whereby a default assignment strategy of assigning masculine is utilized.

3.2. Grammaticality judgment task

3.2.1. Behavioral results

Descriptive results demonstrate (on the aggregate level) that participants had higher accuracy on grammatical than ungrammatical items (Grammatical: $M = 0.91$, $SD = 0.28$, Ungrammatical: $M = 0.53$, $SD = 0.53$). Specifically, participants had higher accuracy for the masculine condition than the feminine one (Masculine: $M = 0.74$, $SD = 0.44$, Feminine: $M = 0.70$, $SD = 0.46$). Additionally, participants had higher accuracy for the transparent condition than the opaque one (Transparent: $M = 0.74$, $SD = 0.44$, Opaque: $M = 0.70$, $SD = 0.46$). Moreover, participants performed worst on the feminine opaque condition (see Figure 3 for overall accuracy on the behavioral portion of the GJT). The output from the generalized mixed effects model (Marginal $R^2 = 0.26$; Conditional $R^2 = 0.52$) corroborates the above-mentioned descriptive results by demonstrating a main effect of grammaticality ($\chi^2 = 1589.36$,

$p < 0.001$), gender ($\chi^2 = 6.39$, $p = 0.012$), transparency ($\chi^2 = 16.44$, $p < 0.001$), and log frequency ($\chi^2 = 11.69$, $p < 0.001$) as well as a significant interaction between grammaticality, gender, and transparency ($\chi^2 = 5.33$, $p = 0.021$). The estimate of these results confirms that participants performed (a) better on grammatical than ungrammatical items ($E = 1.32$, $z = 34.04$), (b) better on masculine than feminine items ($E = -0.11$, $z = -3.04$) as well as (c) better on transparent than opaque items ($E = -0.17$, $z = -4.76$). The significant three-way interaction indicates that the difference in accuracy between grammatical and ungrammatical conditions were both modulated by gender and transparency, following the pattern of what behavioral studies with Spanish homeland natives have also reported (e.g., Pérez-Pereira, 1991; Afonso et al., 2014). Again, taken together, our results indicate a significant role of the morphological exponents of gender at the levels of transparency as well as markedness. In other words, incorrect agreement on the adjective seems easier to judge when there is a clash between the default masculine feature of the head noun and the feminine feature of the adjective in general, and especially so when the head noun is marked with the transparent masculine ending -o.³

³ As suggested by one of our reviewers, we also ran an analysis using the d'prime score from the Signal Detection Theory (Heeger and Landy, 1997), taking into account hits, misses, false alarms, and correct rejection rates. We calculated the d'-prime values subset by Transparency (transparent vs. opaque) and Gender (feminine vs. masculine). Overall, the d'-prime analyses revealed a similar output to the model we ran for behavioral acceptance data: there was a significant effect of Transparency and Gender—with transparent having higher d'-prime score than opaque and masculine having higher d'-prime score than feminine. Similar to the acceptance model reported on this manuscript, we found no significant interaction between Transparency and Gender. Please refer to the R Markdown detailed analysis script available on our public OSF site: https://osf.io/57gac/?view_only=f08cc9da3a384e2ba1995f34980c0890.

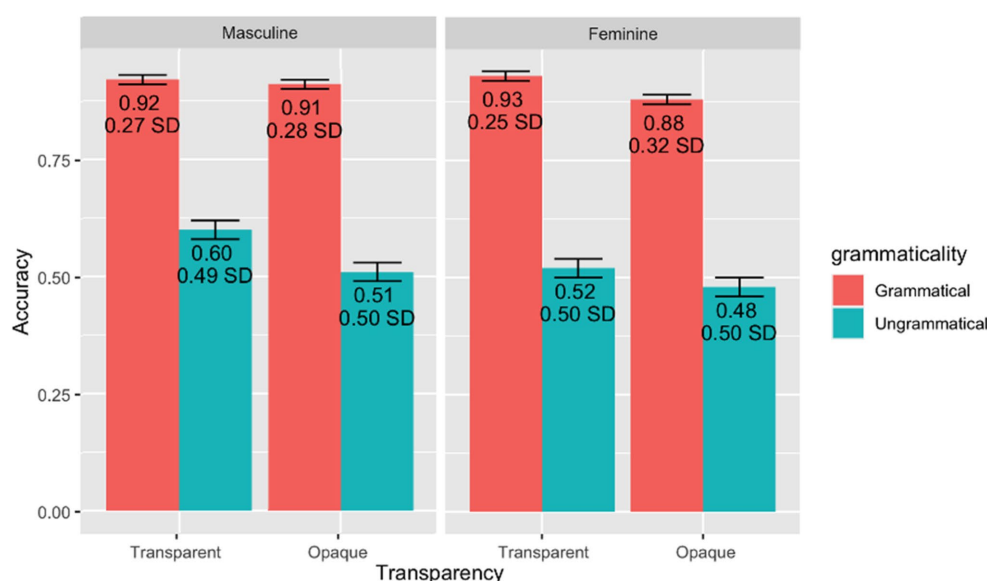


FIGURE 3

Grammaticality judgment task: behavioral results, split by gender, grammaticality, and transparency. Error bars indicate standard error.

3.3. ERP results

EEG data were analyzed in two steps as follows: first, we ran a linear mixed effects model for each moving time window and included only grammaticality (grammatical, ungrammatical) as a fixed effect and subject and electrode as random intercepts. This first step was taken to explore the main ERP components that were elicited by the design. Performing this first step-model was important to identify the ERP signatures in response to the main manipulation (grammatical, ungrammatical) observable in HSs, who might have varied considerably from the ones reported for sequential L2 bilinguals or homeland natives. Recall that this is the first EEG study with Spanish HSs for this domain and given the fact that HSs have been shown to differ significantly from these other groups with respect to performance in behavioral tasks, we did not assume *a priori* that their brain responses would overlap with what has been shown for other groups of Spanish speakers, despite being guided by the components traditionally found within relevant previous studies with homeland natives. If the first model were to indicate a main effect of grammaticality (i.e., a significant amplitude difference between grammatical and ungrammatical conditions), as it did, a second linear mixed effects model would be performed (and was) including transparency (transparent, opaque) and gender (masculine, feminine) as well as interactions as fixed effects and subject and electrode as random intercepts. This measure was taken to examine whether transparency or gender (or their interaction) modulated the effect of the specific ERP components observed in the first step.

For the first linear mixed effects model (with Bonferroni correction), results show a significant main effect of grammaticality in the 400 to 500 ms ($E=0.12$, $t=3.49$), 700 to 800 ms ($E=-0.19$, $t=-4.29$), and 800 to 900 ms windows ($E=-0.14$, $t=-3.06$). With the exception of the 400 to 500 ms window, all estimates are positive, indicating that ungrammatical conditions elicited more positive amplitudes than the grammatical ones. In contrast, in the 400 to 500 ms window the estimates are negative, indicating that ungrammatical conditions elicited more negative amplitudes than grammatical conditions (see Table 3, for summary of results from first linear mixed effects model and see Figures 4, 5 for visual representation of the effects found). In sum, we find clear evidence of a P600 effect as found in the functional monolingual processing literature (and in some of the adult L2 literature as well). This alone demonstrates

sensitivity to grammatical gender in a qualitatively similar way for the present HSs. However, unlike what has been found for homeland natives, the P600 here is accompanied by a clear N400 effect. The N400 effect has been traditionally argued to reflect lexical-semantic processing at the neural level, particularly semantic incongruity or the violation of lexical expectations, in the functionally monolingual processing literature (*cf.*, Kaan, 2007; Kutas and Federmeier, 2011).

While some studies with homeland Spanish natives show a LAN effect, the observed negativity in our data is topographically distributed over central electrodes (as seen in Figure 5) confirming its status as a genuine N400. It is not the case that the N400 has never been observed in Spanish gender processing studies. It has been noted, for example, with other sets of English-dominant bilinguals of Spanish, that is, with Spanish L2 learners (Gabriele et al., 2013), but this occurs when the L2 subjects are at low levels of proficiency and in the absence of any P600 signature. At higher levels of L2 Spanish proficiency, the reported N400 gets replaced by a P600, as shown nicely in the developmental work tracking adult L2 learners over time through the process of Spanish learning (Gabriele et al., 2013; Alemán Bañón et al., 2018). As such, the N400 at lower levels of L2 proficiency could be interpreted as a marker of development, indicating something qualitatively distinct in the processing of gender anomalies (i.e., the recognition of asymmetrical morphological patterns via matching) until reaching higher levels of proficiency where the syntax is in place. Given the high proficiency of our HSs as well as the P600 effect, we do not interpret the present N400 in the same way, a point to which we return downstream.

The results of the second linear mixed effects model (with Bonferroni correction) are provided in Table 4. In the 400 to 500 ms window, we found no significant two-way or three-way interaction, indicating that ungrammatical conditions elicit a greater negativity than grammatical conditions, regardless of transparency or gender. In the 700 to 800 ms window and 800 to 900 ms window, there was a significant two-way interaction between grammaticality and transparency as well as grammaticality and gender. Post-hoc comparisons revealed that (a) transparent conditions elicited greater positivity than opaque conditions ($p < 0.001$) and (b) masculine conditions also elicited greater positivity than feminine conditions ($p < 0.001$).

Starting, then, with the later positivity results, these effects not only indicate a qualitatively similar processing of gender as evidenced in the previous literature for homeland natives as well as advanced L2 Spanish learners, they also indicate that HSs show increased sensitivity to morphological regularity and markedness. The higher positivity noted as being manipulated by transparency suggests that HSs are particularly attuned to, if not reliant on the relatively regular patterns of Spanish gender agreement. This is not at all surprising when we consider that, despite both being Spanish natives, HSs get much less input and opportunities to meaningfully engage with the HL than homeland natives, both over the lifespan as well as in childhood when both types of native speakers would be forming the relevant grammatical representations and the processing strategies for them. It would seem then that quantity and quality of input distinctions between the two sets of natives are not merely responsible for observed differences in how the two assign gender at the lexical level for opaque nouns themselves—they cannot be reinforced by a regular morphophonological rule—but indeed how they process agreement for nouns in real-time when the overt rule cannot have a bootstrapping effect.

TABLE 3 EEG data analyses: summary of the Chi-square and the *p*-values of the main effect of grammaticality across each time-window.

Time window	Chisq	<i>p</i> -value
100 to 200 ms	0.64	0.42
200 to 300 ms	0.35	0.55
300 to 400 ms	0.75	0.38
400 to 500 ms	12.20	< 0.001**
500 to 600 ms	0.66	0.41
600 to 700 ms	0.51	0.47
700 to 800 ms	18.38	< 0.001**
800 to 900 ms	9.39	0.002*
900 to 950 ms	2.05	0.15

* $p < 0.005$. ** $p < 0.001$ (Bonferroni correction applied).

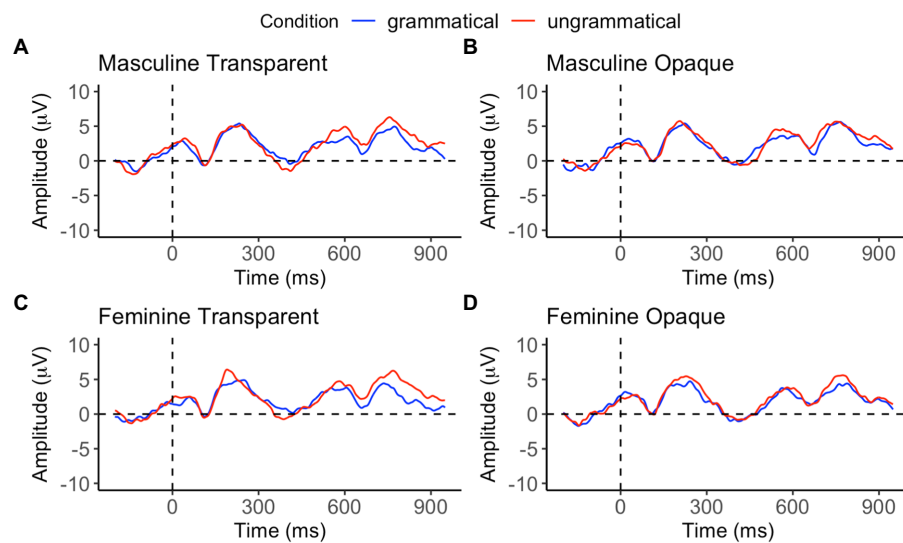


FIGURE 4
ERP waveforms across the transparent (A: masculine; C: feminine) and opaque (B: masculine; D: feminine) gender agreement conditions.

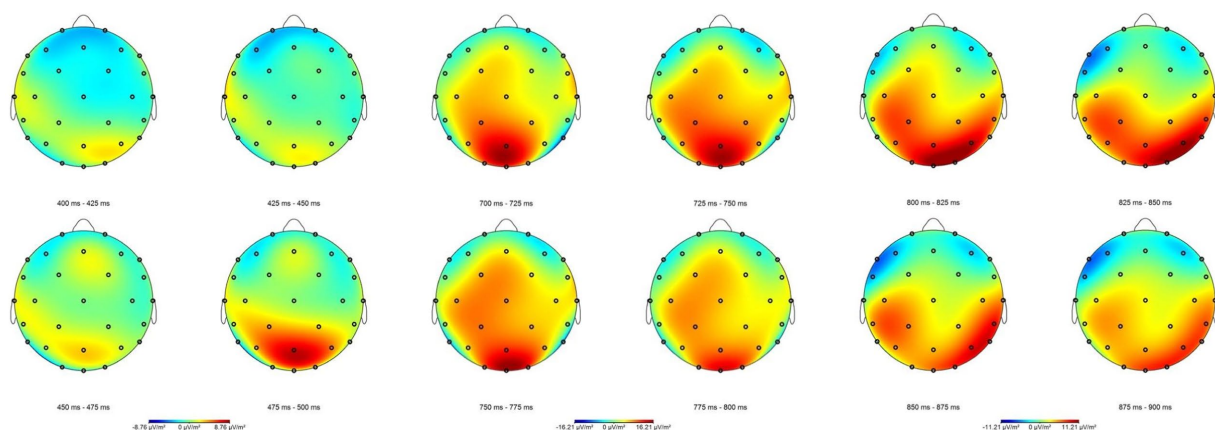


FIGURE 5
Topographic maps illustrating spatial distribution of the averaged brain responses elicited to the ungrammatical condition for the 400–500, 700–800, and 800–900ms time-windows collapsed across the transparent/opaque and masculine/feminine conditions across all participants. It should be noted that while it is standard practice to use consistent scales when plotting scalp maps, we have intentionally employed different scales for each time window represented above. This decision was made to ensure a more representative portrayal of the effect distribution across the different time windows, considering their observed differences in magnitude. We acknowledge this deviation from conventional methodology but assert that by employing individualized scaling for each time window, we aim to provide a more accurate and visually representative depiction of the distribution of the effects found across the different time windows in order to offer a more informative and insightful visualization of the data.

The fact that gender also has an effect means that markedness plays a distinct role, which again is unsurprising. Herein, this means that agreement mismatch errors reflecting a feature-clash was more costly for processing, yielding a more positive P600 effect. Recall that such an effect has also been found for homeland Spanish natives (Alemán Bañón and Rothman, 2016), offering further evidence that HSs processing of gender is qualitatively similar to other Spanish natives. However, given that this markedness effect is found also at the behavioral (in assignment and GJT) and electrophysiological levels (in Alemán Bañón and Rothman, 2016 it is only at the brain level) and is accompanied by the present transparency effect,

we would like to interpret the whole picture as supporting the interpretation offered immediately above: HSs have qualitatively similar gender representations and processing abilities but their context/reality of acquisition and use of the HL over time makes them more sensitive to overt morphological patterns for real-time processing. Such an interpretation is well in line with the argumentation offered in recent behavioral processing studies such as in Di Pisa et al. (2022), where Italian HSs showed similar significant effects for both transparency and markedness despite these same effects not being replicated in the homeland Italian and sequential L2 learner comparison groups.

TABLE 4 EEG Data Analyses: Summary of Results from the Likelihood Ratio Test run as part of the linear mixed effects model across the 400 to 500ms, 700 to 800ms, and 800 to 900ms time windows.

400 to 500ms	Chisq	p-value
Grammaticality	12.12	<0.001**
Transparency	23.24	<0.001**
Gender	11.32	<0.001**
Grammaticality:transparency	0.70	0.40
Grammaticality:gender	1.45	0.22
Transparency:gender	3.80	0.051
Grammaticality:transparency:gender	0.21	0.65
R ² Marginal = 0.006; R ² Conditional = 0.13		

700 to 800ms	Chisq	p-value
grammaticality	18.69	<0.001**
transparency	2.10	0.14
gender	10.37	0.001*
grammaticality:transparency	10.06	0.002*
grammaticality:gender	7.85	0.005*
transparency:gender	0.32	0.57
grammaticality:transparency:gender	2.18	0.14
R ² Marginal = 0.005; R ² Conditional = 0.28		

800 to 900ms	Chisq	p-value
grammaticality	9.60	0.002*
transparency	1.71	0.19
gender	6.52	0.01*
grammaticality:transparency	15.45	<0.001**
grammaticality:gender	13.58	<0.001**
transparency:gender	1.80	0.18
grammaticality:transparency:gender	0.16	0.69
R ² Marginal = 0.005; R ² Conditional = 0.19		

* $p < 0.01$, ** $p < 0.001$ (Bonferroni correction applied).

4. General discussion

The present study investigated the qualitative nature of grammatical gender processing in Spanish as an HL. More specifically, the study aimed to examine whether transparency of the gender cue on the head noun, markedness and/or an interaction between the two would modulate the observed ERP components. Having unpacked the significance of what was observed already in the *Results* section, herein we offer a more general discussion by means of returning to the two research questions offered in the *Introduction*.

Question 1: What are the ERP signatures of grammatical gender processing in Spanish as a Heritage Language?

Overall, the results of our study revealed clear evidence that (our) HSs of Spanish show a P600 effect while processing gender agreement violations. Importantly, such results are consistent with ERPs studies examining the same property in homeland natives and advanced adult

L2 learners. In other words, the present HSs, as a group, are sensitive to grammatical gender violations and process them in a qualitatively similar way to homeland natives. Thus, we interpret the present evidence as HSs having the same underlying grammar in the relevant sense, that is, a system of (morpho)syntactic grammatical gender that is equivalent to other Spanish native speakers.

However, this does not mean that the present HSs show exactly the same effects that have been reported in the homeland native speaker literature. For example, while our data show the classic P600 effect, there was no evidence that the P600 was preceded by a LAN. As discussed in the literature review, many, but crucially not all, studies with homeland natives have shown these two signatures to co-occur. And so, the absence of this co-occurrence is not terribly noteworthy or needing of too much discussion, not least as our methodology follows rather closely that of Alemán Bañón and Rothman (2016) and Alemán Bañón et al. (2017), two studies where the LAN did not accompany the P600 (see Alemán Bañón and Rothman, 2016 for why they concluded this was the case). However, there is a novelty to our data that is worthy of serious consideration, namely, the aggregate biphasic N400-P600. Indeed, this is not attested in the homeland Spanish natives' literature. While this co-occurrence, to our knowledge, is also not reported in the non-native L2 literature either, it is worth noting that in addition to studies with advanced L2 speakers often showing a P600 for gender agreement violations, studies with lower levels of L2 proficiency have shown an N400 for such violations (see Alemán Bañón et al., 2018 for review and discussion the N400 to P600 shift as a function of proficiency). And so, an N400 effect is not unattested for gender processing in the bilingual literature. Yet, in the case of L2 acquisition, not least as it seems to be indicative of lower proficiency, such an effect might signal qualitatively distinct processing related to particularly unstable representations or the lack of a qualitatively similar one. In other words, gender in lower proficiency might not yet be stabilized at the lexical level or might be absent such that the noted effect is more a reflection of the L2 learners doing something else entirely, for example, noting the breakdown of the morphophonological pattern matching.

We reject *a priori* the latter applying to our HSs for several reasons. Firstly, recall that the P600 co-occurs, suggesting that grammatical integration/reanalysis is taking place. Second, if this were applicable, we might expect this only—or at least more significantly—for transparent nouns where the final vowel should match the inflection downstream on the adjective. This is not the case, however. Conversely, if the N400-P600 biphasic effect were only found for opaque nouns, we might be inclined to interpret it as evidence for the former account related to unstable lexical representations since the N400 often occurs in the context of difficulties in lexical processing. Under such a scenario, this explanation would seem reasonable since when the morphology is opaque one is strictly reliant on the lexical representation of gender—no morphophonological rule *per se* can apply. If our HSs have unstable gender assignment representations for such nouns, they might, then, have greater difficulty that would be reflected at the lexical level and thus demonstrable via an N400 effect for such nouns only. Yet, this is also not the case, the biphasic pattern is not conditioned by the transparency of the head noun. In our view, we do not have convincing EEG-related evidence or behavioral evidence for that matter to suggest that the present HSs have unstable gender assignment representations for opaque nouns *per se*. While claiming so is a reasonable argument to make for L2 learners in the process of language acquisition, as has been done with

supporting evidence in the above cited work, one needs to be considered when applying the same logic to the case of HSs precisely because adult HSs are not in an intermediary stage of acquisition when tested. While we do have behavioral evidence showing the HSs are less accurate with feminine opaque nouns, this is not unexpected and, crucially, one need not resort to claims of unstable representations to make sense of it. To the extent that masculine is the default, we would expect what our data bear out: considerably higher accuracy for masculine opaque nouns along with degraded accuracy with feminine counterparts. It is important to make clear that low accuracy and instability are not the same thing, the former does not (necessarily) entail the latter as the source. Instability would appropriately apply if data were to show indeterminacy in gender assignment, for example, if HSs had had to provide the appropriate article for given nouns in the assignment task multiple times and showed inconsistency in doing so. If this were significantly more the case for opaque nouns in general or only for opaque feminine, appealing to instability in their system would have some empirical grounding. However, since our assignment task only had one instance for each item given the sheer size of the list of nouns, it is possible that for 34% of normatively-speaking “feminine” opaque nouns for which a masculine article was provided, HSs have different, yet stable masculine representations. If so, instability to describe this would be descriptively inaccurate. Rather, at most, it would reflect instances of misassignment, although we would be reticent to label it as such given that misassignment (accuracy for that matter) is based on differences to a consensus of non-bilingual norms. Simply put, we have no direct evidence, or at least not the right type of evidence in the present methodology, to suggest unstable gender representations. That the pattern of performance, however, follows what one would expect based on transparency/ markedness considerations and reinforces the importance of them in HS contexts, where input and domains of use are often reduced compared to other early naturalistic acquirers of the language.

We would like to consider, then, two explanations for the HS biphasic effect, not mutually exclusive to each other and both of which require further, future work to best (dis)confirm. The first thing to consider emerges on the coattails of a series of ERP studies addressing the universality and variability behind the neural correlates of morphosyntactic processing (in homeland natives and non-dominant bilinguals alike), where N400s have also been found to be elicited in response to grammatical violations for which P600s are (in theory) expected (*cf.* Pakulak and Neville, 2010; Tanner and Van Hell, 2014; Grey and van Hell, 2017; Kim et al., 2018; Tanner, 2019; Grey, 2023). Findings from these studies have revealed intrinsic and dynamic individual-level variability, both between and within-subjects, present in both L1 and L2 processing during online sentence comprehension. Results show that even when the P600 emerges (and dominates) after grand averaging, there is a need to move away from the traditional interpretation that the P600 alone indexes morphosyntactic violations. Rather, an individual-difference framework that accounts for variability in language processing routes and provides the space to examine its relationship with other learner-internal and external factors should be considered. Thus, we do not want to dismiss the possibility that, in our pool of HSs, there are enough individuals—essentially a balance between the two types—who take an N400 and a P600 route at the individual level that then in our grand averaging conserves both emerging and leaving the impression that there is a true biphasic N400-P600 group effect. If on the right track, then, it would be the case that the biphasic pattern observed is not representative of any (or very few)

HS individuals. At present, we do not have a large enough sample to meaningfully unpack this. Thus, we leave testing this possibility to a future date when we have enough participants, as in Tanner and Van Hell (2014) or Grey (2023), to see if indeed the N400-P600 pattern is truly representative of all our HSs or, if, rather at least some of this pattern is more reflective of a split in individual performance along a continuum whereby some might be more N400-dominant, while others might be more P600-dominant during online gender processing. While it would be worthwhile to pursue the N400-P600 continuum for the present and other, independent reasons in HS processing, we should acknowledge a few things. With a larger sample, this biphasic pattern might not be upheld. Since both signatures co-occur presently at the aggregate level, if it is the case that the biphasic pattern is not truly descriptive of the group's individuals, then it would need to be the case that there is a near equal amount of N400 and P600 dominant processors for both to survive the grand averaging. In this case, we would want to know if our present distribution is, then, merely happenstance or what variables might explain which (and why) individuals fall more and less into one or the other camp. In any case, with more participants the balance might tip in one direction or the other such that the aggregate no longer shows a biphasic grand averaging. Nevertheless, data such as the present underscore the utility and need for doing individual-level EEG analyses when possible.

For now, however, let us offer/consider some potential insights into what we think would underlie a true biphasic N400-P600 response, whether this truly reflects all individuals of the HS aggregate or in the case, it turns out to be only some of them along a continuum as suggested above. As discussed already, our results lead strongly to the conclusion that HSs are quite sensitive to overt morphology. The present study provides converging evidence from both brain (ERPs) and behavior (agreement judgment and assignment) in this respect. While homeland Spanish natives have also been shown to be sensitive to markedness via ERP testing, for example, the degree of sensitivity of the present HSs to both markedness and crucially transparency offline and online not only seems profound but echoes what recent studies have shown for Italian HSs (Di Pisa et al., 2022), where it has been concluded that HSs are likely more sensitive to functional morphology as a compensatory strategy for the very real quantitative differences that their reality of input exposure and opportunities for use imparts. To the extent that HSs are indeed more sensitive to morphology, then the biphasic N400-P600 pattern we observed should not be surprising. Their grammatical representations for gender are qualitatively the same as other types of Spanish native speakers, hence the P600 effect indexes errors in agreement while the N400 itself indexes their enhanced sensitivity to morphology since the locus to establish agreement is lexical in nature at the same time: the gender feature's lexicalization in the mental representation of the head noun. Such an account is not mutually exclusive, as we alluded to there being individual differences. To the extent that all HSs or only some HSs show this novel pattern—unattested in homeland natives and L2 speakers alike—the above might underlie why this is so. If it turns out that, indeed this is only true for some HSs, future research would want to pursue what exponents of particular HS experiences with their HL give rise to their (and not others') greater sensitivity to morphology in syntactic processing. We leave this, then, also as an open question for future research with larger populations done in tandem with teasing out the applicability of this pattern to the many or the few.

Question 2: Do we find evidence of neurophysiological signatures related to the processing of grammatical gender being modulated by various aspects of overt morphology (i.e., transparency and markedness)?

While we have addressed Question 2 in detail above, we summarize the main findings further. Our data, both behavioral and ERP, indicated that our HSs show increased sensitivity to both morphological transparency and markedness when processing gender agreement (violations). While HSs displayed the typical P600 signature for gender processing, indicating that their grammars have qualitatively similar and robust representations for gender, the fact that this typical signature is accompanied by a not-so-typical (in this domain) N400 as well as the fact that their brain responses are significantly conditioned by transparency and markedness effects lead us to the conclusion that morphology has particularly high weighting for this set of natives. We argued that this is likely to be the case because the typical context of HSs involves reduced input and opportunity to use the HL in both real and apparent timeframes: as children when they were stabilizing their HL grammar and over time as they develop. It should come as no surprise that such a reality would have consequences for HL grammars, especially at the level of processing where we believe innovations in our HSs' performances lie—implicitly compared to what homelands have been shown to do. The syntax of gender seems to be well established and in place, whatever input our HSs have had was enough to instantiate this into their HL grammars. Yet, in light of the reduced nature of their exposure and opportunities for engaging with Spanish over time as their dominance shifted toward the majority language, their systems have become optimized to rely more on morphological/morphophonological patterns. We interpret these results, then, in the most positive of lights: the present HS data can be understood as an embodiment of “doing more with less.”

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/57gac/?view_only=362c66d50dd5437cb0696c4116b7a097.

References

- Afonso, O., Domínguez, A., Álvarez, C. J., and Morales, D. (2014). Sublexical and lexico-syntactic factors in gender access in Spanish. *J. Psycholinguist. Res.* 43, 13–25. doi: 10.1007/s10936-012-9236-0
- Alarcón, I. V. (2011). Spanish gender agreement under complete and incomplete acquisition: early and late bilinguals' linguistic behavior within the noun phrase. *Biling. Lang. Cogn.* 14, 332–350. doi: 10.1017/s1366728910000222
- Alemán Bañón, J., Fiorentino, R., and Gabriele, A. (2018). Using event-related potentials to track morphosyntactic development in second language learners: the processing of number and gender agreement in Spanish. *PLoS One* 13:e0200791. doi: 10.1371/journal.pone.0200791
- Alemán Bañón, J., Miller, D., and Rothman, J. (2017). Morphological variability in second language learners: an examination of electrophysiological and production data. *J. Exp. Psychol. Learn. Mem. Cogn.* 43:1509. doi: 10.1037/xlm0000394
- Alemán Bañón, J., and Rothman, J. (2016). The role of morphological markedness in the processing of number and gender agreement in Spanish: an event-related potential investigation. *Language, Cognition and Neuroscience* 31, 1273–1298. doi: 10.1080/23273798.2016.1218032
- Anderson, N. J. (1999). *Exploring second language reading: Issues and strategies* (pp. 53–56). Boston, MA: Heinle & Heinle.
- Antón-Méndez, I., Nicol, J. L., and Garrett, M. F. (2002). The relation between gender and number agreement processing. *Syntax* 5, 1–25. doi: 10.1111/1467-9612.00045
- Arias-Trejo, N., and Alva, E. A. (2013). Early Spanish grammatical gender bootstrapping: learning nouns through adjectives. *Dev. Psychol.* 49, 1308–1318. doi: 10.1037/a0029778
- Baayen, R. H., Davidson, D. J., and Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* 59, 390–412. doi: 10.1016/j.jml.2007.12.005
- Barber, H., and Carreiras, M. (2005). Grammatical gender and number agreement in Spanish: an ERP comparison. *J. Cogn. Neurosci.* 17, 137–153. doi: 10.1162/0898929052880101
- Battistella, E. L. (1990). *Markedness: the evaluative superstructure of language*. Albany, NY: SUNY Press.

Ethics statement

The studies involving human participants were reviewed and approved by Internal Review Board (IRB) at the University of Florida. The patients/participants provided their written informed consent to participate in this study.

Author contributions

AL, ER, and JR designed the study. AL, CR, MN, CL-R, and YR performed the data collection. AL, MK, CR, and ER processed and analyzed the data. AL, MK, CR, MN, CL-R, YR, ER, and JR contributed to the analyses, writing of the paper, and its revisions. All authors contributed to the article and approved the submitted version.

Funding

This work was done as part of the Heritage-Bilingual Linguistic Proficiency in the Native Grammar (HeLPiNG): Charting and Explaining Differences grant, generously funded by the Tromsø Forskningsstiftelse (TFS) foundation (2019–2023).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Bayram, F., Pisa, G., Rothman, J., and Slabakova, R. (2021). "Current trends and emerging methodologies in charting heritage language grammars," in *The Cambridge handbook of heritage languages and linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 545–578.
- Brouwer, H., Fitz, H., and Hoeks, J. (2012). Getting real about semantic illusions: rethinking the functional role of the P600 in language comprehension. *Brain Res.* 1446, 127–143. doi: 10.1016/j.brainres.2012.01.055
- Caffarra, S., and Barber, H. A. (2015). Does the ending matter? The role of gender-to-ending consistency in sentence reading. *Brain Res.* 1605, 83–92. doi: 10.1016/j.brainres.2015.02.018
- Caffarra, S., Barber, H., Molinaro, N., and Carreiras, M. (2017). When the end matters: influence of gender cues during agreement computation in bilinguals. *Lang. Cogn. Neurosci.* 32, 1069–1085. doi: 10.1080/23273798.2017.1283426
- Corbett, G. G. (1991). *Gender*. Cambridge, UK: Cambridge University Press.
- Coulson, S., King, J. W., and Kutas, M. (1998). Expect the unexpected: event-related brain response to morphosyntactic violations. *Lang. Cogn. Process.* 13, 21–58. doi: 10.1080/016909698386582
- Cowper, E. (2005). The geometry of interpretable features: Infl in English and Spanish. *Language* 81, 10–46. doi: 10.1353/lan.2005.0012
- Cuetos, F., Glez-Nosti, M., Barbón, A., and Brysbaert, M. (2012). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica* 33, 133–143.
- De la Cruz Cabanillas, I., Martínez, C. T., Prados, M. D., and Redondo, E. C. (2007). English loanwords in Spanish computer language. *Engl. Specif. Purp.* 26, 52–78. doi: 10.1016/j.esp.2005.06.002
- Di Pisa, G., Kubota, M., Rothman, J., and Marinis, T. (2022). Effects of markedness in gender processing in Italian as a heritage language: a speed accuracy tradeoff. *Front. Psychol.* 13:965885. doi: 10.3389/fpsyg.2022.965885
- Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *Trends Cogn. Sci.* 6, 78–84. doi: 10.1016/s1364-6613(00)01839-8
- Fuchs, Z. (2021). Facilitative use of grammatical gender in heritage Spanish. *Linguist. Approach. Biling.* 12, 845–871. doi: 10.1075/lab.20024.fuc
- Fuchs, Z. (2022). Eyetracking evidence for heritage speakers' access to abstract syntactic agreement features in real-time processing. *Front. Psychol.* 13:960376. doi: 10.3389/fpsyg.2022.960376
- Fuchs, Z., Polinsky, M., and Scontras, G. (2015). The differential representation of number and gender in Spanish. *Linguist. Rev.* 32, 703–737. doi: 10.1515/tlr-2015-0008
- Gabriele, A., Fiorentino, R., and Bañón, J. A. (2013). Examining second language development using event-related potentials: a cross-sectional study on the processing of gender and number agreement. *Linguist. Approach. Biling.* 3, 213–232. doi: 10.1075/lab.3.2.04gab
- Gathercole, S. E. (2002). "Memory development during the childhood years," in *Handbook of memory disorders*. eds. A. D. Baddeley, M. D. Kopelman and B. A. Wilson. 2nd ed (Chichester, UK: John Wiley & Sons, Ltd), 475–500.
- Gathercole, V. C. M., Stadthagen-González, H., Parafita Couto, M. C., De Mulder, H. N., Pérez-Tattam, R. S., Bosma, E., et al. (2022). "Moveable figures and grounds: making the case for the dual nature of motion events as events of motion and change of state" in *Developing language and literacy* (Cham: Springer), 129–172.
- Gathercole, V. C. M., and Thomas, E. M. (2005). "Minority language survival: input factors influencing the acquisition of Welsh" in *Proceedings of the 4th international symposium on bilingualism* (Somerville, MA: Cascadia Press), 852–874.
- Grey, S. (2023). Variability in native and nonnative language: An ERP study of semantic and grammar processing. *Stud. Second Lang. Acquis.* 45, 137–166. doi: 10.1017/S0272263122000055
- Grey, S., and van Hell, J. G. (2017). Foreign-accented speaker identity affects neural correlates of language comprehension. *J. Neurolinguistics* 42, 93–108. doi: 10.1016/j.jneuroling.2016.12.001
- Gunter, T. C., Friederici, A. D., and Schriefers, H. (2000). Syntactic gender and semantic expectancy: ERPs reveal early autonomy and late interaction. *J. Cogn. Neurosci.* 12, 556–568. doi: 10.1162/0899892900562336
- Harris, J. W. (1991). The exponence of gender in Spanish. *Linguist. Inquiry* 22, 27–62.
- Heeger, D., and Landy, M. (1997). *Signal detection theory*. Stanford, CA: Dept. Psych., Stanford Univ.
- Hurr, E., Lopez Otero, J. C., and Sanchez, L. (2020). Gender agreement and assignment in Spanish heritage speakers: does frequency matter? *Languages* 5:48. doi: 10.3390/languages5040048
- Izura, C., Cuetos, F., and Brysbaert, M. (2014). Lextale-Esp: a test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica* 35, 49–66. doi: 10.14691/CPPJ.22.4.19
- Kaan, E. (2007). Event-related potentials and language processing: a brief overview. *Lang. Linguist. Compass* 1, 571–591. doi: 10.1111/j.1749-818X.2007.00037.x
- Kaan, E., and Grüter, T. (Eds.). (2021). *Prediction in second language processing and learning*. Amsterdam, Netherlands: John Benjamins Publishing Company.
- Kim, A. E., Oines, L., and Miyake, A. (2018). Individual differences in verbal working memory underlie a tradeoff between semantic and structural processing difficulty during language comprehension: an ERP investigation. *J. Exp. Psychol. Learn. Mem. Cogn.* 44, 406–420. doi: 10.1037/xlm0000457
- Kupisch, T., Akpunar, D., and Stöhr, A. (2013). Gender assignment and gender agreement in adult bilinguals and second language learners of French. *Linguist. Approach. Biling.* 3, 150–179. doi: 10.1075/lab.3.2.02kup
- Kupisch, T., and Rothman, J. (2016). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Kutas, M., and Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event related event-related brain potential (ERP). *Annu. Rev. Psychol.* 62:621. doi: 10.1146/annurev.psych.093008.131123
- Lenth, R. V. (2022). Emmeans: estimated marginal means, aka least-squares means. R package version 1.7.5. Available at: <https://CRAN.R-project.org/package=emmeans>
- Lew-Williams, C., and Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychol. Sci.* 18, 193–198. doi: 10.1111/j.1467-9280.2007.01873.x
- Li, P., Zhang, F., Yu, A., and Zhao, X. (2020). Language history questionnaire (LHQ3): an enhanced tool for assessing multilingual experience. *Biling. Lang. Cogn.* 23, 938–944. doi: 10.1017/s1366728918001153
- Lipski, J. (1993). "Creoloid phenomena in the Spanish of transitional bilinguals" in *Spanish in the United States*. eds. A. Roca and J. M. Lipski (Berlin, Germany: De Gruyter Mouton), 155–173. doi: 10.1515/9783110804973
- López Ornat, S., Fernández, A., Gallo, P., and Mariscal, S. (1994). *La adquisición de la Lengua española [the acquisition of the Spanish language]*. Madrid, Spain: Edinumen.
- López-Ornat, S. (1997). "What lies in between a pre-grammatical and a grammatical representation? Evidence on nominal and verbal form-function mapping in Spanish from 1;7 to 2;1," in *Contemporary perspectives on the acquisition of Spanish*. eds. A. T. Pérez-Leroux and W. Glass (Somerville, MA: Cascadia Press), 3–20.
- Luck, S. J., and Kappenman, E. S. (Eds.). (2011). *The Oxford handbook of event-related potential components*. Oxford university press.
- Mariscal, S. (2009). Early acquisition of gender agreement in the Spanish noun phrase: starting small. *J. Child Lang.* 36, 143–171. doi: 10.1017/s0305000908008908
- Martohardjono, G., Phillips, I., Madsen, C. N. II, and Schwartz, R. G. (2017). "Cross-linguistic influence in bilingual processing: an ERP study," in *Proceedings of the 41st Boston University Conference on Language Development* (Vol. 2, pp. 452–465). Somerville, MA: Cascadia Press.
- Molinaro, N., Barber, H. A., and Carreiras, M. (2011). Grammatical agreement processing in reading: ERP findings and future directions. *Cortex* 47, 908–930. doi: 10.1016/j.cortex.2011.02.019
- Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: a case of morphosyntactic convergence. *Biling. Lang. Cogn.* 7, 125–142. doi: 10.1017/S1366728904001464
- Montrul, S. (2011). Introduction: The linguistic competence of heritage speakers. *Stud. Second Lang. Acquis.* 33, 155–161. doi: 10.1017/S0272263110000719
- Montrul, S. (2016). Losing your case? Dative experiencers in Mexican Spanish and heritage speakers in the United States. *Adv. Spanish Heritage Lang.* 49:126. doi: 10.1075/sib1.49.06mon
- Montrul, S., Foote, R., and Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition. *Lang. Learn.* 58, 503–553. doi: 10.1111/j.1467-9922.2008.00449.x
- O'Rourke, P. L., and Van Petten, C. (2011). Morphological agreement at a distance: dissociation between early and late components of the event-related brain potential. *Brain Res.* 1392, 62–79. doi: 10.1016/j.brainres.2011.03.071
- Osterhout, L., and Mobley, L. A. (1995). Event-related brain potentials elicited by failure to agree. *J. Mem. Lang.* 34, 739–773. doi: 10.1006/jmla.1995.1033
- Pakulak, E., and Neville, H. J. (2010). Proficiency differences in syntactic processing of monolingual native speakers indexed by event-related potentials. *J. Cogn. Neurosci.* 22, 2728–2744. doi: 10.1162/jocn.2009.21393
- Pérez-Pereira, M. (1991). The acquisition of gender: what Spanish children tell us. *J. Child Lang.* 18, 571–590. doi: 10.1017/s0305000900011259
- Phillips, C., and Ehrenhofer, L. (2015). The role of language processing in language acquisition. *Linguist Approach Biling* 5, 409–453. doi: 10.1075/lab.5.4.01phi
- Polinsky, M. (2008). Gender under incomplete acquisition: heritage speakers' knowledge of noun categorization. *Herit. Lang. J.* 6, 40–71. doi: 10.46538/hlj.6.1.3
- Polinsky, M. (2018). *Heritage languages and their speakers*. Cambridge, UK: Cambridge University Press.
- Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245
- Psychology Software Tools, Inc. [E-Prime 3.0]. (2016). Available at: <https://support.psnet.com/>

- R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Available at: <https://www.R-project.org/>
- Rodina, Y., and Westergaard, M. (2017). Grammatical gender in bilingual Norwegian–Russian acquisition: the role of input and transparency. *Biling. Lang. Cogn.* 20, 197–214. doi: 10.1017/S1366728915000668
- Rossi, E., Pereira Soares, S., Prystauka, Y., Nakamura, M., and Rothman, J. (2022). Riding the (brain) waves! Using neural oscillations to inform bilingualism research. *Biling. Lang. Cogn.* 26, 202–215. doi: 10.1017/S1366728922000451
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J., Gharibi, K., et al. (2022). Monolingual comparative normativity in bilingualism research is out of “control”: arguments and alternatives. *Appl. Psycholinguist.*, 1–14. doi: 10.1017/S0142716422000315
- Sadek, C. S. (1975). Theoretical basis for the development of the language arts curriculum in bilingual programs.
- Sassnagen, J., and Bornkessel-Schlesewsky, I. (2015). The P600 as a correlate of ventral attention network reorientation. *Cortex* 66, A3–A20. doi: 10.1016/j.cortex.2014.12.019
- Scontras, G., Polinsky, M., and Fuchs, Z. (2018). In support of representational economy: agreement in heritage Spanish. *Glossa* 3:1. doi: 10.5334/gigl.164
- Singmann, H., Bolker, B., Westfall, J., Aust, F., and Ben-Shachar, M. S. (2022). Afex: Analysis of factorial experiments. R package version 1.1–1. Available at: <https://CRAN.R-project.org/package=afex>
- Tanner, D. (2019). Robust neurocognitive individual differences in grammatical agreement processing: a latent variable approach. *Cortex* 111, 210–237. doi: 10.1016/j.cortex.2018.10.011
- Tanner, D., and Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia* 56, 289–301. doi: 10.1016/j.neuropsychologia.2014.02.002
- Teschner, R. V., and Russell, W. M. (1984). The gender patterns of Spanish nouns: an inverse dictionary-based analysis. *Hispanic Linguist.* 1, 115–132.
- Unsworth, S., Argyri, E., Cornips, L., Hulk, A., Sorace, A., and Tsimpi, I. (2014). The role of age of onset and input in early child bilingualism in Greek and Dutch. *Appl. Psycholinguist.* 35, 765–805. doi: 10.1017/s0142716412000574
- Van de Meerendonk, N., Kolk, H. H., Chwilla, D. J., and Vissers, C. T. W. (2009). Monitoring in language perception. *Lang. Linguist. Compass* 3, 1211–1224. doi: 10.1111/j.1749-818x.2009.00163.x
- Van Rijswijk, R. V. (2016). *The strength of a weaker first language: language production and comprehension by Turkish heritage speakers in the Netherlands*. Utrecht: LOT.
- Wickham, H. (2016). “Data analysis” in *ggplot2* (Cham: Springer), 189–201.



OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway, Norway

REVIEWED BY

Piotr Romanowski,
University of Warsaw, Poland
Chunxuan Shen,
The University of Queensland, Australia

*CORRESPONDENCE

Yining Wang
✉ yining.wang@hdr.mq.edu.au

RECEIVED 21 October 2022

ACCEPTED 07 June 2023

PUBLISHED 30 June 2023

CITATION

Wang Y, Williams Tetteh V and Dube S (2023)
Parental emotionality and power relations in
heritage language maintenance: experiences of
Chinese and African immigrant families in
Australia. *Front. Psychol.* 14:1076418.
doi: 10.3389/fpsyg.2023.1076418

COPYRIGHT

© 2023 Wang, Williams Tetteh and Dube. This is
an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](#). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Parental emotionality and power relations in heritage language maintenance: experiences of Chinese and African immigrant families in Australia

Yining Wang*, Vera Williams Tetteh and Sithembinkosi Dube

Department of Linguistics, Macquarie University, Sydney, NSW, Australia

Emotionality is increasingly given prominence in the field of language acquisition and socialization in migration contexts. This cross-sectional study explores the emotional experiences of Chinese and African immigrant families in their practices of maintaining their children's heritage languages. We used open-ended interviews, field notes from informal conversations and observations, photographic evidence of children's literacy practices, and language portrait (LP) descriptions, to collect data. Results from an ethnographic analysis of the data revealed a whole range of negative and positive parental emotions (e.g., anxiety, loss, shame vs. enjoyment, accomplishment, and pride), in the discourse of maintaining heritage and minority languages. We discuss the language emotions, whether positive or negative, in light of language ideologies, which specifically points to the significance of profit discourse in the formation of family language policies (FLPs). This materialistic valorization reveals the complexities of power relations between English and minority languages, between Chinese and African languages, and within various Chinese and African languages. Consequently, the distinct hierarchies between English and minority languages and the hidden layers within minority languages further legitimate diasporic ideologies of Chinese and African parents in terms of the emotionality associated with prioritizing, maintaining, and forgoing languages. These findings suggests that language emotionality is of vital importance to the psycho-social wellbeing of immigrant families and has practical implications for policymakers and heritage language research.

KEYWORDS

parental emotionality, language ideology, power relations, heritage language, Chinese migrants, African migrants

1. Introduction

Yeah into this big dream because if you ask, every parent wants their child to learn Shona or Ndebele. But to actually do it practically, it comes down to[sic] ah to our weak economy. The background that we are coming from [...] We are not just working for ourselves, we are working to earn money for ourselves to build our lives but we are also looking after a thousand people that we have left there. So our time with our kids as they grow up to actually nurture them language-wise is very very limited compared to other people. A Chinese person coming here does not have that. They can stay home with their kids, sometimes they will wait until [sic] kids go to junior school even year 5. Or they work normal shifts and just go home without having to do any of that. (Bandi)

Parents such as Lisa and Bandi and their husbands Mandla and Victor migrated from Zimbabwe to seek economic empowerment in Australia. For them, fulfilling obligations with work and/or study as key factors for their migration and living in Australia meant that they were left with limited time to spend with their children and nurture their development in their heritage language. Similar to many other minority immigrant families (e.g., Borland, 2006; Et-Bozkurt and Yagmur, 2022; Romanowski, 2022), the African parents demonstrated a strong desire for passing on their heritage languages to the next generation, but they felt anxious when perceiving that their “big dream” (as voiced by Bandi above) of language nurturing often became stuck due to economic pressures and constraints in the migration context. In fact, emotions, such as desire and anxiety, reveal that language maintenance is not merely a linguistic decision on whether or not to learn a heritage language but is deeply situated within the socio-economic and cultural backgrounds of individual families (Nyarko, 2014) or ethnicities. In explaining how the survival crisis and family burden disadvantaged them from raising their children in relation to African heritage languages (e.g., Shona or Ndebele), Bandi, as shown in the above quotation, made a comparison with Chinese immigrants whose migration and economic situation were perceived to be more advantageous to Chinese heritage language maintenance.

The umbrella term “Chinese” consists of seven major varieties or dialects: Mandarin (the northern), Yue (includes Cantonese), Wu (includes Shanghaiese), Kejia [Hakka], Min [Hokkien], Xiang, and Gan, and many of the dialects are mutually unintelligible (Taylor and Taylor, 2014). Despite the varieties of Chinese languages, only one writing system (Chinese characters) is used in China, and Mandarin is the corresponding spoken form of this written standard (Shen and Jiang, 2023). As the official language of the Chinese government and the medium of instruction in schools, Mandarin has taken precedence over all other varieties and enjoys a unique position of prestige in China (Shen and Jiang, 2023). Accompanied by the rise of China’s economic and political clout in global affairs, Mandarin has replaced Cantonese to become the new lingua franca in the broader Chinese diaspora, such as in the UK (Curd-Christiansen and Huang, 2021), Singapore (Tupas, 2015), Ireland (Liu, 2022), and Australia (Wang, 2020). Largely due to the prestigious position of Mandarin, many Chinese, including all the Chinese participants in the current research, habitually use the term “the Chinese language” as the referent of “Mandarin” Chinese. Thus, unless otherwise specified, the phrase “the Chinese language” mentioned in the excerpts often refers to Mandarin. Given China’s fast economic growth in the twenty-first century and Australian immigration policy with orientation on economic and skill criteria, these recent Chinese immigrants in Australia represent a group of middle-/upper-class Chinese who are highly skilled, highly educated, and in the high-income bracket (Gao, 2015; Colic-Peisker and Deng, 2019). Their usual migration pathway is via the skill or investment visa streams.

When it comes to African migrants, they are from a vast continent of 54 different countries. These countries are not considered to have the economic and political clout that China has. Although the African continent is home to nearly one-third of the estimated 7,000 languages in the world (Wolff, 2021), the languages of power are non-African, rather they are languages

of former colonial powers (e.g., English, French, Portuguese, and Arabic). Nevertheless, Africans are mostly bi/multilinguals, and they rely on oral tradition-based heritage languages as well as socially learned lingua franca for intergenerational communication and socialization. Multilingual repertoires are part of their everyday norms, and traditional African societies are known to have “their ways of educating their children by introducing them, playfully and through language (through tales, songs, riddles, proverbs, and language games), to culturally relevant concepts and value systems” (Wolff, 2016). In terms of language and formal education, the imposition of colonial languages and their subsequent position as formal and/or official language have led to literacy being taught in these foreign languages as the medium of instruction (Obanya, 1999; Ouane and Glanz, 2010).

The Chinese and African immigrants, similar to all other migrants to Australia, with their diverse languages and ethnic backgrounds, have brought vitality to a multilingual and multicultural Australian society. They form an important part of Australian demographic dynamics, especially in terms of identification with their home countries and Australia, language ideologies, educational needs, and orientations. In this study, we aim to deepen the knowledge about Chinese and African families’ experiences with regard to emotionality and heritage language maintenance.

2. Theoretical framework

2.1. Family language policy, heritage language outcomes, and power relations

Family language policy (FLP) refers to “deliberate and observable” as well as “default and invisible” planning in relation to language choices, uses, and practices specifically within the home domain (King et al., 2008; Curdt-Christiansen, 2009, 2018). The tripartite model of FLP (Spolsky, 2012), which comprises language ideology, language practice, and language management, largely frames existing scholarship on heritage/minority language maintenance, parent–child interactions of immigrant families, and child bilingual development (Wang, 2017; Curdt-Christiansen, 2018; Shen and Jiang, 2023). Language practice refers to what families actually do with language, i.e., what choice they make from their linguistic repertoire; and language management is conceptualized as specific efforts or strategies they make to implement their language practice (King et al., 2008; Shen and Jiang, 2023). Underlying the two components is language ideology, the driving force of language policy regarding families’ decisions and planning for the use of languages (Curd-Christiansen, 2018; Shen and Jiang, 2023). Immigrant families, regardless of their ethnic and linguistic backgrounds, conventionally relate the rationale of heritage language maintenance to ideological beliefs of language as the symbol of identities, as the tie of families, and as the vehicle for economic empowerment (Borland, 2006; Et-Bozkurt and Yagmur, 2022; Romanowski, 2022). Taking the ethnic minorities in Australia as an example, most of the second- and third-generation Turkish parents in Melbourne related the responsibility of maintaining heritage Turkish to the

survival of Turkish identity, preservation of Turkish culture, and communication with homeland relatives (Et-Bozkurt and Yagmur, 2022). Similarly, Polish–Australian fathers actively engaged with their children’s Polish learning with the hope to safeguard their Polish identity, maintaining family ties in Poland and gaining bilingual competitiveness (Romanowski, 2022). What strategies the families adopt and what actions they take largely determine whether heritage languages can be maintained or developed in the younger generation. In the case study of three Chinese children and their families in Australia, the confidence and competence of Leo’s (one subject child) heritage Chinese was associated with the high level of parental agency in language management, such as providing books in Chinese classic literature, reading and discussing the characters with the child, and watching television in Mandarin Chinese (Shen and Jiang, 2023). As a result, FLP provides the critical domain (Spolsky, 2012) or the cornerstone (Et-Bozkurt and Yagmur, 2022) of the success of intergenerational language transmission.

In fact, FLP, being a private family matter (Anthonissen and Stroud, 2022), is a socio-political reflection that gives priority to social utility, language prestige, educational empowerment, and socio-economic gains (Wang, 2017; Curdt-Christiansen, 2018; Curdt-Christiansen and La Morgia, 2018). The direction of language shift usually occurs from the minority language to the majority language or from the lower-status language to the more prestigious high-status language (Et-Bozkurt and Yagmur, 2022). Although the majority of Turkish–Australian parents believed that the Turkish language was important as it interwove Turkish identity and culture, they did not consider Turkish to be in a position to compete with English (Et-Bozkurt and Yagmur, 2022). For the best of children’s economic future, they accentuated the significance of higher skills in English, viewing it as key to good education and social mobility in Australia (Et-Bozkurt and Yagmur, 2022). Such value-laden language ideologies are often well-represented from power-inflected language planning and decisions at the family level within the broader global spaces. For instance, middle-class families in China appropriated differentiated agencies in dealing with three languages: Fangyan, Mandarin, and English (Curd-Christiansen and Wang, 2018). Parents often chose to let go of intergenerational transmission of Fangyan but placed great emphasis on their children’s Mandarin and actively invested in their study of English. For parents (and children), Fangyan, though an important vehicle of parental emotionality, was linked to locality and impracticality, Mandarin to prestigious position and symbol of Chinese identity, and English to global mobility and international integration. In African countries, a hierarchy of languages exists in the form of a three-tiered “linguistic pyramid” whereby the languages at the apex [official languages, e.g., English, French, Portuguese, and Arabic (retained languages from colonization and other forms of contact)] are endowed with higher status than languages in the middle (Lingua franca, e.g., Kiswahili) and significantly greater prestige than the base languages (over a 1,000 distinct heritage languages; Obanya, 1999; Wolff, 2021). The languages at the apex are the ones linked to prestigious positions and “used for education, business, and government affairs,” so “mastery of these languages is closely related to educational attainment and occupational/social status” (Obanya,

1999, p. 14) and to international mobility and education. The remaining languages mainly function as important vehicles of affect and parental emotionality and hold grounds for the creation, perpetuation, and maintenance of traditional artifacts, arts, and history, which end up forming the bedrock of information for most scholarly studies (Obanya, 1999). These languages also form the basis of African parental influences and child socialization practices in the Australian migration context (Ndhlovu, 2014; Mugadza et al., 2019; Akosah-Twumasi et al., 2020) where they tend to be mostly invisible in public domains. Thus, the study of family language policy should recognize the relevance and influence of visible and less visible political, social, educational, and economic forces in a given society (Curd-Christiansen, 2013).

2.2. Language emotionality, heritage language maintenance, and societal language mastery

The terms “emotionality,” “emotion,” and “feeling” can be used interchangeably, as shown in the psychology of language learning research on emotion (Sevinç and Mirvahedi, 2022). For this study, we draw on emotional descriptions in the ethnographies of Sevinç (2020) and Wang (2022). Therefore, we use language emotionality when referring to emotional nature or quality in relation to language acquisition and practices. For language emotions, we focus on specific types of feelings about languages such as happiness, excitement, or anger.

The abovementioned ideologies, either associated with family, identity, or power, are intimately involved with people’s internal emotional worlds. They are affected by and prompt different types of emotions (e.g., affection, intimacy, satisfaction, anxiety, stress, and distance), regarding heritage language maintenance or shift. Situated in Australian contexts, the Polish father’s constant engagement with their children’s daily activities, through the use of the heritage Polish, fostered intimate communication and constructed mutually positive feelings (Romanowski, 2022). Similarly, the connection between the Australian-born Chinese children and their grandparent generation relied heavily on heritage language, which served as an expression of love and a bond of affection (Shen and Jiang, 2023). Conversely, the immigrant parents living in Sydney, despite their various ethnic origins, found that the language shift not only erected a kind of fence or barrier between their past and their present but also established emotional distance between them and their children (Tannenbaum, 2005). Parents, regardless of their differences in cultural backgrounds, generally felt depressed or rejected if their children did not speak the minority language that parents addressed them in and felt ashamed at their children’s lack of heritage language proficiency (De Houwer, 2017; Sevinç and Dewaele, 2018; Sevinç, 2020). At the same time, children often felt stressed, unhappy, and even angry at being forced to learn the heritage language (Sevinç, 2020). The emotion-laden conflict in language preference and habitus can be a result of the intergenerational divergence of bilingual repertoires. The first-generation parents generally feel comfortable speaking to their children in their home or minority language as it is natural,

spontaneous, and more connected to their inner world, while their children (e.g., 1.5 or 2nd generation) tend to use majority language more habitually or skillfully (e.g., [Sevinç and Dewaele, 2018](#); [Wang, 2020](#)). Given the intimate link between emotional loading and stronger language, parents and children tend to be less well-attuned to each other's emotional world, leading to a discrepancy affecting the family dynamics ([Pavlenko, 2004](#); [Sevinç and Dewaele, 2018](#)).

The emotional upheavals suggest a potential universal that parents of the minority language, regardless of their ethnic backgrounds, wish their language to be passed on to their children. At the same time, they want their children to do well in the societal language ([De Houwer, 2017](#)). In the exploration of mothers' global satisfaction regarding their bilingual rearing, although there were feelings of awkwardness linked to the assumed failure in transmitting the minority language to their children, there also was a high level of satisfaction largely based on the perceived progress of child bilingualism as a whole ([Leist-Villis, 2004](#)). When enforcing FLP, parents felt insecure about or even torn by how to balance the wish for their children's inheritance of the minority language and the desire for their children's mastery of the societal language ([Sevinç, 2016, 2020](#); [De Houwer, 2017](#)). In immigrant and minority contexts, parental language planning and decisions, which often generated a full range of emotions, were situated in the battlefield of competing priorities of heritage and societal languages ([Sevinç, 2020](#)). In many cases, parental anxiety about their children's integration into mainstream society or their compromise on their children's language shift may cause parents to forego language maintenance goals (e.g., [Tannenbaum and Yitzhaki, 2016](#); [Wang, 2020](#)). For example, Arab transnational families living in Israel tended to send their children to Hebrew-speaking schools, even with an awareness of the potential consequences of emotional prices from the compromise on Arabic language fluency, religious beliefs, and value systems ([Tannenbaum and Yitzhaki, 2016](#)). These educational decisions were primarily based on the parental valorization of Hebrew as a good investment into a more secure education and better assimilation. Therefore, FLP, which prioritized societal Hebrew over heritage Arabic, underscores the significance of power relations in shaping language ideologies, language planning, and decisions.

Thus, the central themes that emerged from the foregoing scholarship suggest that whatever pattern language maintenance takes, decisions usually rest on a rather strong emotional basis. To the best of our knowledge, emotion research in relation to the heritage language is gaining currency, but it is heavily shaped by quantitative frameworks (e.g., [Xiao and Wong, 2014](#); [Luo, 2015](#); [Jee, 2016, 2020](#)). The few lived experiences presented in qualitatively informed language-related emotionality are usually limited to a few ethnic groups, such as the Turkish families' language anxiety about the use of minority Turkish and majority Dutch in the Netherlands ([Sevinç and Dewaele, 2018](#)) and Mongolian women's emotional relief when translanguaging in Australia ([Dovchin, 2021](#)). Our study provides an ethnographic exploration of the emotional nuances of Chinese and African families in the context of Australia.

As mentioned earlier, language emotionality is frequently prompted by language ideologies, language behaviors, and perceived outcomes (e.g., [Tannenbaum, 2005](#); [De Houwer, 2017](#)). Thus, the study examines the attitudes and practices that Chinese

and African families have held and employed. This will give context to interpreting the resulting emotions in the enforcement of FLP. Due to the significance of FLP in the emotional and linguistic stability of transnational families and their children ([Romanowski, 2022](#)), this study broadens the scope by investigating the intricacies of language-related emotionality experienced by Chinese and African families in Australia. Our study also provides a comparative ethnography of language ideologies, language practices, and parental emotions, in relation to the heritage language maintenance of these two immigrant ethnic groups. In particular, it reveals how power relations play out in the similarities and differences of language practices and emotional experiences of these Chinese and African families when supporting bilingualism in relation to their heritage languages. The research addresses the following questions:

1. What language maintenance attitudes and practices can be observed in Chinese and African families?
2. What language emotions are emergent on the part of Chinese and African parents in the process of heritage language maintenance?
3. How does the study's Chinese and African parents' emotionality interact with their language ideologies and power dynamics?

3. Methodology

Our research is a comparative study of language emotionality emergent in two ethnographies of Chinese and African migrants living in Australia. In this study, we reuse, share, and analyze data pooled from the two ethnographies. The first ethnography documented Author-1's investigation of specific emotional discourse related to FLP and maintenance experiences (see [Wang, 2022](#)). This formed an extension of her PhD project which investigated Chinese heritage language maintenance trajectories in Australia ([Wang, 2020](#)). The second ethnography was drawn from Author-2 and Author-3's Hidden Oracles project. It was an extension of Author-2's PhD thesis ([Williams Tetteh, 2015](#)), which investigated African families' language maintenance and language use in their settlement trajectories, particularly the extent to which these hitherto invisible languages are used in Australia.

Our methodological approach of sharing and reusing data follows this emergent trend within the humanities and social sciences where qualitative data are being pooled, shared, and reanalyzed to paint a broader picture in ethnographic research about language and migration (see [Piller et al., forthcoming](#)). As such, while language emotions in interactions (e.g., when happy or angry and when satisfied or disappointed) were not key foci for both research projects we draw from, language-related emotionality did, in fact, loom large in both as we found in our field notes and through various discussions about our projects. Initial conversations we had as research colleagues showed some commonalities and differences in our datasets worth pursuing as a broader and comparative study. As mentioned in the introduction, some African parents would at times make references to Chinese families when comparing their families' linguistic and migration challenges. Thus, we pooled both datasets together for reuse, and we systematically analyzed the data, which brought forth the

parents' overt and implicit emotions in relation to heritage language maintenance as migrants in Australia. These formed the basis of numerous follow-up discussions we had about the interpretation of our shared data and the research findings we present in this study.

3.1. Participants

The participants in the study are from Chinese and African families recruited through referrals from the community or research colleagues who know the families and the criteria set out in our recruitment advertisements and by word of mouth. The families we engaged with for the study are well-educated middle-class Chinese and African parents who immigrated to Australia in recent decades. There were 25 migrant parent participants in the study (see Table 1). In total, 13 parents (three fathers and 10 mothers) migrated from China. Of the remaining, 12 (four fathers, five mothers, an uncle, and an aunt) migrated from different parts of sub-Saharan Africa, namely Zimbabwe, Ghana, South Sudan, and Rwanda, and one of the fathers was Australian-born. In total, 21 parents held bachelor's degrees or above, three (one Chinese and two African) held vocational diplomas, and one African parent had up to year 10 schooling equivalent. Notably, 22 of them migrated to Australia between 2000 and 2017, and only three (one Chinese and two African) migrated in the 1990's. Before migration, all of them worked in professional roles in academia, government, NGO, finance, IT, or health. These families had 24 school-aged children in total, ranging in age from 8 to 21 years. They attended either primary school or high school at the time of the interview with the exception of two who were university students. All the names used in the research are pseudonyms. Chinese participants' pseudonyms include both the family and given names, and African counterparts' pseudonyms only include given names.

3.2. Data collection

As mentioned above, data for this study are pooled from two ethnographies of African migrants and Chinese migrants in Australia. Ethnographic data gathered for both studies include transcriptions of open-ended semi-structured interviews with parents and children, field notes from informal conversations and observations, photographic evidence of children's literacy resources

and practices, and language portrait (LP) descriptions. The LP method derives from a multimodal research tool, where both the visual and verbal modes play a role in constructing the participants' identity, language ideology, and attitudes as well as their lived language experiences and emotional states (Busch, 2012, 2016; Obojska and Purkarthofer, 2018). It goes beyond the languages used to express cognitive, emotional, and lived experiences (Busch, 2012, 2016; Wolff, 2016). For the present study, data from children and LPs were not included in the analysis.

The data for Chinese families were collected between 2017 and 2019 and for African families in 2019–2020. All interviews with Chinese parents except one (with Ge Chang) were conducted in Mandarin Chinese. Ge Chang preferred to be interviewed in English. Interviews with African parents were in English. All interviews were transcribed verbatim. Field notes from informal conversations and observations were noted down in Chinese and English, respectively. The non-English data selected for analysis were translated into English.

3.3. Data analysis

Data analysis followed previous ethnography models from previous studies (Tannenbaum, 2005; Tannenbaum and Yitzhaki, 2016; Sevinç and Backus, 2019). We have used inductive thematic analysis as the major analytical method to establish patterns of language use and participants' interpretation of their repertoires in relation to their settlement in Sydney and Australian society. The analysis in this study mainly addresses the themed areas based on the centrality of the abovementioned research questions: What feelings do parents express about languages and how do these reflect their emotive states? The transcript and field note data that conveyed parents' emotionality were initially coded into concrete themes such as oral language use, literacy language practice, children's favorable attitudes, children's resistance, proficiency outcomes, language as investment, parental happiness, and parental struggles—in NVivo. The emotional expressions were visible through the parents' use of sentimental words (e.g., regret, annoyed, upset, enjoy, proud, and amazing) or through the emotional behavior they displayed (e.g., speaking with tears in their eyes or with laughs and beaming with smiles) when they recounted their language maintenance journey. Since it was not always possible to thematize data in a clear-cut way, some data were coded with more

TABLE 1 Summary of parents' migration backgrounds and languages.

Country of origin	Migration period	Parents in study	Children in study	Languages spoken
China	2007–2014	Three fathers; 10 mothers	Seven sons; seven daughters	Mandarin, English, Cantonese, Shanghaiese, Sichuanese, Hakka, and Hokkien
Zimbabwe	2000–2017	Two fathers; two mothers	Three daughters	Shona, Ndebele, and English
South Sudan	2000–2008	Father; mother; aunt	Two sons	Arabic, Madi, Luganda, Kuku, Swahili, and English
Ghana	1990–1994	Father; mother	Daughter; son	Ewe, Ga, Akan, Pidgin English, French, Spanish, and English
Rwanda	2006	Father*; mother, uncle	Daughter	French, Swahili, Kinyarwanda, Kirundi, Auslan, and English

*Father is Australian born of Anglo origin.

than one theme. These themes were then allocated to the main categories including heritage language practices, parents' language ideologies, and parents' emotional experiences, as reflected from the titles of the following data analysis sections. In addition, data from collected evidence of FLP and maintenance results, such as photographic images provided by Chinese families, were placed into categorized files and titled "Chinese literature books," "Chinese writing samples," "certificates and awards," "school reports," and so on. The purpose of the thematization and categorization was to conceptualize immigration narratives, language use patterns, language attitudes, and negative/positive feelings and then to identify associations between heritage language issues and familial emotions of parents in a migration and minority status.

4. Findings: language maintenance and parental emotionality in the discourse of Chinese and African families

4.1. Language maintenance attitudes and practices of Chinese and African families

In the exploration of heritage language maintenance experiences of the subject families, there emerged similarities as well as noticeable differences between Chinese and African families in terms of their attitudes to and practices of language maintenance. Both Chinese and African families aspired to pass on their heritage languages to the next generation and the parents typically expressed their desires as follows:

She [Cai Xi] should speak Chinese. Or it would be so weird that a Chinese person can't speak Chinese. (Cai Wei)

I think it's always been dreams[sic] like to keep in my culture, my language. (Jeanette)

Across the data, parents, regardless of their ethnicity, clearly stated their affection for their heritage languages. Both Chinese and African families had made efforts in maintaining their children's heritage languages, primarily the oral skills, and parents talked about how they pushed their children to speak their languages in daily communication:

I always say, "no English at home." They [Ge Si and Ge Bai] are not allowed to speak English to each other. When they speak English, I say "STOP." (Ge Chang)

I tried to speak Shona to [Child name] every Friday when I don't go to work [...] Just to make sure that this child keeps speaking Shona but I can FEEL it[sic] that I'm fighting against [...] all odds. (Bandi)

As shown above, parents usually needed to fight against a child's habitual use of English when they endeavored to maintain the child's heritage language oracy. The parental struggle in language maintenance reveals how difficult it is to keep minority languages in a monolingual mindset society (Clyne, 2008; Piller, 2016), even

at a basic communication level, let alone the aspect of reading and writing.

However, in terms of heritage language literacy maintenance, there emerged a striking difference between Chinese and African families in their investment in their children's reading and writing. In this study, Chinese parents' heavy investment in their children's Chinese literacy forms a contrast with their African counterparts' more lax attitudes to the literacy development of their languages. The Chinese immigrant families widely involved their children into various literacy practices, which included reading Chinese literature, writing Chinese characters and essays, practicing calligraphy, and doing Chinese math (also see Wang, 2020). In the process of literacy involvement, these parents used Chinese textbooks, exercise books, and literature materials from China as important resources for a home tutoring or for assisting with community school assignments (see Figure 1), as referred to by Ji Ran—Ji Ming's mother:

Every time my friends went back to China, I asked them to bring us Chinese books, like the[sic] textbooks, math books, and lots of novels. My son [Ji Ming] is requested to copy one Chinese text each day and to do math exercise in[sic] school holidays. He is also encouraged to read more Chinese novels—whatever he likes. He read quite a few sets of Gongfu [功夫, Chinese martial arts] novels written by Jinyong [金庸—a well-known Gongfu novelist in Hongkong]. He also read all the[sic] Four Great Classic Novels I recommended. That's why his Chinese still improves[sic] in Australia, especially in the aspect of comprehension and general knowledge. (Ji Ran)

Ji Ming's parents' effort of obtaining various social (friend's help) and linguistic (e.g., Chinese and math textbooks and classic literature) resources is not exceptional among Chinese families. Across the data, Chinese textbooks (including math) and/or different kinds of reading materials, usually brought from China or bought online, were used to different degrees by Chinese families. For example, the families (e.g., Ge Bai's and Shi Diwen's) who used math Chinese books often emphasized the dual function of Chinese learning and math advancement, which was often described as "一举两得[kill two birds with one stone]." Most Chinese parents evidently considered their children's literacy proficiency as the crucial marker of the success or failure of their family language policy or language maintenance outcome. When Xu Li's mother, Xu Dai, admired other children's ability in reading sophisticated books, she concluded with a sense of loss that her language tutoring was unsuccessful because "Xu Li's vocabulary remained stuck on grade-one level." In the Chinese diaspora, parents' utilization of multiple resources, their consideration of language planning, and their emotions from children's language proficiencies, as manifested in the above families, reflected a typical aspiration for Chinese language proficiency, especially literacy competency (Li, 2006, 2007; Wang, 2020).

In contrast with the recorded rich resources employed and the tight schedule made by Chinese families in literacy practices, African families scarcely reported their aspiration for or efforts of developing their children's literacy competence. In the data, Ruth is the only parent who reported that she had



FIGURE 1

A sample from Ji Ming's copied texts (left), math mark (middle), and reading collection (right).

attempted to source reading materials in Madi, her heritage language, for her children, Isaiah and David. Even so, her focus on language development with the children was by speaking in Madi at home and making them participate in weekend community activities to maintain the spoken language and ethnic culture. In effect, the divergent attitudes in relation to literacy practices between Chinese and African families are deeply grounded in the linguistic and historical background of their communities as well as the entrenched linguistic hierarchies in the transnational market (see details in the Conclusion section).

In sum, both Chinese and African parents attached importance to maintaining their children's heritage languages and attempted to implement the rule of speaking-only-Chinese/African language(s) at home, but with reference to literacy practices that these parents employed in Australia, Chinese parents, in comparison with their African counterparts, demonstrated greater concerns for their children's reading and writing and made significantly more efforts into developing their children's literacy skills in the heritage language. This noticeable difference should be situated in both the micro-discourse of Chinese and African families as well as the macro-discourse of Chinese and African communities as well as broader societal contexts.

4.2. Parental emotions of heritage language maintenance

In the exploration of the language maintenance experiences of all these Chinese and African families, a full range of parental emotions emerged. This section illustrates the emotional complexities, underlying factors, and arising consequences, by exploring first negative emotions typically known as anger, disappointment, and shame, which then shifts to positive sentiments such as joy, accomplishment, and pride.

As mentioned in the previous section, maintaining Chinese and African heritage languages is a desired family action,

but the maintenance endeavors are often accompanied by stress-triggering experiences, such as children's unfavorable language attitudes and perceived difficulties in achieving optimal results, which often generate unpleasant feelings and impinge negatively on family cohesion. Parents, such as Shi Fan and Bandi, experienced such sense of anger and frustration:

My son [Shi Diwen] doesn't want to work hard at[sic] Chinese language. His dad at times got[sic] annoyed and said[sic] loudly: "Oh, your handwriting looks so ugly, you must WRITE, WRITE, and WRITE!" (Shi Fan)

You can hear Sandile is very fluent in Shona. But at that age, she also feels that she's got to be like others so she PRETENDS that she CAN'T speak Shona when I spoke Shona to her [frowned]. (Bandi)

The emergent conflict between parental expectations (e.g., doing regular Chinese homework or speaking Shona at home) and children's language behavior (e.g., reluctance to write or speak) becomes a frequent cause of unpleasant emotions or intergenerational clashes. When parents confront undesirable results of children's heritage language performance, they may blame themselves or may experience being blamed for their inadequate parenting. Ruth provided an example of how her sister-in-law's twin children felt upset at being unable to fulfill an undergraduate classroom task in relation to heritage language use and how they felt unhappy about not being brought up in their heritage language. Ruth's concluding comments on the story were "The mother and the father doing big mistake. Now, they regret it."

When expressing a sense of disappointment, parents also blamed the school system for not providing (proper) heritage language education. Perceiving that the potential loss of the heritage language was due to parents' heavy work commitments and limited time availability, Mandla blamed the school for failing to take the leading

responsibility of teaching African languages in formal school settings:

We don't have that privilege [economically] and [clears throat] so as much as we want that's why if it was taught in school it would be an advantage to us. We actually need that help to augment our efforts to make it happen. Because we are economic refugees. So our time with our kids as they grow up to actually nurture them language wise is very very[sic] limited compared to other people. (Mandla)

What we have seen is how parental anxiety about the unfeasibility of enacting language maintenance practices is caused by the perceived disadvantage of migration status (e.g., heavy workload and economic difficulties), changes in family dynamics, and a challenge in parental authority when migrant children assume the role of “language brokers” (Renzaho et al., 2017, p. 14), as well as the widely acknowledged lack of minority language support from the institutional level (Lo Bianco, 2009; Piller, 2016). For parents, such as Mandla, the financial burden has barely left them time for nurturing their children's heritage languages, and schools' neglect of African heritage languages further sped up children's heritage language loss. In fact, Australia has laid out various policies to support community language education, but resources are heavily focused on the languages ascribed with more economic status, such as Chinese, Japanese, Korean, and Indonesian (see Lo Bianco, 2009). This means that for African languages and many other minority languages with less economic capital, institutional and societal support is actually limited. Therefore, for families from those language backgrounds that are typical in south–north migrant realities, the intergenerational transmission of their heritage languages becomes greatly challenged (Kamwangamalu, 2013).

However, the maintenance of Chinese and African heritage languages in a migration context may not always lead to stress and anxiety. A sense of enjoyment, fulfillment, and pride, in relation to language maintenance practices and achievements, has also been identified in parental discourses. Across the data, parental positive feelings were closely related to the progress of children's fluency and literacy, as well as the endorsement of parental efforts within and beyond family domains. Mandla, Sandile's father, recounted a pleasant surprise at Sandile's improvement in speaking Ndebele after she was sent to stay with her grandparents:

The vocabulary that she'll be speaking[sic] you'll be like wow. That's when I realized, my parents had a huge influence on her language. And she would speak words that her mother being half Ndebele half Shona sometimes she wouldn't understand. (Mandla)

In the migration context, where generational communication is often disrupted due to children's loss of heritage languages or shift to dominant languages (Fillmore, 1991), the intergenerational transmission of heritage languages which should have strengthened (grand)parent–child ties and family cohesion is often missing. Parents generally feel close to their children when both parties speak the same language and/or feel respected when children

endorse parental language policy and show favorable attitudes to their heritage languages. Ruth revealed such enjoyment with a tone of contentment, “All the time I speak in Madi with my children. Yeah. And they like it.”

Children's achievements in their heritage languages, if acknowledged by their transnational families, ethnic communities, or within institutional settings, do generate a greater sense of parental pride of accomplishment. Ruth's sense of pride seemed ignited when the whole family cheered for her sons' using fluent Madi to talk with family members on the phone:

Uncle and grandma[sic] they are VERY happy. Yeah. they say I'm very proud of you. You teach your kids with[sic] our language. My uncle in Botswana, when he ring and talk[sic] to my kids, to[sic] in my language, and he's so happy. He say[sic], [name] I'm very proud of you because you'll never forget to teach your kids with the Madie. Thank you for yours [...] Oh, my God. You can't even believe, it is so nice. (Ruth)

What has been conveyed, from the frequent use of interjections such as “very happy,” “very proud,” “so nice,” and “my God,” is not only parental feelings of joy, gratification, excitement, and pride but also the important role of language in connecting family members and strengthening family ties. More importantly, the wider endorsement of the heritage language from the social and institutional level significantly enhances parental motivation for achieving higher-level proficiency and begets further success. Li Ni, for example, when expressing her satisfaction with the result of her family language policy, proudly showed evidence of this in a couple of certificates awarded to her daughter—Li Long, in various Chinese language competitions (Figure 2). She related her gratitude specifically to the support from Li Long's Chinese community schools and other language organizations, as she said, “A word of praise from teacher or a small reward from school is more than a thousand words from parents.” Against the widely assumed fact of language loss among three generations (Alba et al., 2002), the potential for a benign circle to operate confirms the feasibility of intergenerational transmission of heritage languages and underscores the significance of concerted efforts from institutions, communities, and families.

It should be noted, as mentioned earlier, that due to different historical backgrounds and linguistic situations, Chinese parents, compared with their African counterparts, demonstrated greater aspirations for and efforts into maintaining their children's literacy competency which can explain the reason why parental emotionality related to Chinese language maintenance is largely associated, in addition to oral-based skills, with parental expectations for children's literacy competency.

In sum, the experience of heritage language maintenance is fraught with emotions, negative (e.g., anger, frustration, regret, and disappointment) and positive (e.g., enjoyment, excitement, fulfillment, and pride). Language emotionality seems to interplay with parental expectations, maintenance results, and children's language performances. In fact, what underly parental emotionality of the heritage language is their language ideologies, particularly, ideologies of power relations. We will demonstrate this in the following section.



FIGURE 2

A sample from Li Long's awarded certificates in Chinese language activities and competitions.

4.3. Parental emotionality and language ideology in relation to power

Across the data, parents' emotional realities and maintenance practices, though related to the consideration of ethnic identity and familial bond (Wang, 2022), are heavily influenced by the concern of economic returns from language investment. This capitalist appraisal of the heritage language reflects the multifaceted layers of inter-power relations between English and minority languages, Chinese and African languages, and intra-power relations within various Chinese or African languages.

Chinese parents, almost in common, relate their motivations for Chinese language maintenance, for the purpose of reaping the economic, occupational, and educational benefits of Chinese in transnational diasporas. These parents, such as Xia Tian's father—Xia Ming, specified the usefulness of learning Chinese with a focus on the socio-economic prospects of China in the global world:

China plays a more and more important role in international[sic] world, whether in economic or political position. When children grow up, they definitely have chances to work with Chinese, whether in Australia or in China. Chinese is useful and I have confidence. (Xia Ming)

The above quote shows how the political and economic status of a home country (China) empowers its social agents (Chinese immigrants) in migration contexts (Australia) to take action of maintaining their language. In fact, Chinese parents across the data demonstrated a detectable sense of pride in the emergent prominence of their heritage language (Mandarin) as well as a

sense of urgency to harness the economic edge in the growing Chinese market. However, no matter how desirous Chinese parents were for their children's competence in Chinese, particularly a functional or high level of literacy, they were often caught in a dilemma when struggling to balance their children's learning of heritage Chinese and school subjects—usually taught in English. As a result, parents generally compromised the value of Chinese for the purpose of achieving academic excellence in schools as they feared that the time spent on the Chinese language would jeopardize their children's performance in high-stake assessments [e.g., tests for opportunity classes (classes in years 5 and 6 which are designed for gifted and talented students), tests for selective high schools (high schools for academically gifted students), and HSC (the higher school certificate)¹ examination]. Although Xu Li's mother, Xu Dai, admitted that it was “a pity” to discontinue, for quite a few times, Xu Li's Chinese learning during preparation toward critical examinations, she firmly stated that the suspension was “the only choice” they could make because Xu Li needed to “make concentrated efforts” to be well-prepared in year 4 for the test of opportunity class, in year 5 for the test of selective high school, and from year 10 for HSC examination. Such inconsistency or conflict in maintenance practices, though implemented with more or less feelings of regret (as shown by Xu Dai), was generally described by parents as “a wise decision for the best of child's education.” Parental attitudes, decisions, and practices in relation to their

1 The Higher School Certificate (HSC) is the culmination of the school career and the highest educational award you can achieve at secondary school in New South Wales.

heritage languages seemed well-constructed on the power structure between majority languages and minority languages, in which English—the majority language in Australia and the lingua franca in the world—enjoys unique prestige above all other languages in and beyond educational discourses; that is, while Chinese is much valued for its rising currency, English is unanimously recognized as the language carrying the maximum weight in child's immediate education and in the more distant future. The stereotypical view of “superior” over “inferior” languages is explicitly or implicitly represented in language attitudes in the broader African diaspora and influenced the subject African families. The view of English as superior seems deeply ingrained in the social and educational discourses of the African population. Mandla pointed out the pervasive admiration of English back in his home country as follows, “English is admired, everything English feels prestigious. Everything. English is better.” Due to the widespread “English fever” (e.g., see [Cho, 2021](#)) observation in the context of South Korea, English is prioritized as the desired means of children's education, as Mandla continued:

Everybody, who has a little bit of money back home sends their child to a group A school [top rating private educational institutions]. There are even Schools[sic] now back home that don't even consider Shona as a subject as a pass. If you don't have a Shona at your O level, it doesn't matter, if you fail Shona at your grade seven, it doesn't matter. You know, so that culture is ingrained in many Zimbabweans. (Mandla)

With the African families living in Australia, parents were divided into those who desired to keep their African languages and those who seemed to make flexible language policies forgoing their heritage languages. Parents admitted that they “never put them under pressure to learn Ndebele” (Lisa) or “If they can't speak you know Ndebele, then let them speak English. I'm okay with that” (Mandla). In terms of the causes to the lax attitude of maintenance policy, parents' own English/French medium education background and the limited social-economic gains from their heritage languages, especially in a society with English as the dominant language, posed as significant factors. Situated in a market with entrenched linguistic hierarchies, African parents, in general, did not hold an optimistic view about the prospect of their own heritage languages. Mandla, for example, felt sad about the unfavorable situation of maintaining the heritage language—Shona within the second generation:

If the mentality [admiring English and ignoring Shona] of the Zimbabwean here in Australia doesn't change, [sic] The next generation won't be speaking any Shona [...]. The future of Shona in Australia is very dark. (Mandla)

In fact, African parents' pessimistic sentiments about the prospect of an African ethnic language form a contrast with Chinese parents' positive expectations of the Chinese heritage language in a migration context. The contrastive sentiment also reflects the hierarchical relationship between the Chinese language (more precisely, Mandarin Chinese) and African ethnic languages in the profit discourse where

the desired former carries heavier currency than the lesser-desired latter. The instrumental hierarchies largely grounded peoples' attitudes to and practices of these languages in both parental and public discourses, as pointed out by Mandla:

You see, even Indians or Chinese people whatever Asians. Their young kids, you'll see them communicating in their vernacular languages. But if a Zimbabwean mom is speaking to a Zimbabwean kid, you'll think they're all Australians behind you if you don't look back. [...] And when you turn back, they're just as black as you are.[sic] (Mandla)

In fact, influences on parental ideologies and emotionality not only arise from the globally entrenched power structure of English, Chinese, and African languages but also from the regionally based competition within minority Chinese languages and within minority African languages.

It should be noted, and as mentioned earlier, that the term “the Chinese language” is tacitly approved by all the Chinese participants as “Mandarin Chinese” rather than any other Chinese varieties. In the fieldwork, there emerged a clear consensus that, from the perspectives of many Chinese, Mandarin is considered a language while other Chinese varieties are considered regional dialects which may retain some economic or symbolic value but are not considered at par with Mandarin. The stereotypical conception of superior Mandarin vs. inferior others also explains why the Chinese parents spontaneously equated the maintenance of the so-called heritage language with that of Mandarin Chinese, regardless of their actual language backgrounds. This habitual use of terminology (e.g., which is regarded as the Chinese heritage language) reveals the hierarchical order between Mandarin—the prestigious national language, and other varieties—usually indexed to locality and lower status (see [Wang, 2020](#)). Based on the instrumental appraisal, Chinese parents' heavy investment into the heritage language was predominantly focused on Mandarin Chinese rather than their own heritage others if non-Mandarin. In the data, only two parents, Jie Yu (Cantonese) and Li Ni (Shanghainese), maintained the oracy of their heritage Cantonese and Shanghainese on a regular basis, while most other non-Mandarin heritage parents had foregone their mother tongues such as Hokkien (Xia Tian's family), Hakka (Mo Jie's family), Shanghainese (Cai Xi's family), and Sichuanese (Yang Mei's family). Parents' lax attitudes to their own spoken languages, in contrast with their devotion to Mandarin, further entrench the power gap between national Mandarin and regional others.

In fact, the aforementioned emotion, either joy or sadness, is in general related to the maintenance outcomes of the privileged Mandarin other than the inferiorized mother tongues. Apart from the pervasive favor of Mandarin, parents at times showed delicate (dis)favor to some specific regional “others,” which seems to reveal a delicate stratification between Chinese varieties other than Mandarin, as indicated by Li Long's mother:

We Shanghai people, more or less, have a sense of pride in being Shanghainese. So, I still want to[sic] my daughter to keep our language. But most of my Shanghainese friends have

given up speaking Shanghaiese with their children because they think Shanghaiese is not that useful and Mandarin is the most important. (Li Ni)

Parents' (e.g., Li Ni) nuanced overtone of some regional Chinese (e.g., Shanghaiese) and an undertone of others (e.g., non-Shanghaiese other than Mandarin) not only define the distinction of Mandarin but also reveal an implicit power layer, which seems to put some sets of "regional varieties," such as Shanghaiese and Cantonese, ahead of similar others (also see Wang, 2020). The embodied language (dis)favor, largely power-oriented, further exposes the intricacies of hierarchical orders existing in Chinese languages, which put Mandarin at the unique top layer, then followed by Shanghaiese/Cantonese due to their economic importance or symbolic value in China, and more others at the bottom level.

The nuanced layers of superiority vs. inferiority in minority African languages were observed from African subjects' language attitudes and emotional responses. For example, the Ewe parents, Phoebe and Efo, revealed their irritability at intra-community linguistic hierarchies that persistently positioned their heritage language as inferior within the Ghanaian community, what Efo described as being spoken to "as if who you are doesn't matter." Both attributed this positioning to politics in their home country which continue to shape their interactions with the majority Twi or Akan speakers even in a migration context where both languages are constructed as minority languages. Experiences that they narrated included interruptions by onlookers at Ghanaian community gatherings where the majority of Twi-speaking community members expected Twi to be spoken. Phoebe told of how on one occasion an Akan woman butted into a private Ewe conversation yelling and demanding that they speak Twi, "HEY HEY HEY NO EWE NO EWE, SPEAK TWI." Efo also recalled "several instances" at church where private Ewe conversations with his wife, Phoebe, were met with admonitions to speak in Twi, "Hey don't. Speak in Twi." Efo explained further that because people knew them to be bilingual in Ewe and Twi, "We can understand their language, but they cannot understand ours," and some of the Twi speakers felt suspicious when they chose to use Ewe and not Twi. This minority positioning within a minority language community is seen as demeaning and threatening to the upkeep of the Ewe language and their speakers' identity/dignity. This negative positioning concerned the parents so much that they expressed relief and praised the study for looking into shedding some light on such power-led linguistic issues faced by minority language speakers.

As illustrated in this section, parents' language ideologies, maintenance practices, and emotional responses are deeply grounded in the power relationships of languages both in global and regional discourses. The power structure revealed in the research not only entrenches the unique prestige of English and features the rising currency of Mandarin Chinese in the global world but also reveals a delicate stratification of minorities within minorities, either in terms of Chinese varieties or African languages.

5. Discussion and conclusion

This study documented language emotionality experienced by Chinese and African immigrant parents in their practices of maintaining their children's heritage languages. It explored how these parents' different emotions interplayed with their language ideologies in relation to power dynamics. Parental emotionality of heritage language maintenance manifested by these Chinese and African families accentuates three characteristics: shared aspiration for language maintenance and divided action in literacy practices, complexities of emotional experiences, and significance of power-inflected ideology in parental emotionality.

First, the shared aspiration for Chinese/African language fluency between Chinese and African families echoes with heritage language desires across ethnic and minority groups in the context of Australia and beyond (see Et-Bozkurt and Yagmur, 2022; Romanowski, 2022). The comparative investigation of the attitudinal divide in the aspect of heritage language literacy offers additional dimensions to FLP from linguistic, educational, historical, and political perspectives. In the research, Chinese parents, compared with their African counterparts, have displayed a stronger drive toward developing their children's reading and writing skills and have made heavier investments into their children's literacy development in the heritage language. This distinct divide in literacy desires and practices of their heritage languages is deeply grounded within the educational and historical backgrounds of Chinese and African diasporas. It reflects the hierarchical relations of languages in contexts before and after migration and closely associates with the differences in linguistic features between the Chinese language (Mandarin Chinese as referred to) and African languages. The Chinese parents received most of their education in China where Mandarin was predominantly used as the medium of instruction. The parents spoke Mandarin, along with some regional dialects if they had any, either in institutions or in private domains. However, the African parents, due to their home countries' colonial history, received their education in the medium of a European language, i.e., English, or French before their migration. The African languages were mainly learned through subject learning in school, Bible reading at the church, or daily communication, as some African participants (e.g., Phoebe) mentioned. For them, literacy gained in formal education is mainly tied to non-African languages (e.g., English and French), and heritage language maintenance tends to be oral-based and is usually tied to informal learning. Thus, the linguistic status in the educational systems of China and African countries underscores, respectively, the significance of Mandarin and European languages such as English and French. For African families, their previously held language habitus which prioritizes English seems further entrenched in Australia where literacy remains legitimately linked to English, and the use of heritage languages is largely confined to private domains. Where the written form is concerned, unlike Mandarin Chinese, which is standardized in simplified Chinese, many African languages do not have identifiable scripts to which their cultures and identities would have been tied.

This constitutes another reason that their cultures, traditions, and values were handed down through oral communication and interaction, such as singing songs, telling tales, and remembering proverbs.

Second, the varieties and complexities of emotion types of the Chinese and African parents in this study have enriched the studies of language emotionality by complementing inquiries usually dominated by negative feelings, such as language anxiety experienced by Turkish families in the Netherlands (Sevinç, 2020) and by Korean families in Australia (Jee, 2020). This research brings forth a whole range of negative (e.g., frustration, disappointment, and shame) and positive (e.g., joy, fulfillment, and pride) emotions in FLP, as well as underlying reasons for such emotional dynamics. Parents' unpleasant feelings are mostly triggered by their children's resistance, undesirable outcomes, and perceived lack of societal support, while parental enjoyment and pride are attributed to their children's endorsement of FLP, their achievements of and progress in heritage language fluency, and/or literacy. The language-related emotionality, which looms large in migration contexts, indicates that heritage language transmission is significant to the psycho-social wellbeing of immigrant parents and their family cohesion (also see Wang, 2022). The difficulty of heritage language maintenance reveals the lack of institutional and societal support for many minority languages, especially those with limited instrumental capital, such as the African languages recorded in this study, while the positive feelings about the maintenance result suggest the potential for heritage language maintenance at the family level. The contrasting emotional experiences underscore the significance of combined efforts for heritage language maintenance from families, communities, and institutions.

Third, the value-laden ideology represented by parental emotions confirms the significance of power relations in the formation and implementation of FLP across ethnic diasporas (see Curdt-Christiansen and Wang, 2018; Et-Bozkurt and Yagmur, 2022). In a previous language research study, the documented power structure falls into the distinction between lesser status and more prestigious languages, typically between the majority language and the minority language, or between the official/"national" language (e.g., Mandarin) and the dialectal language (e.g., non-Mandarin). This research adds a new dimension by revealing the delicate stratification within lesser-role minority languages/varieties used in the Chinese/African diaspora. It is also the first study exploring the intersection of Chinese and African families in the same migration context (Australia) in relation to their languages and emotions. In the research, the multifaceted layers, which are based on linguistic utility, are reflected from the distinct hierarchy between majority English and minority Chinese/African languages, from a materialistic comparison between more profitable Mandarin Chinese and lesser "useful" African languages, and from the hidden tiers within Chinese/African languages/varieties. This practical ideology has significantly shaped the families' language decisions and practices. Both Chinese and African families prioritized "prestigious" English over their heritage languages through all stages of their children's education in Australia,

though they must bear emotional costs arising from a child's language and culture loss. In addition to the linguistic and historical factors, the divergent aspiration for literacy transmission between Chinese and African parents can be a result of their practical appraisal of Chinese and African heritage languages. The rising currency of Mandarin Chinese strengthens Chinese parents' desires for literacy transmission, while the perceived "dark future" (as voiced by Mandla) of African languages (e.g., Shona) lowered parents' expectations for their children's heritage language proficiency. In addition, as the "national" Mandarin enjoys superior status over all other "regional" dialects in the Chinese language market, Chinese parents are willing to acknowledge Mandarin as the legitimate heritage that they should maintain rather than their own heritage varieties that are not Mandarin. It is largely the success or failure of Mandarin Chinese maintenance that generates a parental sense of fulfillment and pride or anxiety and shame. Even in terms of various "regional" Chinese, parents tend to elevate certain Chinese varieties (e.g., Cantonese and Shanghaiese), which carry more materialistic or iconic weight and generate greater pride than other "regional" dialects (e.g., Hokkie and Hakka). In effect, whether or not to invest in Chinese heritage languages and which is the proper heritage language to invest in largely depends on the perceived economic returns in the market of Chinese languages. With their African counterparts, the intricacies of power structure embedded in African heritage languages in their home countries and in the diaspora deeply influence their choice of language maintenance and their emotional fluctuations toward their languages.

Therefore, themes emerging from the research suggest that the pattern of language maintenance and decisions in this regard are more than mere technical linguistic planning but could generate strong emotional reactions and reveal power hierarchies. Hence, the impact of language emotionality is essential for the psycho-social wellbeing of migrant families and has implications for policymakers and heritage language research.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Macquarie University Human Research Ethics Committee. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

YW collected data of Chinese migrant families and took the lead in writing the manuscript. VWT and SD collected data of African migrants, analyzed the data,

provided critical feedbacks, and helped to shape the paper. All authors contributed to the article and approved the submitted version.

Funding

This research was partly supported by Macquarie University's Linguistics Department's Chitra Fernando 2019 Award.

Acknowledgments

We gratefully acknowledge the participants who let us into their emotional worlds regarding their heritage language maintenance and migration experiences.

References

- Akosah-Twumasi, P., Alele, F., Smith, A. M., Emeto, T. I., Lindsay, D., Tsey, K., et al. (2020). Prioritising family needs: a grounded theory of acculturation for sub-saharan African migrant families in Australia. *Soc. Sci. 9*, 17. doi: 10.3390/socsci9020017
- Alba, R., Logan, J., Lutz, A., and Stults, B. (2002). Only English by the third generation? Loss and preservation of the mother tongue among the grandchildren of contemporary immigrants. *Demography* 39, 467–484. doi: 10.1353/dem.2002.0023
- Anthonissen, C., and Stroud, C. (2022). "Family time(s): migrant temporalities in family language planning in the urban African South," in *Multilingualism across the Lifespan, 1 Edn, Vol. 1*, eds U. Royneland and R. Blackwood (London: Routledge), 104–123. doi: 10.4324/9781003125815-8
- Borland, H. (2006). Intergenerational language transmission in an established Australian migrant community: what makes the difference? *Int. J. Sociol. Lang.* 180, 23–41. doi: 10.1515/IJSL.2006.038
- Busch, B. (2012). The linguistic repertoire revisited. *Appl. Linguist.* 33, 503–523. doi: 10.1093/applin/ams056
- Busch, B. (2016). "Biographical approaches to research in multilingual settings: exploring linguistic repertoires," in *Researching Multilingualism: Critical and Ethnographic Perspectives*, eds M. Martin-Jones and D. Martin (London: Routledge), 60–73.
- Cho, J. (2021). English fever and American dreams: the impact of Orientalism on the evolution of English in Korean society. *Engl. Tod.* 37, 142–147. doi: 10.1017/S026607841900052X
- Clyne, M. (2008). The monolingual mindset as an impediment to the development of plurilingual potential in Australia. *Sociolinguist. Stud.* 2, 347–366. doi: 10.1558/sols.v2i3.347
- Colic-Peisker, V., and Deng, L. (2019). Chinese business migrants in Australia: middle-class transnationalism and 'dual embeddedness'. *J. Sociol.* 55, 234–251. doi: 10.1177/1440783319836281
- Curd-Christiansen, X. L. (2009). Invisible and visible language planning: ideological factors in the family language policy of Chinese immigrant families in Quebec. *Lang. Pol.* 8, 351–375. doi: 10.1007/s10993-009-9146-7
- Curd-Christiansen, X. L. (2013). Family language policy: sociopolitical reality versus linguistic continuity. *Lang. Pol.* 12, 1–6. doi: 10.1007/s10993-012-9269-0
- Curd-Christiansen, X. L. (2018). "Family language policy," in *The Oxford Handbook of Language Policy and Planning*, eds J. Tollefson and M. Perez-Milans (Oxford: Oxford University Press), 420–441. doi: 10.1093/oxfordhb/9780190458898.013.21
- Curd-Christiansen, X. L., and Huang, J. (2021). "Pride" and "profit": a sociolinguistic profile of the Chinese communities in Britain. *Int. J. Sociol. Lang.* 2021, 47–72. doi: 10.1515/ijsl-2020-0005
- Curd-Christiansen, X. L., and La Morgia, F. (2018). Managing heritage language development: opportunities and challenges for Chinese, Italian and Pakistani Urdu-speaking families in the UK. *Multilingua* 37, 177–200. doi: 10.1515/multi-2017-0019
- Curd-Christiansen, X. L., and Wang, W. (2018). Parents as agents of multilingual education: family language planning in China. *Lang. Culture Curricu.* 31, 235–254. doi: 10.1080/07908318.2018.1504394
- De Houwer, A. (2017). "Minority language parenting in Europe and children's well-being," in *Handbook on Positive Development of Minority Children and Youth*, eds N. J. Cabrera and B. Leyendecker (Berlin: Springer), 231–246. doi: 10.1007/978-3-319-43645-6_14
- Dovchin, S. (2021). Translanguaging, emotionality, and English as a second language immigrants: Mongolian background women in Australia. *TESOL Quart.* 55, 839–865. doi: 10.1002/tesq.3015
- Et-Bozkurt, T., and Yagmur, K. (2022). Family language policy among second- and third-generation Turkish parents in Melbourne, Australia. *J. Multiling. Multicult. Dev.* 43, 821–832. doi: 10.1080/01434632.2022.2044832
- Fillmore, L. W. (1991). When learning a second language means losing the first. *Early Childh. Res. Quart.* 6, 323–346. doi: 10.1016/S0885-2006(05)80059-6
- Gao, J. (2015). *Chinese Migrant Entrepreneurship in Australia From the 1990s : Case-Studies of Success in Sino-Australian Relations*. Amsterdam: Elsevier.
- Jeon, M. J. (2016). Exploring Korean heritage language learners' anxiety: 'we are not afraid of Korean!'. *J. Multiling. Multicult. Dev.* 37, 56–74. doi: 10.1080/01434632.2015.1029933
- Jeon, M. J. (2020). Heritage language anxiety and major language anxiety experienced by Korean immigrants in Australia. *Int. J. Bilingual Educ. Bilingual.* 2020, 1–17. doi: 10.1080/13670050.2020.1799321
- Kamwangamalu, N. M. (2013). Rural-urban and south-north migrations and language maintenance and shift. *Int. J. Sociol. Lang.* 222, 33–49. doi: 10.1515/ijsl-2013-0031
- King, K. A., Fogle, L., and Logan-Terry, A. (2008). Family language policy. *Lang. Linguist. Compass* 2, 907–922. doi: 10.1111/j.1749-818X.2008.00076.x
- Leist-Villis, A. (2004). *Zweisprachigkeit im Kontext sozialer Netzwerke: unterstützende Rahmenbedingungen zweisprachiger Entwicklung und Erziehung am Beispiel griechisch-deutsch*. München: Waxmann Verlag.
- Li, G. (2006). "The role of parents in heritage language maintenance and development: Case studies of Chinese immigrant children's home practices," in *Heritage Language Development: Focus on East Asian Immigrants, Vol. 32*, ed K. Kondo-Brown (Amsterdam; Philadelphia, PA: John Benjamins Publishing Company), 15–31.
- Li, G. (2007). Second language and literacy learning in school and at home: An ethnographic study of Chinese Canadian first graders' experiences. *Literacy Teach. Learn.* 11, 1–31. Available online at: <https://files.eric.ed.gov/fulltext/EJ899617.pdf>
- Liu, Y. (2022). Commodification of the Chinese language: investigating language ideology in the Chinese complementary schools' online discourse. *Curr. Iss. Lang. Plan.* 23, 319–342. doi: 10.1080/14664208.2022.2037290
- Lo Bianco, J. (2009). *Second Languages and Australian Schooling*. Australian Council. Victoria, VIC: ACER Press. Available online at: <https://research.acer.edu.au/cgi/viewcontent.cgi?article=1007&context=aer> (accessed November 15, 2022).
- Luo, H. (2015). Chinese language learning anxiety: a study of heritage learners. *Heritage Lang. J.* 12, 22–47. doi: 10.46538/hlj.12.1.2
- Mugadza, H. T., Stout, B., Akombi, B. J., Williams Tetteh, V., and Renzaho, A. (2019). The concept of a child within sub-Saharan African migrant homes: reconciling culture and child rights. *Child Fam. Soc. Work* 24, 519–528. doi: 10.1111/cfs.12632

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Ndhlovu, F. (2014). *Becoming an African Diaspora in Australia: Language, Culture, Identity*. London: Palgrave Macmillan.
- Nyarko, K. (2014). *Childrearing, Motherhood and Fatherhood in Ghana* (Dordrecht: Springer Netherlands), 231–239.
- Obanya, P. (1999). "Education, equity and transformation from the perspectives of language education," in *Comparative Perspectives on Language and Literacy. Selected Papers from the work of the 10th World Congress on Comparative Education Societies*, ed L. Limag (Cape Town: UNESCO BREDA), 17–30.
- Obojska, M. A., and Purkharthofer, J. (2018). 'And all of a sudden, it became my rescue': language and agency in transnational families in Norway. *Int. J. Multilingual. Multicult. Dev.* 33, 3–11. doi: 10.1080/01434632.2011.638072
- Ouane, A., and Glanz, C. (2010). *Why and How Africa Should Invest in African Languages and Multilingual Education: An Evidence-and Practice-Based Policy Advocacy Brief*. Hamburg/Tunis: UIL and ADEA.
- Pavlenko, A. (2004). 'Stop doing that, ia komu skazala!': language choice and emotions in parent–child communication. *J. Multiling. Multicult. Dev.* 25, 179–203. doi: 10.1080/01434630408666528
- Piller, I. (2016). *Linguistic Diversity and Social Justice : An Introduction to Applied Sociolinguistics/Ingrid Piller, 1st Edn*. Oxford: Oxford University Press.
- Piller, I., Butorac, D., Farell, E., Lising, L., Motaghi-Tabari, S., and Williams Tetteh, V. (forthcoming). *Life in a New Language*. Oxford: Oxford University Press.
- Renzaho, A. M. N., Dhinra, N., and Georgeou, N. (2017). Youth as contested sites of culture: the intergenerational acculturation gap amongst new migrant communities-parental and young adult perspectives. *PLoS ONE*. 12, 1–19. doi: 10.1371/journal.pone.0170700
- Romanowski, P. (2022). Paternal agency in heritage language maintenance in Australia: Polish fathers in action. *Int. J. Bilingual Educ. Bilingual.* 25, 3320–3332. doi: 10.1080/13670050.2022.2050994
- Sevinç, Y. (2016). Language maintenance and shift under pressure: three generations of the Turkish immigrant community in the Netherlands. *Int. J. Sociol. Lang.* 242, 81–117. doi: 10.1515/ijsl-2016-0034
- Sevinç, Y. (2020). "Language anxiety as a negative emotion in home language maintenance and development," in *Handbook of Home Language Maintenance and Development: Social and Affective Factors*, eds A. C. Schalley and S. A. Eisenclas (Berlin: De Gruyter Mouton), 84–108. doi: 10.1515/9781501510175-005
- Sevinç, Y., and Backus, A. (2019). Anxiety, language use and linguistic competence in an immigrant context: a vicious circle? *Int. J. Bilingual Educ. Bilingual.* 22, 706–724. doi: 10.1080/13670050.2017.1306021
- Sevinç, Y., and Dewaele, J.-M. (2018). Heritage language anxiety and majority language anxiety among Turkish immigrants in the Netherlands. *Int. J. Bilingual.* 22, 159–179. doi: 10.1177/1367006916661635
- Sevinç, Y., and Mirvahedi, S. H. (2022). Emotions and multilingualism in family language policy: introduction to the special issue. *Int. J. Bilingual.* 27, 145–158. doi: 10.1177/13670069221131762
- Shen, C., and Jiang, W. (2023). Parents' planning, children's agency and heritage language education: re-storying the language experiences of three Chinese immigrant families in Australia. *Front. Psychol.* 13, 1083813–1083813. doi: 10.3389/fpsyg.2022.1083813
- Spolsky, B. (2012). Family language policy—The critical domain. *J. Multiling. Multicult. Dev.* 33, 3–11. doi: 10.1080/01434632.2011.638072
- Tannenbaum, M. (2005). Viewing family relations through a linguistic lens: symbolic aspects of language maintenance in immigrant families. *J. Fam. Commun.* 5, 229–252. doi: 10.1207/s15327698jfc0503_4
- Tannenbaum, M., and Yitzhaki, D. (2016). 'Everything comes with a price...': family language policy in Israeli Arab families in mixed cities. *Lang. Intercult. Commun.* 16, 570–587. doi: 10.1080/14708477.2016.1195395
- Taylor, I., and Taylor, M. M. (2014). *Writing and Literacy in Chinese, Korean and Japanese, Vol. 14*. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/swll.14
- Tupas, R. (2015). Pragmatism, Mandarin and political culture in Singapore: recent reprises of an ideology. *J. World Lang.* 2, 94–105. doi: 10.1080/21698252.2016.1183269
- Wang, X. (2017). Family language policy by Hakkas in Balik Pulau, Penang. *Int. J. Sociol. Lang.* 2017, 87–118. doi: 10.1515/ijsl-2016-0058
- Wang, Y. (2020). *The Heritage Language Maintenance of Chinese Migrant Children and Their Families*. Sydney, NSW: Macquarie University. Available online at: <http://minerva.mq.edu.au:8080/vital/access/manager/Repository/mq:71673> (assessed April 1, 2022)
- Wang, Y. (2022). Speaking Chinese or no breakfast: Emotional challenges and experiences confronting Chinese immigrant families in heritage language maintenance. *Int. J. Bilingual.* 27, 232–250. doi: 10.1177/13670069221126043
- Williams Tetteh, V. (2015). *Language, education and settlement: A sociolinguistic ethnography on, with, and for Africans in Australia* (Unpublished PhD), Macquarie University, Sydney, NSW, Australia.
- Wolff, E. (2016). *Language and Development in Africa: Perceptions, Ideologies and Challenges*. Cambridge: Cambridge University Press.
- Wolff, E. (2021). The 'de-indigenization' of African languages. *Acad. Lett.* 2, 2702. doi: 10.20935/AL2702
- Xiao, Y., and Wong, K. F. (2014). Exploring heritage language anxiety: a study of Chinese heritage language learners. *Modern Lang. J.* 98, 589–611. doi: 10.1111/modl.12085



OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway, Norway

REVIEWED BY

Özgür Aydın,
Ankara University, Türkiye
Yao Yao,
Hong Kong Polytechnic University,
Hong Kong SAR, China
Anna-Lena Scherger,
Technical University Dortmund, Germany

*CORRESPONDENCE

Gita Martohardjono
✉ gmartohardjono@gc.cuny.edu

RECEIVED 06 March 2023

ACCEPTED 15 June 2023

PUBLISHED 14 July 2023

CITATION

Martohardjono G, Johns MA, Franciotti P,
Castillo D, Porru I and Lowry C (2023) Use of
the first-acquired language modulates pupil
size in the processing of island constraint
violations.

Front. Psychol. 14:1180989.

doi: 10.3389/fpsyg.2023.1180989

COPYRIGHT

© 2023 Martohardjono, Johns, Franciotti,
Castillo, Porru and Lowry. This is an open-
access article distributed under the terms of
the [Creative Commons Attribution License](#)
(CC BY). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

Use of the first-acquired language modulates pupil size in the processing of island constraint violations

Gita Martohardjono^{1,2*}, Michael A. Johns³, Pamela Franciotti²,
Daniela Castillo², Ilaria Porru² and Cass Lowry²

¹Department of Linguistics and Communication Disorders, Queens College, New York, NY, United States,

²Second Language Acquisition Laboratory, Linguistics Program, The Graduate Center of the City
University of New York, New York, NY, United States, ³Institute for Systems Research, University of
Maryland, College Park, MD, United States

Introduction: Traditional studies of the population called “heritage speakers” (HS) have treated this group as distinct from other bilingual populations, e.g., simultaneous or late bilinguals (LB), focusing on group differences in the competencies of the first-acquired language or “heritage language”. While several explanations have been proposed for such differences (e.g., incomplete acquisition, attrition, differential processing mechanisms), few have taken into consideration the individual variation that must occur, due to the fluctuation of factors such as exposure and use that characterize all bilinguals. In addition, few studies have used implicit measures, e.g., psychophysiological methods (ERPs; Eye-tracking), that can circumvent confounding variables such as resorting to conscious metalinguistic knowledge.

Methodology: This study uses pupillometry, a method that has only recently been used in psycholinguistic studies of bilingualism, to investigate pupillary responses to three syntactic island constructions in two groups of Spanish/English bilinguals: heritage speakers and late bilinguals. Data were analyzed using generalized additive mixed effects models (GAMMs) and two models were created and compared to one another: one with group (LB/HS) and the other with groups collapsed and current and historical use of Spanish as continuous variables.

Results: Results show that group-based models generally yield conflicting results while models collapsing groups and having usage as a predictor yield consistent ones. In particular, current use predicts sensitivity to L1 ungrammaticality across both HS and LB populations. We conclude that individual variation, as measured by use, is a critical factor that must be taken into account in the description of the language competencies and processing of heritage and late bilinguals alike.

KEYWORDS

pupillometry, heritage speakers, late bilinguals, current use, historical use, group vs. individual analyses

1. Introduction

Research on heritage speakers (HS) over the past 20 years has claimed that these childhood bilinguals whose home language is a societal minority language, differ qualitatively in the competence of their first-acquired language¹ (henceforth L1) when compared to other bilinguals (Benmamoun et al., 2013a; Montrul, 2016b; Polinsky, 2018). Specifically, HS are said to diverge in their L1 production (e.g., Fenyvesi, 2005), comprehension (e.g., Polinsky, 2006), lexical diversity (e.g., Hulsen, 2000), and grammatical intuition (e.g., Montrul and Bowles, 2009). This observed variation has led some researchers to the hypothesis that HS are a distinct type of bilingual due to the early age of initial exposure to the L2, although other factors, such as proficiency and attrition have also been suggested (Polinsky, 2016; Montrul, 2022).

While the majority of the HS literature documents behavioral outcomes in the L1, much less is known about heritage language processing from a psycholinguistic perspective, though initial investigations provide evidence that HS processing has both similarities and differences with the processing patterns of other bilingual populations who share their L1 (Madsen, 2018; Martohardjono et al., 2021). The goal of our study is to further the investigation into HS processing by investigating the role of *relative language use* in Spanish/English bilinguals and how it may affect processing of the first-acquired language, Spanish. We take as our starting point the widely stated observation that the bilingual experience is largely determined by the relative interaction with the two languages, and that this interaction can vary greatly from one speaker to the next (Grosjean and Li, 2013), suggesting that individual variation plays a critical role (see also Rothman et al., 2023). We argue that while the categorization of bilingual speakers into distinct types, such as childhood/early/heritage on the one hand, and adult/late on the other, may be intuitively appealing, especially when viewed from the perspective of critical or sensitive periods of language acquisition, it critically ignores the fact that the bilingual experience varies systematically along many dimensions other than age, such as linguistic environment, exposure, input and use. These factors have only recently been included as variables in experimental studies of bilingualism (see section 2.2.2) and our study aims to contribute to this line of inquiry.

More so than any other bilingual “type,” HS bilinguals have primarily been described in the literature as being dominant in the later-acquired, societal majority language (henceforth L2, e.g., Benmamoun et al., 2013b). But language dominance is itself a complex concept determined by a number of factors, such as age of onset (AoO), proficiency, lifetime exposure, use and contexts of use (Montrul, 2016a). Turning specifically to use factors, the variable of focus in our study, we note that dominance in one language, more often than not, entails diminished use of the other, which in turn may affect its processing (Putnam and Sánchez, 2013). The question that arises then

is, does relative use (of the L1 and the L2) affect processing of the L1, and if so, how? Moreover, while L2 dominance may characterize many or most heritage speakers, use of the heritage language (HL) can vary widely. For some, use of the HL is restricted to a limited number of domains, such as family and in particular, elders, thus also limiting the scope of its use. Others, however, are raised and continue to live in a vibrant bilingual community where the HL, in spite of being a societal minority language, is used daily and in a variety of contexts. For these HS, use of the HL may remain high. Therefore, there is likely to be variability in HL use across HS populations, something that has largely been ignored in the HL literature. But HS are not the only bilingual population susceptible to variable use of the L1, as has been amply attested in L1 attrition studies (e.g., Schmid, 2011). Late bilinguals (LB), i.e., those whose acquisition and active use of the L2 occurs only later in life for a variety of reasons, such as university study, work, migration, etc. may also experience variable L1 use. A first step then, is to investigate to what degree relative use of the two languages affects processing of L1 in two groups of bilinguals, HS and LB, who are otherwise only distinguished by age of onset of the L2. If it turns out that use factors affect the two groups in similar ways, the classification of HS as a distinct bilingual “type” becomes less compelling as it may simply be the case that increased use of the L2 has affected processing of the L1 while keeping competence relatively intact. The main innovation we bring to the field of heritage speaker studies, then, is the inclusion of *relative use* as a potential predictor of how the L1 is processed. A second innovation is the application of a methodology that has only recently been introduced in the study of bilingualism and indeed, language in general, namely pupillometry.

The current study is part of a larger project investigating HS and LB who are fluent in both their L1, Spanish, and their L2, English. The HS recruited for this project were either born or had arrived in the US before age 5 and had Spanish as their home and community language. They were schooled in the L2 English starting around age 4 (pre-Kindergarten) and while some became dominant in the L2, they continued to maintain and use their L1. This group was therefore classified as having an early onset of bilingualism. The participants grouped as LB, on the other hand, were born in a Spanish-speaking country, were schooled in Spanish and immigrated to the US in adulthood. While some had limited classroom instruction in English as part of their high school curriculum, this did not occur before age 12. They were fully immersed in English only upon arrival to the US, which for most occurred in their 20s. While everyone in this group had become fluent in the L2 English by the time of testing, they had a late onset of bilingualism, both because they were first exposed to the L2 after age 12 and because they did not have active use of the L2 until adulthood.

The overall purpose of the project is to investigate differences and similarities between HS and LB in the processing of complex sentences (relative clauses and wh-questions) in L1 Spanish. Both implicit (Visual World Paradigm (VWP), EEG, pupillometry) and explicit measures (response accuracy, metalinguistic/acceptability judgments) were taken and compared. Individual-level characteristics were collected in an extensive questionnaire. In the study reported here, we present data from pupillary responses to grammatical and ungrammatical wh-questions involving island constraints (see section 2.3.2). Previous analyses of ERP data on similar structures have been reported in Phillips et al. (2021), and of pupillometry data in Martohardjono et al. (2021). Relevant results from these studies will be discussed in comparison to the results of the present study.

¹ We use the term “first-acquired language” to mean the language first acquired in the home, consistently used by caregivers and in the community in which the family resides. We intend to distinguish it from the more commonly used L1 because of the connotations that this latter term has with monolingual child language and with the much-contested notion of “native speaker.” Nonetheless, for efficiency’s sake and at the suggestion of a reviewer, we use the abbreviation L1 to denote this first-acquired language.

We begin with the characterization of heritage language speakers typically adopted in the literature, as a distinct bilingual “type.” This is followed by a summary of studies that investigate the role of relative use and exposure as determining factors in bilingualism. We then motivate the present study and describe two previous studies we conducted on the processing of *wh*-questions. This section also includes a description of the use of pupillometry in language studies. We then lay out the present study, including analyses and results. We conclude with a discussion of the results and general conclusions.

2. Background and rationale

2.1. Heritage speakers as a cognitively distinct bilingual type

The group commonly known as Heritage Speakers consists of children of immigrants in a particular situation of first language acquisition, involving majority vs. minority language settings. As such, they are raised in the home language, which is the societal minority language, until they reach school age, when they begin education in the societal majority language. Many, though not all, heritage speakers become dominant in that language. Nonetheless, we note that heritage speakers often retain fluency in the home language, depending on their particular linguistic environment—for example if they live in a community where maintenance of the minority language is prevalent, leading to sustained use. This is often the case in Hispanic communities in the US (Otheguy and Zentella, 2011).

Early studies described HS (Benmamoun et al., 2013a) as being qualitatively distinct in their bilinguality² from LB, who are thought to have a more uniform and continuous experience of their first language, are schooled in that language, and acquire the other language only later in life. For example, it was argued that heritage speakers are distinct from child first language learners, and that the particular conditions under which they learn the home language often leads to interrupted, “incomplete acquisition” of that language (see for example Montrul, 2008, 2022). In recent years, this deficit-framing of heritage speakers’ acquisition of their home language has faded in the literature, being replaced with more neutral terms such as “differential acquisition” (Kupisch and Rothman, 2018), and “divergent attainment” (Polinsky and Scontras, 2020). Furthermore, the notion of incompleteness has been challenged by some (e.g., Bayram et al., 2019; Higby et al., 2023) and several studies have reported full acquisition of various aspects of the heritage language grammar (e.g., Guijarro-Fuentes and Schmitz, 2015; Schmitz et al., 2016; Schmitz and Scherger, 2019).

While all bilinguals are susceptible to attrition and cross-linguistic influence—two phenomena common in cases of language contact—HS are in general thought to be even more so (but for counter-examples, see Chang et al., 2011³) since in the process of becoming

dominant in the L2, the mental representation and processing of the L1 can weaken (e.g., Gallo et al., 2021). But the claim that HS bilinguals are *as a group* distinct from other bilingual types implies a significant degree of homogeneity, presumably of a cognitive nature, due to early exposure to the L2. While some argue that this cognitive difference is representational (Polinsky, 2016), others argue that it is primarily located in the processing mechanism (Putnam and Sánchez, 2013; Hopp and Putnam, 2015). Our study does not seek to address that debate directly. It is indeed possible that restructuring of the L1 grammar occurs in some heritage speakers, and that this is likely due to the demands of having to process the two languages continuously. However, restructuring is by no means a phenomenon that is unique to heritage speakers. Competing demands are faced by all bilinguals, including those who acquire the L2 late in life but become fluent in it. As a result, restructuring of the L1 grammar may occur, i.e., attrition. Here we focus instead on *processing* of the L1 and contrast two factors that could arguably affect it. The first is Age of L2 immersion (e.g., Kałamała et al., 2022), which we use as the criterial factor separating HS and LB, early for HS (usually around 6) later for LB (usually after a purported critical period). This comparison will involve a group analysis. The second factor is relative use of L1/L2 which will involve a continuous variable analysis collapsing HS and LB. As there is ample evidence from neuro- and psycholinguistic studies that proficiency in a language modulates its processing (McLaughlin et al., 2010; Morgan-Short et al., 2012a,b; Alemán Bañón et al., 2018) we keep proficiency constant across all participants, including only those who have a self-rated score of 4/5 or higher in the L1.

2.2. Language use as a variable in bilingual studies

2.2.1. Neurolinguistic studies

Although research into relative language use in bilinguals is fairly recent, it has yielded interesting results in a variety of domains. For example, in a number of neurolinguistic studies, Pliatsikas and colleagues have shown that use has structural repercussions. Pliatsikas et al. (2020) proposed a three-stage model for language acquisition and use. When participants are first exposed to a second language, gray matter volume in vocabulary-learning and language-control regions increases (stage 1) but proliferation of these regions fades with L2 experience. During stage 2, language-controlling subcortical and cerebellar adjustments emerge (Abutalebi and Green, 2016) but these adaptations should also fade, possibly resulting in pruning processes and white matter adaptations, indicating less frontal lobe engagement and, consequently, more automation (stage 3).

DeLuca et al. (2019) investigated the effect of exposure and use in bilinguals with a wide range of age of second language acquisition (AoL2A; 0–22 yrs) living in an L2 English majority environment. Two models were compared: the first model included duration (L2 AoA and Length of L2 immersion) and degree/extent of bilingual language use (i.e., L2 exposure and use in the home and other social contexts) as predicting variables. The second model investigated active use of the L2 (total number of years actively using the L2) and immersion (length of

² We use this term as defined in Hamers et al. (2009), to mean the ability of an individual to speak two languages, as distinct from bilingualism, which refers to the effect of two languages in contact on society as a whole.

³ In early work, Chang et al. (2011) found that HS were better at maintaining language-internal and cross-linguistic contrasts than homeland native speakers

and late learners.

time actively using the L2 in immersion settings). Results from both models predicted adaptations to subcortical structure. Specifically, results indicated that sustained active use of the L2 induces structural changes thought to optimize efficacy in L2 processing and production.

The effect of language use on brain structures is also evident in late sequential bilinguals. In two studies comparing highly proficient bilinguals with either high or limited immersion against two groups of monolinguals, [Pliatsikas et al. \(2017\)](#) found subcortical expansion changes in the highly immersed bilingual group compared to the monolingual group, whereas the non-immersion group showed insubstantial changes in comparison to the monolingual speakers. These results suggest that amount of immersion in a bilingual environment has structural correlates in the brain.

2.2.2. Behavioral and psycholinguistic studies of relative language use

A number of studies using behavioral and psycholinguistic measures have investigated whether higher language use leads to faster language processing (e.g., [De Bruin et al., 2016](#)); and whether language use interacts with proficiency regardless of age of first language exposure ([De Carli et al., 2015](#)). Other studies investigated the role of language use from a methodological perspective arguing for this factor to be included when quantifying bilingualism through language background questionnaires (e.g., [Luk and Bialystok, 2013](#); [Kałamała et al., 2022](#)).

[De Bruin et al. \(2016\)](#) investigated the effect of language use in three groups of older Gaelic-English speakers whose L1 is Gaelic. They were categorized as *active* bilinguals (equal use of both languages), *inactive* bilinguals (higher use of English than Gaelic) and monolinguals (very little use of Gaelic across the lifespan). Accuracy and response times (RTs) of the three groups were compared while performing a picture-word matching task in both English and Gaelic. In the English task, they found that while all groups were highly accurate, differences emerged in terms of processing speed. When self-rated English use was treated as a continuous variable, the authors report a significant effect of current language use, namely participants who reported a higher use of English had faster RTs in the English task. In the Gaelic task, findings showed that the inactive group was less accurate than the active group of bilinguals and that the RT difference between Gaelic (the L1) and English (the L2) was larger than in the active group, suggesting that current language use plays a more significant role than early use.

[De Carli et al. \(2015\)](#) investigated and compared the effect of language use and age of acquisition (AoA) on the language proficiency of bilinguals. They administered a sentence recognition task to two groups of speakers: Italian-Spanish bilinguals and highly proficient Spanish and Italian L2 speakers with L1 Italian and L1 Spanish, respectively. Based on current use of each language (Italian and Spanish) across different contexts and according to their responses, participants were classified into two subgroups of users, *occasional* and *intensive* users. In the sentence recognition task, participants were presented with an Italian or Spanish sentence (i.e., “Me gustaría dar un paseo,” *I would like to take a walk*) together with two alternative translations in the other language, an incorrect one (i.e., “Mi piacerebbe dare un passaggio,” *I would like to give a ride*) and a correct one (i.e., “Mi piacerebbe fare una passeggiata”). Findings showed no effect of AoA but a significant effect of language use in both RTs and accuracy. Early bilinguals who keep using both languages intensively were faster and more accurate, as were L2 speakers who were also

intensive users of both languages, with no significant differences between the two groups. This suggests that AoA had little if any effect for these groups. Intensive bilingual users were also significantly faster and more accurate in their responses when compared to occasional bilingual users who did not statistically differ from the L2 speakers.

In a study on Polish-English bilinguals living in Poland and using English on a daily basis [Kałamała et al. \(2022\)](#) investigated the relationship between different measures of bilingualism: Onset of Bilingualism (L2 AoA), L2 Age of Active Communication (AoAC), L2 proficiency, daily use of L2 (time spent using the L2) and patterns of language use (language entropy/diversity of language use, code mixing, code switching).⁴ More specifically, the authors aimed to establish which aspects of bilingualism best predict L2 abilities. Language use and diversity of language use were assessed through two questionnaires each asking about use in several contexts. Many findings were reported, but significant for the purposes of our study were the following: while AoA predicted self-confidence in using the L2 (earlier AoA, higher self-confidence), higher L2 use was a significant predictor of greater vocabulary knowledge; bilinguals with a more diverse language use tend to be more confident in the use of the L2 but have poorer vocabulary knowledge. Finally, frequent language switchers tended to have better vocabulary knowledge, though the effect was modulated by AoA and found only in late bilinguals. Overall, [Kałamała et al.](#)'s findings suggest that diversity of language use (language entropy) and AoA affect self-confidence in using the L2 and that diversity of language use, greater language use, and language switching practices (in late bilinguals only) have an impact on vocabulary knowledge.

The picture that emerges from the above is that the degree of interaction with a language, whether defined as use, current use, diversity in use (language entropy), exposure, or immersion, has distinct outcomes in neural structure, processing (reaction times), and proficiency (accuracy) in both early and late bilinguals. In the following section we return to the question of how this plays out in two purportedly distinct Spanish/English bilingual populations, HS and LB, focusing on processing of the L1 Spanish.

2.3. Preliminary experimental evidence on HS processing

There is preliminary evidence that HS process their L1 differently from both native speakers and late bilinguals, due to early exposure to and use of their L2. Auditory perception studies show that balanced early bilinguals, compared to late bilinguals, have more difficulty processing their L1 in noisy environments ([Weiss and Dempsey, 2008](#)) or discriminating phonological categories ([Peltola et al., 2012](#)). Semantic judgment tasks show that early bilinguals are slower to categorize semantically anomalous items than late bilinguals and monolinguals with the same L1 ([Proverbio et al., 2007](#)).

Our own studies suggest that HS show divergent L1 processing patterns compared to LB in both grammatical and ungrammatical sentences. In an VWP experiment, HS of L1

⁴ Note that [Kałamała et al. \(2022\)](#) refer to language use also as language exposure.

Spanish did not show an expected sensitivity to relative clause type (subject vs. object RC), which LB did (Madsen et al., 2019). Similarly, in pupillometric studies of relative clause processing, late bilinguals showed an expected increase of processing cost for object relative clauses (increased pupil diameter), but HS did not (Madsen, 2018). In a study using event-related potentials, HS showed a sensitivity to different relative clause types, but their pattern of ERP components differed from that of LB (Madsen, 2018). Importantly, the HS tested in these studies all had high levels of proficiency in their L1, similar to that of the LB comparison group. As already mentioned, this was intentional, as we wanted the variable of comparison to be Onset of Bilingualism (AoA of the L2), not L1 proficiency. Taken together, these results suggest that HS' increased dominance in their L2 due to increased early exposure to their L2 has large effects in their syntactic processing of the L1 (see also Montrul, 2016a). However, we subsequently found that when predictor variables of use are included, a more nuanced picture emerges. In a series of studies comparing L1 Spanish/L2 English HS and LB groups we investigated knowledge and processing of L1 ungrammaticality through metalinguistic judgments, EEG, and pupillometry. As in our previous studies, we only included participants who were fluent in both L1 Spanish and L2 English since our critical variable was onset of bilingualism, proxied as age of arrival in the US (HS/early vs. LB/late) and importantly NOT L1 proficiency. A second reason to have fluency as a criterion is the complexity of the particular structures we tested, namely grammatical and ungrammatical wh-questions containing different types of subordinate clauses. Participants classified as LB started active use of English in adulthood while those classified as HS did so at school age. As these studies are relevant to the current one, we will describe them in some detail below.

2.3.1. ERP responses to L1 (un)grammaticality

In an ERP study investigating the processing of syntactic structures that contrast in grammaticality between the L1 Spanish and the L2 English, Phillips et al. (2021) performed two analyses on the same dataset of aurally presented wh-questions in Spanish. The first analysis was based on group differences of L2AoA (LB vs. HS); the second on individual variables of language history and use across the two groups. Spanish and English show a contrast in the obligatory use of the complementizer *que/that* in questions containing embedded clauses.⁵

- 1) Sarah said (that) Lindsey is going to the party.
Who_i did Sarah say (*that) _____i is going to the party?
- 2) Isabel dijo *(que) Julieta va a la fiesta.
¿Quién_i dijo Isabel *(que) _____i va a la fiesta?

(examples from
Phillips et al., 2021)

⁵ Following linguistic convention, brackets () around material indicate optionality of that material; (*) indicates ungrammaticality of the material and *() indicates obligatory inclusion of the material.

Results showed that Spanish wh-questions without a complementizer, evoked an N400 in the LB group but not in the HS group. This suggests that HS processing of these L1 structures is influenced by the L2 English, where an N400 component would not be expected for the equivalent English sentence, supporting the claim that HS as a group hold qualitatively different representations of the L1 Spanish than LB.

The second analysis examined whether individual variables collected in an extensive questionnaire for the same participants were predictive of sensitivity to the (un)grammaticality of these sentences. Predictor variables included current use of L2 English, exposure to L2 English over time, in different settings and with different interlocutors, and L2AoA (LB or HS). Results show that N400 amplitude to ungrammatical L1 Spanish sentences decreased as English use and exposure increased, indicating that increased L2 use diminished sensitivity to ungrammaticality in the L1 Spanish. Crucially, the group variable was not predictive. That is, regardless of whether a subject had early L2AoA (was grouped as HS) or late (was grouped as LB), the amount of L2 English exposure and use influenced processing of L1 Spanish. This result aligns with previous studies using eye-tracking and showing cross-linguistic influence from the L2 on the processing of L1 relative clause attachment (Dussias and Sagarra, 2007) evidencing “permeability” of the L1 after prolonged exposure to an L2.

2.3.2. Pupillometric responses to L1 violations of island constraints

The data we present in the current report are based on a previous pupillometry study which we describe here, comparing LB and HS on island constraints. In that study, we used two separate tasks, administered in separate sessions, 10 to 14 days apart: an acceptability judgment task and a pupillometry task on auditorily presented Spanish sentences varying in (un) grammaticality along a hierarchy known in the syntactic literature as “strong” and “weak” islands (Martohardjono et al., 2021). These structures have been extensively studied in the L2 acquisition literature (e.g., Belikova and White, 2009; Kush and Dahl, 2022), within a native speaker processing framework (e.g., Hofmeister et al., 2013) and within the framework of experimental syntax (Sprouse and Hornstein, 2013). Strong islands included wh-questions out of relative clauses and temporal adverbials which result in a high degree of unacceptability. Weak islands included wh-questions out of wh-islands (e.g., when/how/why) and noun complements. Samples of strong (indicated with **) and weak (indicated with *) islands as illustrated in 3) below were tested against their grammatical counterparts in auditory mode and participants were asked to judge them on a scale of 1–5 for acceptability.

3) Strong Island

Grammatical:

- a. ¿Qué niño comió el dulce mientras que su tía buscaba la comida?
‘Which child ate the candy while his aunt looked for food?’

Strong ungrammatical:

- b. **¿Qué tía, el niño comió el dulce mientras que _____i buscaba la comida?
‘Which aunt did the child eat the candy while looked for food?’

Weak Island

Grammatical:

- a. ¿Qué enfermera confirmó Ignacio que había llevado la medicina?

‘What nurse did Ignacio confirm had brought the medicine?’

Weak ungrammatical:

- b. *¿Qué enfermera, confirmo Ignacio por qué __, había llevado la medicina?

‘What nurse did Ignacio confirm why had brought the medicine?’

Results of the acceptability judgment task showed almost parallel behavior for LB and HS, with significantly higher rejection rates for all ungrammatical structures in both weak and strong conditions, when compared to their grammatical counterparts. This was interpreted as the two groups sharing metalinguistic intuitions about these sentences.

The pupillometry results were more complex, with group means for LB and HS showing partly different pupil dilation patterns. For wh-islands, a weak condition, neither LB nor HS showed the expected increase in pupil dilation for ungrammatical sentences. In fact, both groups showed the reverse pattern, with larger dilation for grammatical than ungrammatical sentences. LB and HS showed slightly different patterns for the other weak constraint, noun complements, although neither in the expected direction. LB showed no significant differences between grammatical and ungrammatical conditions, while HS showed again the reverse pattern, with grammatical sentences eliciting larger pupil dilation than ungrammatical ones, an unexpected result. For the strong constraints, LB and HS converged only in the relative clause condition, with both groups showing a significant increase in pupil dilation for ungrammatical sentences compared to grammatical sentences. In the temporal adverbial type, LB showed the expected pattern, while HS showed no significant differences between grammatical and ungrammatical conditions.

The conclusion we drew from the group analysis of the judgment and pupillometry tasks was that (1) in bilingual populations, processing patterns do not always align with metalinguistic patterns, (2) that the greater between-group differences in processing for LB vs. HS may be reflective of age of L2 acquisition differences, although this was not seen in acceptability judgments, and (3) that the unexpected dilation patterns may be related to the (un)interpretability, rather than the (un)grammaticality of a sentence. Together, our two previous studies indicate that while explicit, metalinguistic knowledge (as measured by judgments) largely coincide across fluent Spanish/English bilinguals, regardless of onset of bilingualism (i.e., HS/LB), processing patterns may in fact diverge across the two groups, lending credence to the claim that the two groups can indeed be considered distinct at some level. However, when use variables are factored in, as they were in the ERP study, these turn out to have an influence on syntactic processing that overrides that of group categorization.

The present study is a follow-up to the ERP and pupillometry studies we just described. In particular, given that metalinguistic judgments of island violations did not differ between HS and LB, but group analyses of the pupillometric data gave inconsistent and even puzzling results; and given further that in the ERP study we found

usage factors significantly modulating L1 processing of ungrammaticality (N400 amplitude) in a structure of L1/L2 contrast (obligatory vs. optional complementizer), we wanted to see (1) whether usage factors might also play a role in determining sensitivity to violations that hold in both languages and (2) whether a model using only usage factors as terms might shed light on the unexpected and puzzling (group) results found in the previous study. Before delving into the details of the present study in section 3, we give a brief description of how pupillometry has been applied in language studies, since it is a fairly recent addition to the methodologies used in the field (e.g., Scherger et al., 2021).

2.4. Pupillometry in linguistic research

Pupillometry is known to be an implicit measurement that allows to track cognitive processes online without relying on explicit responses. Pupil dilation has long been associated with higher cognitive load when completing a task, i.e., the higher the effort the greater the change in pupil dilation. This has been well-attested for roughly half a century in several pioneering studies using this methodology in non-linguistic research (e.g., Hess and Polt, 1964; Kahneman and Beatty, 1966; Schmidtke, 2018 for a review). More recently, a variety of studies have demonstrated that changes in pupil size are linked not only to changes in luminance, but also to aspects of the sympathetic and parasympathetic nervous systems. This includes attention, mobilization, and allocation (Seropian et al., 2022), general arousal levels (Ayasse and Wingfield, 2020), task-evoked changes in arousal (Hopstaken et al., 2015), fatigue (Alhanbali et al., 2021), effortful processing (McGarrigle et al., 2017; Zhao et al., 2019), and surprisal (Zekveld et al., 2018). In linguistic research, pupillometry has gained more prominence only in the past decade, now increasingly used in research on both native and non-native language processing. A great number of studies measured pupil dilation in combination with linguistic tasks testing word and sentence language processing in either auditory mode (e.g., picture-matching tasks, VWP), sentence reading and speech production (Schmidtke, 2018). Scherger et al. (2021) used pupillometry in combination with a production and a comprehension task to investigate potential effects of early and late child bilingualism on double-object constructions in German.

With regard to sentence comprehension, pupil responses are seen to indicate processing overload modulated by syntactic complexity. Engelhardt et al. (2010) tested whether prosody alone and prosody together with visual context has an effect on the online processing of garden-path sentences: they administered two spoken language comprehension tasks to English monolingual speakers, one in which the prosody of the auditory stimuli was manipulated to mismatch the syntactic structure of the garden-path sentence and one in which the task also included pictures either matching or mismatching the intended meaning. Their findings indicate that while a prosodic mismatch tends to elicit greater pupil dilatation, hence higher processing overload during sentence comprehension, the effect of prosody is modulated when combined with visual context. Piquado et al. (2010) compared pupillary responses of younger and older English monolingual adults during a sentence listening and recall

task. The study tested relative clauses manipulated by complexity (i.e., subject and object RC type) and length (with and without modifiers) to test whether processing load was modulated by syntactic complexity. While the younger group had greater pupil dilation when recalling both the more complex (object RC) and longer structures (object RC with modifiers), pupil dilation in the older group was affected only by sentence length. The authors argue that the lack of an effect of syntactic complexity in pupillary responses in the older group supports the hypothesis of “an age-specific dissociation of memory load vs. syntactic complexity effects” (Piquado et al., 2010, p. 12; see also Just and Carpenter, 1993 for a similar study).

In bilinguals, pupil responses have been shown to be modulated by the language experience of the L2 (e.g., Yao et al., 2023). Schmidtke (2014) compared the performance of monolingual and bilingual speakers of English during a word recognition task to test the effect of language experience (among other factors) on lexical retrieval efforts. Schmidtke (2014) found delayed pupil responses in bilinguals at lower level of proficiency, which was interpreted as evidence that lexical retrieval comes at a cost for bilinguals with less experience in the target language. This study is relevant to ours as it at least implicitly addresses use via the measure of experience.

The use of pupillometry in bilingualism research has also been applied to the study of code-switching in Spanish/English bilinguals. A pioneer pupillometry study comparing the online processing of single-word insertion and multi-word alternation in nominal phrases revealed a larger pupil response for the language mixing conditions compared to a unilingual baseline condition and a difference between single-word insertions and alternations in the female condition only, suggesting that the observed difference in pupil dilation is modulated by the gender of the noun (Johns and Dussias, 2022). Pupillometry as a methodology could also have a potentially positive impact in bilingual language assessment in the early diagnosis of developmental language disorders. This methodology has been used for the first time to compare sentence processing in (presumably) monolingual children already diagnosed with Specific Language Impairment (Lum et al., 2017) and proposed as an optimal tool to detect bilingual children at risk early in their linguistic development under the assumption that children with a language disorder may not show an increase in pupil dilation across grammatical and ungrammatical conditions compared to typically developing children (Scherger, 2022). Given its recent flourishing in language studies and its many applicabilities, pupillometry poses as a promising research tool to study cognitive processes in typical and atypical bilingual populations. In our study we use pupil dilation as an indicator of the increased processing load associated with ungrammaticality.

3. The present study: comparing group-level (L2AoA) and individual-level (usage) analyses

The conflicting group results of the AJT and pupillometry tasks in Martohardjono et al. (2021) coupled with the insights gained on the role of L2 use in L1 processing from the ERP study (Phillips et al., 2021) led us to the present study where we performed additional

analyses on a subset of the data collected in the pupillometry task.⁶ In particular, we were interested in comparing group to individual level analyses, whereby the group analysis separated HS and LB by onset of bilingualism, early for HS, late for LB, while in the individual analyses use is measured as a continuous variable across all participants. Secondly, we were interested in investigating how two calculations of use, historical use over time and current use, affect processing of ungrammaticality in the L1. Based on the results of the studies summarized in section 2.2., showing that use variables significantly impact neurological, psycholinguistic, and behavioral outcomes in bilinguals, we hypothesized that relative language use would be predictive of recognition of ungrammaticality in the L1 Spanish: the greater the use of the L1, the greater the recognition of ungrammaticality as measured in relative pupil dilation. Specifically, we expect that due to increased processing load, ungrammatical items will elicit larger pupil dilation than grammatical items across the three conditions tested, wh-islands, temporal adverbial islands, and relative clause islands. However, given that wh-islands are considered weak violations, compared to the other two islands which are considered strong violations, we expect this relative weakness to be reflected in pupil size differential as well. Furthermore, we expect the ungrammatical-grammatical differential to manifest across all participants, modulated by usage. This would show that language use plays a significant role in the processing of the L1 regardless of onset of bilingualism.

3.1. Materials and methods

3.1.1. Participants

Of the 60 participants that took part in the larger study (see section 2.3.2), data from 51 were included in this reanalysis. All were Spanish-English bilinguals between ages 18–45 ($M_{Age} = 28.02$, $SD_{Age} = 7.41$). To assess their eligibility, all participants completed a language history questionnaire and provided self-ratings for their comprehension fluency in Spanish on a five-point scale ($M = 4.88$, $SD = 0.32$). Because our focus was on comparing age of onset of bilingualism to use factors and because LB tend to be more L1-proficient than HS, only participants who rated their fluency as four or higher were included in the study. That is, we did not want variation in L1 proficiency to act as a confound in the design of our study. Based on age of arrival, participants who were either born in the United States or arrived in the country during early childhood were categorized as Spanish heritage speakers (HS: $N = 30$; mean age: 26; Mean AoA Spanish = 0; Mean AoA English = 4.4 (school-age) whereas those whose L2 acquisition occurred after age 15 were considered late bilinguals (LB: $N = 21$; mean age 32; Mean AoA Spanish = 0; Mean AoA English = 15 (instructed learning abroad); Mean AoArrival = 26.

3.1.2. Language background questionnaire

All participants were administered a Language Background Questionnaire (LBQ) in two separate sections (see [Supplementary material](#) for the complete LBQ). The first section,

⁶ Note that this does not include a reanalysis of data from the acceptability judgment task.

based on Li et al. (2006), was administered before the experimental session and included questions about historical language background. Specifically, participants stated their native language and all languages spoken, as well as the Age of Acquisition (AoA), Context and Mode of Acquisition (i.e., where and how) and also self-rated their level of proficiency on a scale from 1 (i.e., *I have limited knowledge of the language*) to 5 (i.e., *I am a native speaker/user of the language*). Participants were asked about their first-learned language, any additional languages they were exposed to in their household while growing up, the degree of the exposure (i.e., languages most spoken), and languages used among members of the household. This first section of the questionnaire also covered questions about participants' educational background, country of residence, and primary language(s) used in their communities and schools attended.

The second section of the LBQ was administered at the end of the experimental session and collected participants' demographic data (i.e., sex, profession, social class) as well as data about participants' current language use and attitudes. The items for this part were created in our lab and focused on relative language ability and use. Participants listed all the languages in which they read and write, the learning age and self-rated their reading/writing ability for each language on a scale from 1 (i.e., *I have a limited reading/writing ability in the language*) to 5 (i.e., *I am a native reader/writer of the language*). Participants were asked about their current language use preferences (i.e., *English, Spanish, Both, N/A*) with members of their family (i.e., father, mother, siblings, children, significant other), work (i.e., boss, co-workers), friends, classmates; and they quantified their use of Spanish (i.e., *mostly, little, none, N/A*) in seven different contexts (home, school work, social activities, reading, listening to the radio/music, watching TV). Participants then quantified their everyday use of both Spanish and English in percentages and specified the contexts in which the interactions typically occur. The final part of the LBQ asked about participants' traveling practices in Spanish-speaking countries and their preferred language (English or Spanish). The LBQ was administered in English.

3.2. Stimuli

The stimuli analyzed for this study consisted of 3 of the 4 structures tested in the original pupillometry study (Martohardjono et al., 2021): Wh-islands, Temporal Adverbial islands, and Relative Clause islands.⁷ All stimuli sentences were recorded in Spanish by a female native speaker and created in couplets, each presenting a declarative statement as context [see example 4–6 (a)], followed by a wh-interrogative [see examples 4–6 (b) and (c)]. Different items were created for each island type in both grammatical [examples in (b)] and ungrammatical [examples in (c)] versions by questioning a noun phrase (NP) inside a syntactic island. The grammatical conditions differed from their ungrammatical counterparts in changing the

questioned NP. The wh-island and temporal adverbial island each had 30 items while the relative clause island condition had 45, totaling 105 target sentences. Each ungrammatical experimental sentence was timestamped for the epoch of interest, i.e., where the ungrammaticality surfaces, whereas in the grammatical sentences, the timestamp was located at the point where the structure of interest begins. The sample stimuli indicate these boundaries with “||.” All participants were presented both the grammatical and ungrammatical versions of each item.

4) Wh-island

- a. Ignacio confirmó por qué la enfermera
 Ignacio confirm.PRET.3SG why the nurse
 había llevado la medicina.
 have.IMP.3SG bring.PART the medicine
 ‘Ignacio confirmed why the nurse had brought the medicine.’
- b. ¿Qué enfermera confirmó Ignacio || que
 what nurse confirm.PRET.3SG Ignacio COMP
 había llevado la medicina?
 have.IMP.3SG bring.PART the medicine
 ‘What nurse did Ignacio confirm had brought the medicine?’
- c. *¿Qué enfermera confirmó Ignacio || por qué
 what nurse confirm.PRET.3SG Ignacio why
 había llevado la medicina?
 have.IMP.3SG bring.PART the medicine
 ‘What nurse did Ignacio confirm why had brought the medicine?’

5) Temporal adverbial island

- a. El niño comió el dulce mientras que su tía
 the child eat.PRET.3SG the candy while COMP his aunt
 buscaba la comida.
 search.IMP.3SG the food
 ‘The child ate the candy while his aunt looked for food.’
- b. ¿Qué niño comió el dulce || mientras que su
 what child eat.PRET.3SG the candy while COMP his
 tía buscaba la comida?
 aunt search.IMP.3SG the food
 ‘What child ate the candy while his aunt looked for food?’
- c. *¿Qué tía || el niño comió el dulce mientras
 what aunt the child eat.PRET.3SG the candy while
 que buscaba la comida?
 COMP search.IMP.3SG the food
 ‘What aunt did the child eat the candy while looked for food?’

6) Relative clause island

- a. Paola hizo el gesto que causó
 Paola make.PRET.3SG the joke COMP cause.PRET.3SG
 la controversia
 the controversy
 ‘Paola made the joke COMP caused the controversy.’

⁷ Complex noun phrases were not included based on syntactic literature showing indeterminacy of judgments of these structures in native speakers.

- b. ¿Qué gesto hizo Paola || que causó
 what joke make.PRET.3SG Paola COMP cause.PRET.3SG
 la controversia?
 the controversy.
 ‘What joke did Paola make that caused the controversy?’
- c. *¿Qué controversia hizo Paola || el gesto
 what controversy make.PRET.3SG Paola the joke
 que causó?
 COMP cause.PRET.3SG
 ‘What controversy did Paola make the joke that caused?’

3.3. Procedure

Stimuli sentences were presented in the aural modality given its suitability for heritage speakers. In each trial, the context sentence was followed by the target sentence, and trials were pseudorandomized over five blocks. Throughout the auditory blocks, participants fixated their gaze on a white “+” marker centered on a black screen. To ensure task engagement, yes/no comprehension probes followed 40% of the trials.⁸ Participants read the task instructions in the language of preference (Spanish or English) and were given a practice block to familiarize themselves with the task.

Tobii TX300 infrared cameras were used to record the pupil diameter and gaze location for each eye separately. Data were gathered at 60Hz for the whole trial (one sample every 16.67 milliseconds) during both the context and target sentences, as well as for the preceding and following 1,000 ms before and after each trial.

3.4. Analysis

3.4.1. Pre-processing

For each trial, any samples that were marked as invalid during recording (a Tobii validity code of 1, 2, 3, or 4) were excluded; this includes the pupil diameter and x- and y-gaze positions for both the left and right eyes. Missing samples were not interpolated as interpolation can increase autocorrelation in the residuals leading to anti conservative models (see van Rij et al., 2019, p. 5). Next, the pupil diameter and x- and y-gaze positions were averaged for the left and right eyes. Data were time-locked to the point of ungrammaticality (and the corresponding position in each grammatical counterpart) with the epoch of analysis extending 2,000 ms (120 samples) from this point. This 2,000-ms window was chosen for two reasons: First, since the onset of the epoch was unique for each sentence, the duration of the epoch was also variable. This time window ensured that >90% of all trials had

data up to this point. Second, 2,000 ms was determined to be sufficient to capture the task-evoked pupil response, given previous research that suggests that the pupillary response emerges roughly 500 to 1,500 ms post-stimulus onset (Hoeks and Levelt, 1993; Winn et al., 2015, 2018; Winn, 2016). The average pupil size was calculated during the 200-ms (12-sample) period before the onset of this epoch, and baseline subtraction was performed to account for non-stimulus-related changes in pupil size during the course of the experiment. Trials where more than 35% of all samples were marked for exclusion were removed,⁹ resulting in 37% of all trials being removed. Participants with an insufficient number of trials within each structural condition were likewise excluded from the analysis for that particular condition only (wh-island: 14 participants; temporal adverbial: 13 participants; relative clause: 9 participants).

3.4.2. Generalized additive mixed models

Data were analyzed using generalized additive mixed-effects models (GAMMs) using the *bam* function in the *mgcv* package (v. 1.8-33; Wood, 2011; Wood et al., 2016) with further model criticism and visualization performed using the *itsadug* package (v. 2.3; Van Rij et al., 2020). GAMMs are ideal for analyzing time-series data, like the task-evoked pupil response (TEPR), as it is able to capture non-linear dependencies in the data as well as account for autocorrelation using an embedded autoregressive (AR-1)—an added benefit over using other modeling techniques such as growth curve analysis. Data for each of the island types was analyzed separately, but all followed the same procedure (see [Supplementary material](#) for the full analysis scripts). First, a maximally specified reference model was fit without the inclusion of an embedded AR1 model in order to determine the appropriate value for the autocorrelation coefficient ρ , which was extracted using the *start_value_rho* function in the *itsadug* package. Next, the model was re-run with an embedded AR1 model with this specified ρ value. The *acf_resid* function in the *itsadug* package was used to ensure that autocorrelation in this final model was within acceptable levels; if not, the ρ value was manually adjusted until the autocorrelation at lag 1 was sufficiently low (<0.2). The *gam.check* function in the *itsadug* package was used to determine the appropriate number of knots, k , for each smooth term in the model. All models were specified to use a scaled- t distribution to account for the non-normal distribution of the data. Time was entered into the model as the sample number, which was re-numbered such that sample 1 was the first sample that corresponded to the start of the epoch of analysis. Given that the epoch extended for 2,000 ms and each sample was approximately 16.67 ms, the total number of samples for the epoch of analysis was 120. In all models, a smooth term for gaze position was included to account for its effects on pupil size (Gagl et al., 2011). This smooth term modeled the x- and y-gaze position as a continuous, non-linear interaction, allowing for the effects of gaze position on pupil size to be modeled directly as a covariate. Lastly,

⁸ For example, after hearing the item *El niño comió el dulce mientras que su tía buscaba la comida* the statement *La tía comió el dulce* appeared on the screen, followed by a Verdadero (in green) and a Falso (in red) button.

⁹ While this amount of excluded data may seem high compared to behavioral methods, where more than 10% of data excluded would be rare, this is not the case for pupillometry data. For a discussion, see [Schmidtke \(2018: 542–543\)](#).

random smooths by participant and by item were included as well (van Rij et al., 2019).

For each island type, two different models were run. The first was a binary coded model (see Wieling, 2018) that estimated the differences between the two groups (LB, HS) and the two conditions (grammatical, ungrammatical) as well as the interaction between them. Given that binary-coded variables represent specific contrasts within the model, this model was subsequently relevelled—in the same way that a linear model might be relevelled—to examine all contrasts of interest. The comparisons of interest were:

1. LB, ungrammatical minus LB, grammatical
2. HS, ungrammatical minus HS, grammatical
3. HS, grammatical minus LB, grammatical
4. HS, ungrammatical minus LB, ungrammatical
5. The difference in the grammaticality effect between HS and LB.

The second model sought to examine the grammaticality effect not as a function of group but rather as a function of current and historical usage of Spanish. Both usage variables were continuous predictors derived from different questions in the LBQ. Current usage was derived from the following question and its subcomponents: “How much Spanish do you use in/at: 1) home, 2) school, 3) work, 4) social activities, 5) reading, 6) listening to the radio/music, and 7) watching TV?” Possible answers were “Mostly,” “Both” (meaning both Spanish and English in equal amounts), “Little,” “None,” and “Not Applicable” (which was excluded). These answers were converted to numeric values (3, 2, 1, and 0, respectively), the average was taken across the seven domains in the question, and the value was rescaled between 0 and 1, where 0 indicated “exclusively Spanish” and 1 indicated “no Spanish.” Historical usage was derived from the following questions: “What languages were spoken in your house growing up?,” “Which of the languages from [the previous question] were used most often?,” “What was the primary language spoken in your local community?,” and “What was the language of instruction?” Possible answers were “Spanish,” “Both Spanish and English,” and “English.” These answers were converted to numeric values (0, 0.5, and 1) respectively and the average was taken across these four questions such that 0 indicated “exclusively Spanish” and 1 indicated “exclusively English.” Two-sample t-tests revealed that, while there was a significant difference in historical usage between the two groups ($t = -10.1$, $p < 0.001$, Figure 1A), there was no difference in current usage between the two groups ($t = 1.11$, $p = 0.27$; Figure 1B). This shows that language use over time separates the late bilinguals from heritage speakers, with LB having more Spanish use, while current use of both Spanish and English overlaps between the two groups.

To model current and historical usage as continuous predictors, they were included as two decomposed tensor product interactions, which allowed pupil size to be modeled not only as a non-linear function of time but also as a non-linear function of usage. This way, it is possible to determine how each term modulates the grammaticality (coded as a binary variable) effect in each of the three island types. Likewise, both current and historical usage were included in the same model so they could be compared against each other directly while also controlling for the other. For example, if the interaction between current usage and grammaticality is significant, but the interaction between historical usage and grammaticality in that same model is non-significant, it suggests that the former is a

better predictor of the grammaticality effect even when the latter is taken into account. Lastly, a significant interaction term indicates that the effect of the usage variable on the grammaticality effect is significantly ‘wiggly’, that is, has a non-zero and non-linear effect on the pupil size as it changes over time. Given that the two models for each island type were non-nested, model comparison was not performed. All figures below are model estimates plotted using the *itsadug* package. R code for all of the analyses and visualizations below can be found in [Supplementary material](#).

3.5. Results

3.5.1. Wh-islands: group differences in the grammaticality effect

The summary of the model with LB, Grammatical as the reference level is provided in Table 1; summary tables of the model when revealed are provided in [Supplementary material](#). Fitted smooths are presented in Figure 2. Model summary tables present the binary-coded difference smooths, represented by terms beginning with ‘Is’, and indicate whether a given difference smooth is significantly different from zero. Difference smooths are always compared back to the reference level, represented by “s(Sample),” which is congruent to the intercept in a linear model. For example, in Table 1, the term ‘s(Sample)’ represents the fitted smooth for late bilinguals (LB) in the grammatical condition. The second term, “IsUngram,” then estimates the difference smooth between ungrammatical and grammatical items for LB; that is, when the only change vis-à-vis the reference level is from grammatical to ungrammatical. Interaction terms (“IsUngramHS”), through the same logic, represent the difference in the grammaticality effect (ungrammatical minus grammatical) between the two groups.

The model suggested that the two groups did not differ from one another in neither the grammatical ($F = 0.03$, $p = 0.99$) nor ungrammatical ($F = 0.66$, $p = 0.58$) conditions. However, there was a significant interaction between Group and Grammaticality ($F = 3.48$, $p = 0.03$) such that HS showed a significant difference between the grammatical and ungrammatical items ($F = 14.03$, $p < 0.001$) while LB did not ($F = 1.79$, $p = 0.19$). However, the effect was in the opposite direction from that expected: grammatical items elicited *larger* pupillary responses than ungrammatical items.

3.5.2. Wh-islands: effects of current and historical usage

The model revealed a significant interaction between current usage and grammaticality ($F = 4.35$, $p < 0.001$; Figure 3), but the interaction between historical usage and grammaticality was non-significant (see Table 2 for model summary). Figure 3 provides the heatmap showing the estimated strength of the grammaticality effect as a function of current usage of Spanish; that is, the difference of ungrammatical minus grammatical, where positive values indicate larger pupil sizes in response to ungrammatical vs. grammatical items. This is also indicated by the coloration: warmer colors indicate a larger positive difference, while cooler colors indicate a smaller (or negative) difference. The x-axis shows the time into the trial, with 0 corresponding to the onset of the epoch. The y-axis displays the usage variable, with lower values indicating more usage of Spanish and higher values indicating more usage of

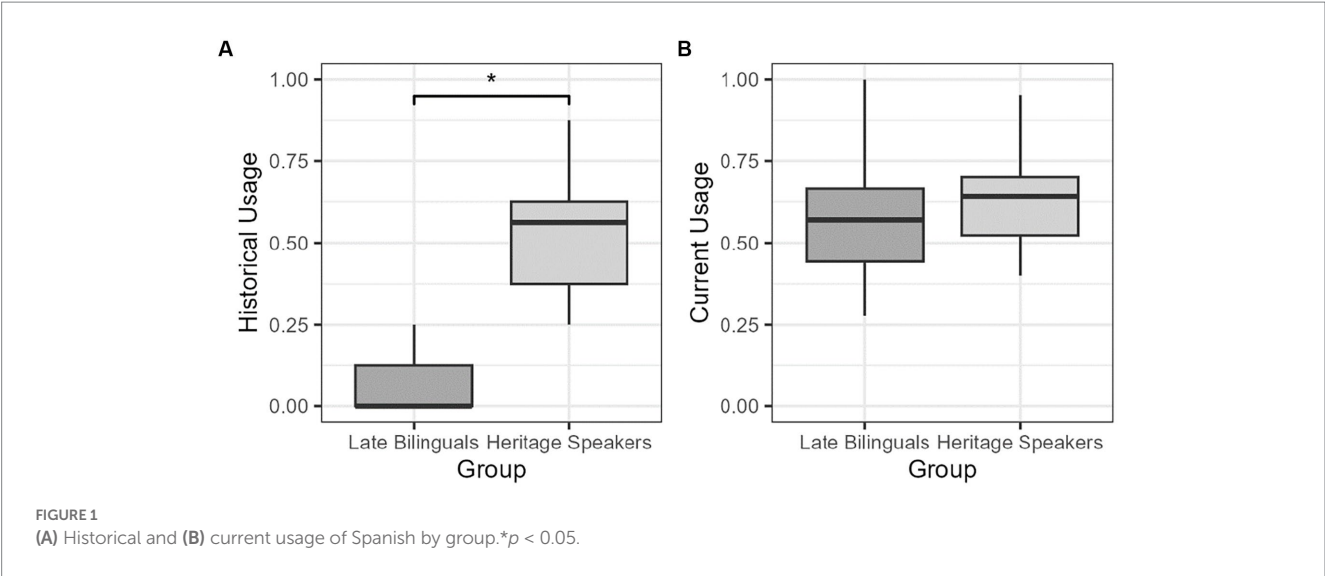


TABLE 1 Wh-islands model summary (reference: LB, grammatical).

Parametric coefficients	β	SE	t	p	
(Intercept)	0.00	0.00	−0.36	0.72	
Smooth terms	EDF	Ref.DF	f	p	
s(Sample)	3.92	4.74	5.34	<0.001	*
s(Sample): IsUngram	2.01	2.01	1.79	0.18	
s(Sample): IsHS	2.57	2.80	0.03	0.92	
s(Sample): IsUngramHS	2.01	2.01	3.48	0.02	*
s(X Gaze, Y Gaze)	37.19	38.80	91.72	<0.001	*
s(Sample, Subject)	174.89	458.00	1.76	<0.001	*
s(Sample, Item)	37.07	299.00	0.64	<0.001	*

* $p < 0.05$.

English. The other three panels present ‘slices’ of the heatmap at different values of Current Usage (noted in the titles), showing the pupillary responses to grammatical and ungrammatical items at these values. More current usage of Spanish (lower values) was associated with a strong grammaticality effect, with ungrammatical items eliciting larger pupil sizes than grammatical items. More current usage of English (higher values), however, was associated with a *reverse* grammaticality effect, with grammatical items eliciting larger pupil sizes than ungrammatical items.

3.5.3. Temporal adverbial islands: group differences in the grammaticality effect

The summary of the model with LB, Grammatical as the reference level is provided in Table 3; summary tables of the model when releveled are provided in Supplementary material. Fitted smooths are presented in Figure 4. The model revealed a significant interaction between Group and Grammaticality ($F = 8.74, p < 0.001$). LB showed a significant effect of grammaticality, with ungrammatical items

eliciting larger pupillary responses than grammatical items ($F = 20.53, p < 0.001$). There was no difference between grammatical and ungrammatical items for the HS.

3.5.4. Temporal adverbial islands: effects of current and historical usage

The model revealed a significant interaction between current usage and grammaticality ($F = 3.48, p = 0.02$; Figure 5), but the interaction between historical usage and grammaticality was non-significant (see Table 4 for model summary). Nonetheless, historical usage did have an overall effect on pupil size that did not differ based on grammaticality ($F = 4.93, p = 0.03$; Figure 6): decreasing historical use of Spanish (i.e., higher values) are associated with overall larger pupillary responses. As for the interaction between current usage and grammaticality, individuals who reported more current use of Spanish showed a strong, late grammaticality effect, while those who reported more current use of English showed a small reversal of this effect late in the epoch.

3.5.5. Relative clause islands: group differences in the grammaticality effect

The summary of the model with LB, Grammatical as the reference level is provided in Table 5; summary tables of the model when releveled are provided in Supplementary material. Fitted smooths are presented in Figure 7. The model revealed a significant interaction between Group and Grammaticality ($F = 3.95, p = 0.02$). While both the LB ($F = 26.96, p < 0.001$) and HS ($F = 9.16, p < 0.001$) showed a significant effect of grammaticality, with ungrammatical items eliciting larger pupillary responses than grammatical items, this effect was larger for the LB than the HS.

3.5.6. Relative clause islands: effects of current and historical usage

The model revealed a significant interaction between current usage and grammaticality ($F = 2.52, p = 0.03$; Figure 8), but the interaction between historical usage and grammaticality was non-significant (see Table 6 for model summary). In this case,

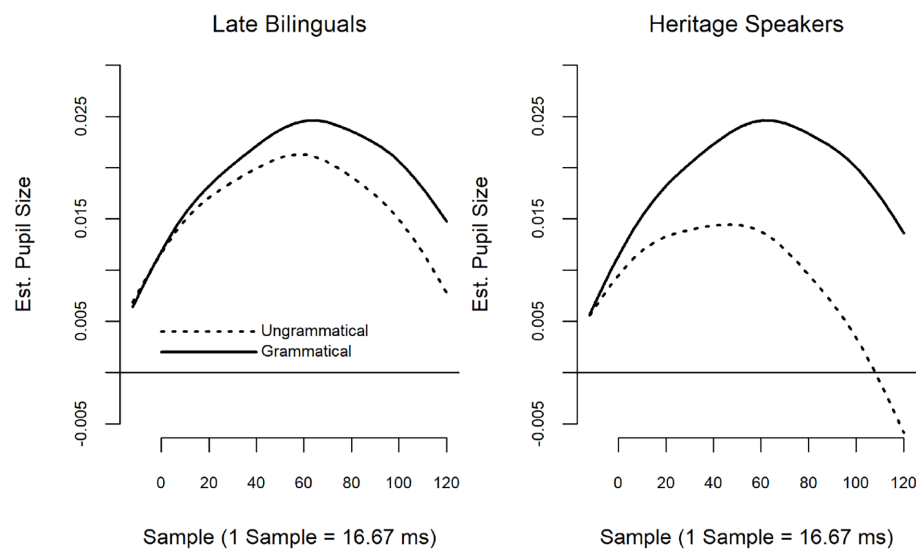


FIGURE 2
Wh-islands fitted smooths: group by grammaticality.

individuals who reported higher current usage of Spanish than English *and* individuals who reported higher usage of English than Spanish both showed late grammaticality effects in the pupillary response. Individuals who reported roughly equal amounts of Spanish and English showed little-to-no differences between the grammatical and ungrammatical conditions.

4. Discussion

A comparison of the group-based and the usage-based models yielded divergent results across the 3 island types. We discuss these separately for each type:

For wh-islands, considered a “weak” island violation in the syntactic literature (e.g., [Torrego, 1984](#)) the model comparing HS to LB as a group did not detect differences in the way either grammatical or ungrammatical items were treated. However, although the interaction between Group and Grammaticality was significant, neither HS nor LB showed patterns that follow predictions of (un)grammaticality. For the HS, the grammatical items elicited significantly larger pupil dilation than ungrammatical ones; for the LB the difference between grammatical and ungrammatical items, while similar to the pattern seen for the HS, did not result in significance (see [Figure 2](#)). This would indicate that neither group perceived the ungrammaticality of wh-island violations. In contrast, the usage-based model showed a significant effect for current, though not for historical use. Moreover, in the usage-based model, the effect was seen in the expected direction, i.e., ungrammatical sentences eliciting larger pupil dilation than grammatical sentences. As illustrated in the heat map and slice diagrams, pupil dilation to ungrammatical Spanish sentences was modulated by whether Spanish or English was used more. As expected, more current Spanish use elicited greater pupil dilation for ungrammatical items, indicating greater processing load and greater sensitivity to Spanish ungrammaticality. Conversely, more

current English use elicited smaller pupil dilation to ungrammatical items, indicating less sensitivity to Spanish ungrammaticality. This reversal is reminiscent of what was found in the ERP study ([Phillips et al., 2021](#)) where overall increased English use was inversely related to N400 amplitude to Spanish ungrammaticality. The fact that divergent results were obtained in the two models points to the importance of looking at data from different angles, in this case, both with a group as well as an individual-level analysis.

For the Temporal Adverbial islands, considered a strong violation in the syntactic literature (here indicated by **), the group model showed the expected effect for LB but not HS. This would indicate that HS are not sensitive to ungrammatical sentences like 7.

- 7) **Que. tía, el niño comió el dulce mientras que ____, buscaba la comida?

** What aunt did the boy eat the candy while ____ looked for food?

On the other hand, when grammaticality is examined via usage variables, we see again that current, but not historical usage is predictive of sensitivity to ungrammatical TA islands. Similar to what we saw in the wh-islands, the heat map and “slice” diagrams showed that greater current use of Spanish elicits a pupil response to these ungrammatical sentences. This effect is again reversed with increased English use where grammatical items elicit greater dilation than ungrammatical ones. Here again, the usage-based results for the TA islands align with the ERP results in [Phillips et al. \(2021\)](#) and stand in contrast to the results from the group-based model, which suggested no sensitivity to strong L1 ungrammaticality for the HS group. Different from the wh-island results, the usage-based model for TA islands did show an effect for historical usage, which was, however, independent of grammaticality but modulated by decreased use of Spanish over time. This suggests that less Spanish use over the lifetime incurred greater processing load for Spanish sentences containing temporal adverbial clauses overall.

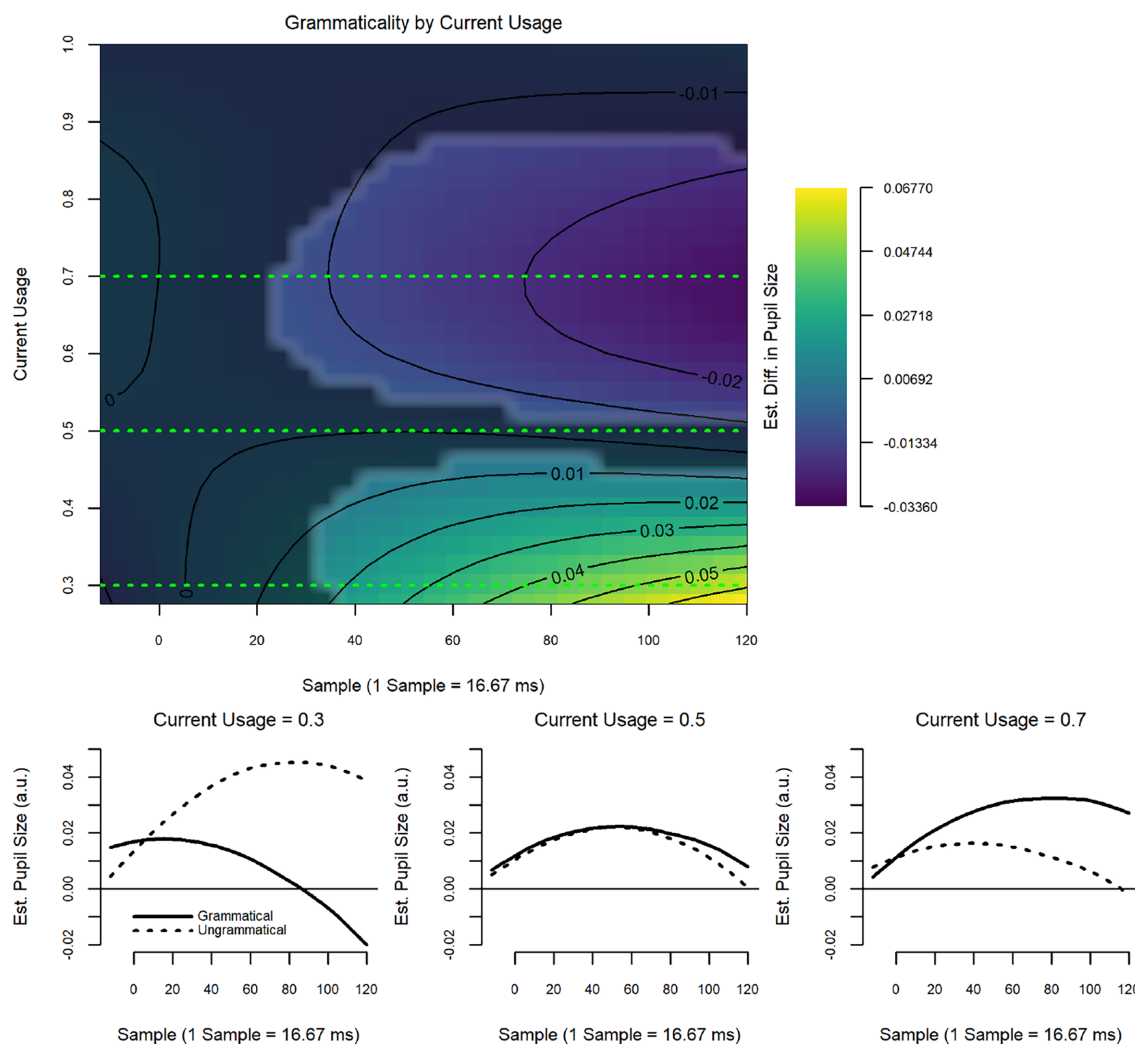


FIGURE 3
Wh-islands: current usage by grammaticality.

Finally, for RC islands, which in the syntactic literature are considered the strongest island violation, the group model showed a grammaticality effect for both LB and HS, although this effect was significantly larger for LB than for HS. From this, one might conclude that HS are less sensitive to these strong violations than are LB. As in the case of TA-island violations, this would suggest a qualitative difference in the way L1 ungrammaticality is processed by HS, compared to LB.

Results from the usage-based model, on the other hand, revealed again that use is a significant variable, modulating detection of ungrammaticality. Here, as for the other two types, current, though not historical usage was predictive of increased pupil dilation for RC island violations. However, in this case, this was true for both more Spanish as well as more English use. While the result for more Spanish is expected, the result for more English is puzzling. A possible interpretation could be that this is a consequence of the strength of this type of violation. Questioning a noun located inside a relative clause (el crítico in 8a below) arguably results not only in a strong violation but in a virtually unparseable and therefore uninterpretable structure (seen in b).

8) Declarative

- a. El cine mostró el documental que **el crítico** odiaba.

'The cinema showed the documentary that **the critic** hated.'

Question:

- b. ** ¿Qué crítico mostró el cine el documental que ____odiaba.

'Which critic did the cinema show the documentary that ____hated?'

Considering that this is true in both Spanish and English, it may be the case that the uninterpretability of such sentences requires increased processing effort regardless of which language is used more, which is what could be reflected in the heat maps and diagrams. What is important for our purposes, is that this was true for individuals in both the HS group and the LB group.

Finally, how can the different results between current and historical use be interpreted? Historical use was measured by questions that asked about language use over the lifetime. As shown in Figure 1A, historical use separates LB participants from HS participants, as would be expected, since onset of bilingualism occurs later for LB than for HS. That is, LB use Spanish more over their

TABLE 2 Wh-islands model summary: usage by grammaticality.

Parametric coefficients	β	SE	t	p	
(Intercept)	0.0002	0.0035	0.07	0.94	
Smooth terms	EDF	Ref.DF	f	p	
s(Sample)	3.88	4.63	5.36	<0.001	*
s(Sample): IsUngram	1.00	1.01	0.36	0.55	
s(Historical Usage)	1.00	1.00	1.52	0.22	
s(Historical Usage): IsUngram	1.01	1.01	1.87	0.09	
s(Current Usage)	1.40	1.46	1.99	0.10	
s(Current Usage): IsUngram	3.79	4.26	7.88	<0.001	*
ti(Sample, Historical Usage)	1.01	1.02	1.18	0.28	
ti(Sample, Historical Usage): IsUngram	2.32	2.79	1.06	0.27	
ti(Sample, Current Usage)	2.04	2.19	2.93	0.05	
ti(Sample, Current Usage): IsUngram	4.35	5.80	4.20	<0.001	*
s(X Gaze, Y Gaze)	37.73	38.90	122.36	<0.001	*
s(Sample, Subject)	159.47	387.00	1.43	<0.001	*
s(Sample, Item)	65.15	299.00	0.38	<0.001	*

* $p < 0.05$.

TABLE 3 Temporal adverbial islands model summary (reference: LB, grammatical).

Parametric coefficients	β	SE	t	p	
(Intercept)	-0.19	0.00	-4.72	<0.001	*
Smooth terms	EDF	Ref.DF	f	p	
s(Sample)	3.24	3.78	3.51	0.01	*
s(Sample): IsUngram	2.35	2.58	20.53	<0.001	*
s(Sample): IsHS	2.01	2.01	7.05	<0.001	*
s(Sample): IsUngramHS	3.45	4.08	8.74	<0.001	*
s(X Gaze, Y Gaze)	38.57	38.99	307.93	<0.001	*
s(Sample, Subject)	208.83	469.00	1.72	<0.001	*
s(Sample, Item)	105.26	300.00	0.96	<0.001	*

* $p < 0.05$.

lifetime than do HS over theirs. Current use, on the other hand, was measured by questions related to language use at and around time of testing. Figure 1B shows that there was overlap between HS and LB in the use of Spanish and English. The result we obtained showing that historical use does not play a role in the detection of ungrammaticality, while current use does, suggests that recent use of the HL affects sensitivity to ungrammaticality, while cumulative use does not.

5. Conclusion

We began this paper with the observation that research on Heritage Speakers has typically labeled these bilinguals as being distinct from other bilinguals, a characterization that is primarily based on age of acquisition (of the L2), L2 dominance, and group analyses. This separation into type and group, we have argued,

largely ignores the heterogeneity that must necessarily hold across all bilingual speakers. This heterogeneity might even be greater for HS than for LB, given the greater linguistic and societal experiences HS encounter. It is reasonable to assume that the great variability and large number of factors, linguistic and extra-linguistic, influencing the bilingual experience of HS should defy attempts of strict categorization of this population, at least from a cognitive perspective.¹⁰ At the same time, the

¹⁰ We are of course not addressing group categorization along social parameters which may very well be applicable to the bilingual labeled “heritage speaker” by virtue of being the child of immigrants whose home language is minoritized. Minoritization of one’s first-acquired language will certainly have various implications, but of a social nature.

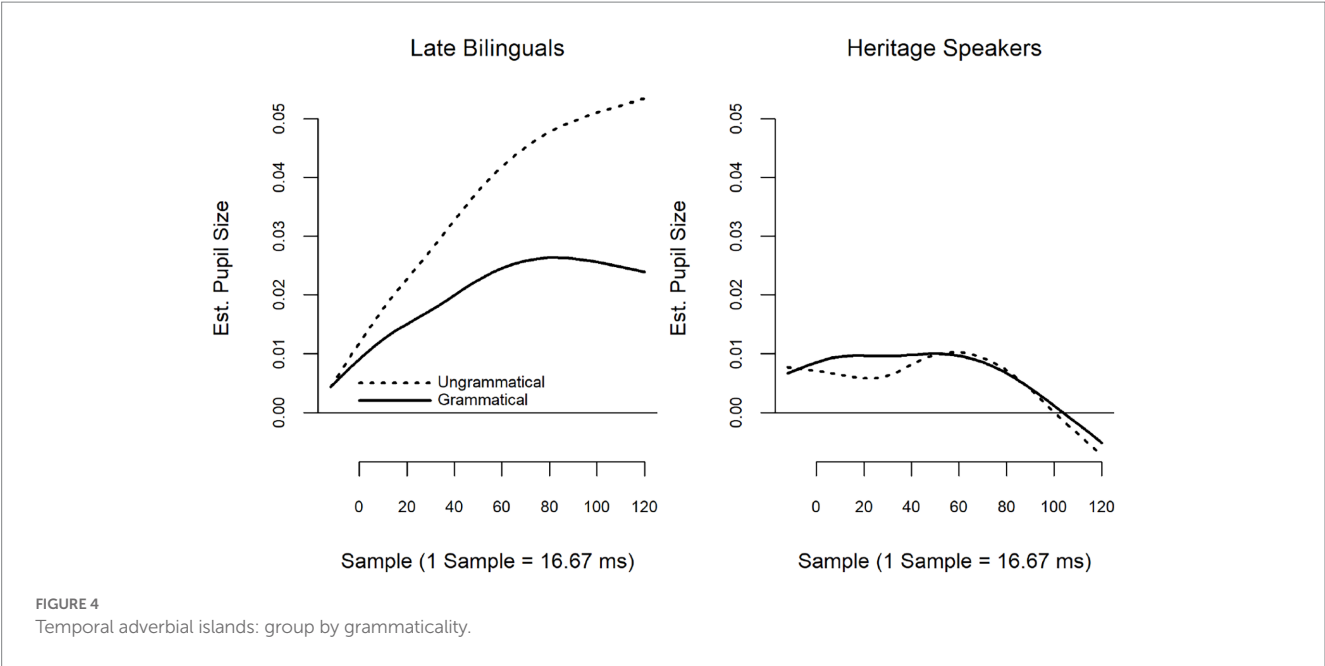


TABLE 4 Temporal adverbial islands model summary: usage by grammaticality.

Parametric coefficients	β	SE	t	p	
(Intercept)	−0.02	0.00	−6.05	<0.001	*
Smooth terms	EDF	Ref.DF	f	p	
s(Sample)	3.02	3.55	1.63	0.16	
s(Sample): IsUngram	2.00	2.00	1.59	0.20	
s(Historical Usage)	1.00	1.00	7.81	0.01	*
s(Historical Usage): IsUngram	2.91	3.31	2.52	0.05	
s(Current Usage)	2.72	2.79	1.35	0.16	
s(Current Usage): IsUngram	1.00	1.01	4.74	0.02	*
ti(Sample, Historical Usage)	1.01	1.02	4.93	0.03	*
ti(Sample, Historical Usage): IsUngram	1.03	1.05	0.50	0.51	
ti(Sample, Current Usage)	2.23	2.37	1.69	0.34	
ti(Sample, Current Usage): IsUngram	2.39	2.92	4.04	0.01	*
s(X Gaze, Y Gaze)	38.40	38.98	267.25	<0.001	*
s(Sample, Subject)	164.96	398.00	1.37	<0.001	*
s(Sample, Item)	91.57	299.00	0.71	<0.001	*

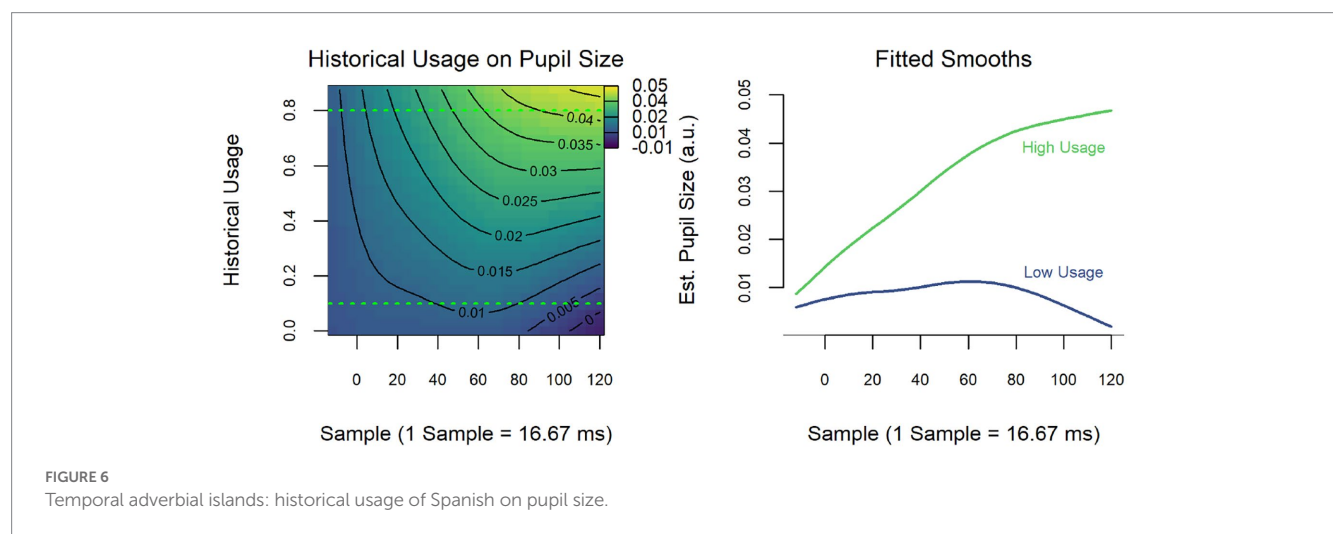
* $p < 0.05$.

TABLE 5 Relative clause islands model summary (Ref: LB, grammatical).

Parametric coefficients	β	SE	t	p	
(Intercept)	−0.02	0.00	−5.98	0.00	*
Smooth terms	EDF	Ref.DF	f	p	
s(Sample)	3.87	4.50	5.71	<0.001	*
s(Sample): IsUngram	2.01	2.01	26.96	<0.001	*
s(Sample): IsHS	2.01	2.02	2.20	0.11	
s(Sample): IsUngramHS	2.01	2.02	3.95	0.02	*
s(X Gaze, Y Gaze)	38.49	38.98	499.95	<0.001	*
s(Sample, Subject)	238.47	508.00	2.23	<0.001	*
s(Sample, Item)	169.52	449.00	1.08	<0.001	*

* $p < 0.05$.

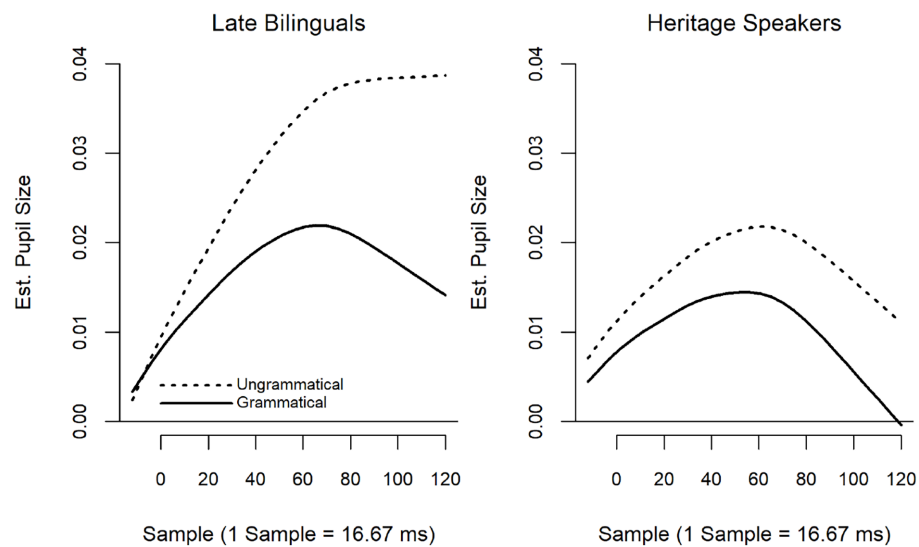


FIGURE 7
Relative clause islands fitted smooths: group by grammaticality.

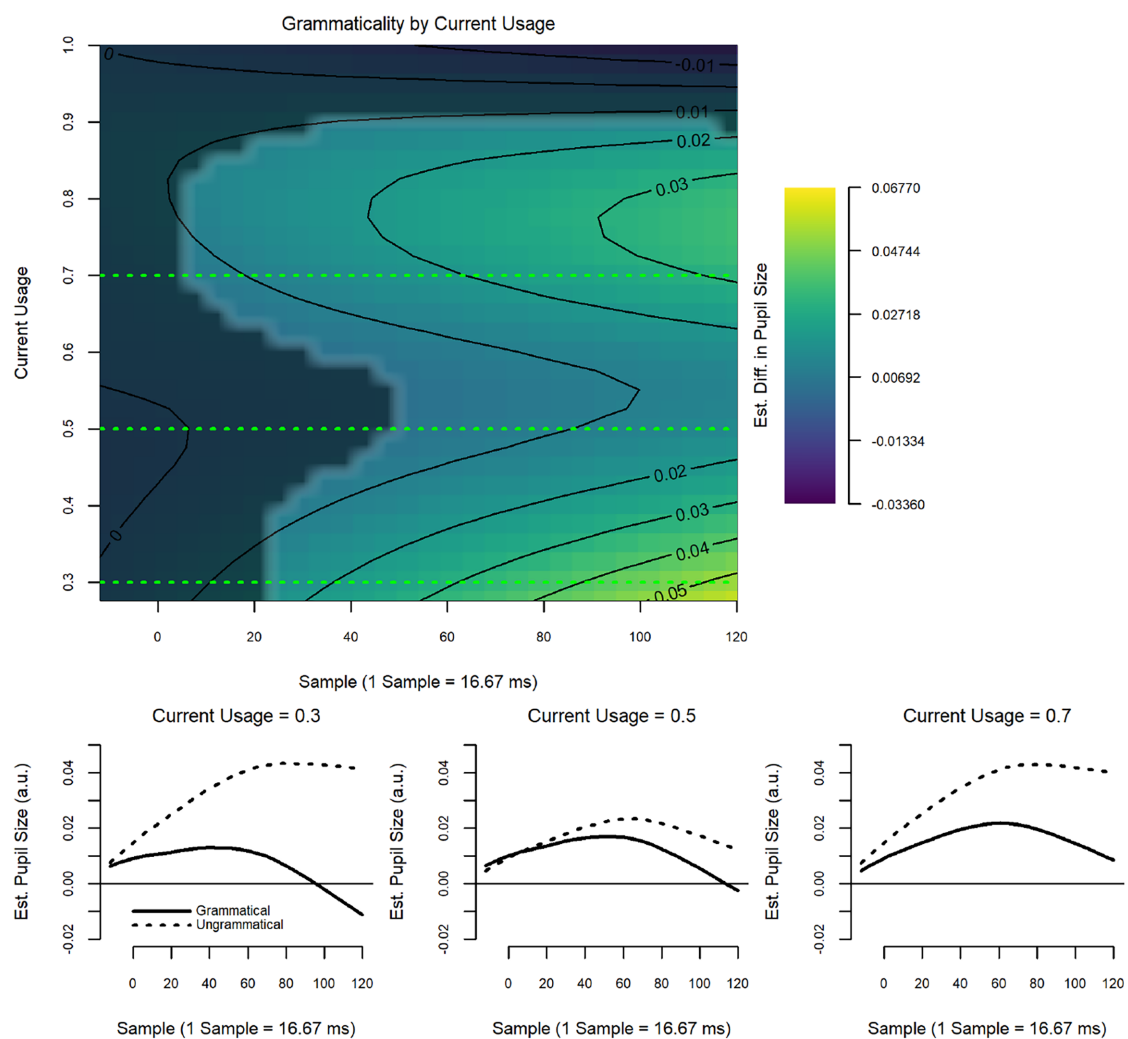


FIGURE 8
Relative clause islands: current usage by grammaticality.

TABLE 6 Relative clause islands model summary: usage by grammaticality.

Parametric coefficients	β	SE	t	p	
(Intercept)	−0.02	0.00	−7.20	<0.001	*
Smooth terms	EDF	Ref.DF	t	p	
s(Sample)	4.11	4.82	6.06	<0.001	*
s(Sample): IsUngram	2.00	2.02	6.77	<0.001	*
s(Historical Usage)	1.01	1.01	0.80	0.37	
s(Historical Usage): IsUngram	1.44	1.50	0.34	0.67	
s(Current Usage)	1.00	1.00	1.51	0.22	
s(Current Usage): IsUngram	4.21	4.54	4.73	<0.001	*
ti(Sample, Historical Usage)	1.01	1.02	2.13	0.14	
ti(Sample, Historical Usage): IsUngram	1.02	1.03	2.76	0.10	
ti(Sample, Current Usage)	1.45	1.57	1.91	0.09	
ti(Sample, Current Usage): IsUngram	3.44	3.98	2.52	0.03	*
s(X Gaze, Y Gaze)	38.43	38.98	524.52	<0.001	*
s(Sample, Subject)	179.76	407.00	1.96	<0.001	*
s(Sample, Item)	145.06	449.00	0.81	<0.001	*

* $p < 0.05$.

heterogeneity of experiential factors determining the bilinguality of the HS individual would lend itself better to a perspective that views that individual as being on a continuum. While it is impossible to address all or even the majority of these factors in an empirical study, we noted that usage variables have largely been ignored in the HS literature, even as there is increasing evidence of its significance in the general bilingual literature. We have therefore chosen to use the continuum of usage in our analysis. Furthermore, the literature has often ignored the inclusion of fluent HS populations and by doing so has risked confounding the effect of age of L2 exposure, L2 dominance and heritage language proficiency. Fluent HS are abundant in areas where there is a vibrant community speaking the HL. We sought to address these issues by comparing exclusively fluent HS and LB and focusing on variables of relative L1 (Spanish) and L2 (English) use, both historical and current. We chose an implicit method that is gaining increased use in experimental studies of language, pupillometry, on the detection of ungrammaticality in the L1 because of its fine temporal resolution and its ability to provide moment by moment data while at the same time being less invasive than neurophysiological methods such as EEG.

Our findings indicate that current use of the L1 Spanish plays a significant role in the detection of ungrammaticality in that language. Specifically, with greater current use of the L1, both weak and strong violations of island constraints in that same language produced increased pupil dilation, indexing greater processing loads for these sentences than when hearing their grammatical counterparts. In addition, the usage-based models showed a reverse grammaticality effect with increased L2 English use, indicating diminished sensitivity to Spanish ungrammaticality in two of the three island types. This indicates that sensitivity to ungrammaticality

in the L1 is attenuated by increased use of the L2, even when the ungrammaticality holds in both languages, as it does in the three island types we investigated, further suggesting that ungrammaticality in the L2 does not reinforce ungrammaticality of equivalent structures in the L1. In general, our findings align with the results reported in other studies investigating the effect of use on neurolinguistic, psycholinguistic and behavioral measures. Importantly, the effect of current use was found across participants, regardless of group adherence, while the group-based models revealed inconsistent and sometimes incoherent results, as in the case of Wh- and TA-island violations for which group analyses failed to reveal sensitivity to L1 ungrammaticality. While we do not discount group analyses as a valid method, we note that the group results obtained here may have masked the sensitivity to L1 ungrammaticality in the heritage speaker participants. Our results support the growing concern in the field that group analyses should not be the only way of investigating language processing across the speaker spectrum, i.e., for monolinguals (Tanner and Van Hell, 2014), bilinguals (Bice and Kroll, 2021) and L2 learners (Grey, 2022). Finally, we take our results to support the idea that the characterization of HS as cognitively distinct from other bilinguals is unwarranted, at least in terms of L1 processing. Usage factors have increasingly shown themselves to be significant in studies of language and should be added to other individual-level characteristics, such as relative proficiency and dominance that are likely to affect all bilingual speakers in the same way.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by IRB Graduate Center, CUNY. The patients/participants provided their written informed consent to participate in this study.

Author contributions

GM, MJ, and CL contributed to the conception and design of the study. MJ and CL organized the database. MJ performed the statistical analysis. GM wrote the first draft of the manuscript. MJ, PE, DC, IP, and CL wrote sections of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was partially funded by the New York State Education Department (NYSED) Grant #016-042, to Martohardjono.

References

- Abutalebi, J., and Green, D. W. (2016). Neuroimaging of language control in bilinguals: neural adaptation and reserve. *Biling. Lang. Cogn.* 19, 689–698. doi: 10.1017/S1366728916000225
- Alemán Bañón, J., Fiorentino, R., and Gabriele, A. (2018). Using event-related potentials to track morphosyntactic development in second language learners: the processing of number and gender agreement in Spanish. *PLoS One* 13:e0200791. doi: 10.1371/journal.pone.0200791
- Alhanbali, S., Munro, K. J., Dawes, P., Carolan, P. J., and Millman, R. E. (2021). Dimensions of self-reported listening effort and fatigue on a digits-in-noise task, and association with baseline pupil size and performance accuracy. *Int. J. Audiol.* 60, 762–772. doi: 10.1080/14992027.2020.1853262
- Ayasse, N. D., and Wingfield, A. (2020). Anticipatory baseline pupil diameter is sensitive to differences in hearing thresholds. *Front. Psychol.* 10:2947. doi: 10.3389/fpsyg.2019.02947
- Bayram, F., Kupisch, T., Cabo, D. P., and Rothman, J. (2019). Terminology matters on theoretical grounds too: coherent grammars cannot be incomplete. *Stud. Second. Lang. Acquis.* 41, 257–264. doi: 10.1017/S0272263119000287
- Belikova, A., and White, L. (2009). Evidence for the fundamental difference hypothesis or not?: island constraints revisited. *Stud. Second. Lang. Acquis.* 31, 199–223. doi: 10.1017/S0272263109090287
- Benmamoun, E., Montrul, S., and Polinsky, M. (2013a). Heritage languages and their speakers: opportunities and challenges for linguistics. *Theoret. Linguist.* 39, 129–181. doi: 10.1515/tl-2013-0009
- Benmamoun, E., Montrul, S., and Polinsky, M. (2013b). Defining an “ideal” heritage speaker: theoretical and methodological challenges. Reply to peer commentaries. *Theoret. Linguist.* 39, 259–294. doi: 10.1515/tl-2013-0018
- Bice, K., and Kroll, J. F. (2021). Grammatical processing in two languages: how individual differences in language experience and cognitive abilities shape comprehension in heritage bilinguals. *J. Neurolinguistics* 58:100963. doi: 10.1016/j.neuroling.2020.100963
- Chang, C. B., Yao, Y., Haynes, E. F., and Rhodes, R. (2011). Production of phonetic and phonological contrast by heritage speakers of mandarin. *J. Acoust. Soc. Am.* 129, 3964–3980. doi: 10.1121/1.3569736
- De Bruin, A., Della Sala, S., and Bak, T. H. (2016). The effects of language use on lexical processing in bilinguals. *Lang. Cogn. Neurosci.* 31, 967–974. doi: 10.1080/23273798.2016.1190024
- de Carli, F. A. B. R. I. Z. I. O., Dessi, B., Mariani, M., Girtler, N., Greco, A., Rodriguez, G., et al. (2015). Language use affects proficiency in Italian–Spanish bilinguals irrespective of age of second language acquisition. *Biling.* 18, 324–339. doi: 10.1017/S1366728914000054
- DeLuca, V., Rothman, J., Bialystok, E., and Pliatsikas, C. (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proc. Natl. Acad. Sci.* 116, 7565–7574. doi: 10.1073/pnas.1811513116
- Dussias, P., and Sagarra, N. (2007). The effect of exposure on syntactic parsing in Spanish–English bilinguals. *Biling. Lang. Cogn.* 10, 101–116. doi: 10.1017/S1366728906002847
- Engelhardt, P. E., Ferreira, F., and Patsenko, E. G. (2010). Pupillometry reveals processing load during spoken language comprehension. *Q. J. Exp. Psychol.* 63, 639–645. doi: 10.1080/17470210903469864
- Fenyvesi, A. (2005). “Hungarian in the United States” in *Hungarian language contact outside Hungary: Studies on Hungarian as a minority language*. ed. A. Fenyvesi (Amsterdam: John Benjamins), 265–318.
- Gagl, B., Hawelka, S., and Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: analysis and correction. *Behav. Res. Methods* 43, 171–1181. doi: 10.3758/s13428-011-0109-5
- Gallo, F., Ramanujan, K., Shtyrov, Y., and Myachikov, A. (2021). Attriters and bilinguals: What's in a name? *Front. Psychol.* 12:558228. doi: 10.3389/fpsyg.2021.558228
- Grey, S. (2022). Variability in native and nonnative language: an ERP study of semantic and grammar processing. *Stud. Second. Lang. Acquis.* 2022, 1–30. doi: 10.1017/S0272263122000055
- Grosjean, F., and Li, P. (2013). *The psycholinguistics of bilingualism*. West Sussex, UK: John Wiley & Sons.
- Guijarro-Fuentes, P., and Schmitz, K. (2015). The nature and nurture of heritage language acquisition. *Lingua* 164, 239–250. doi: 10.1016/j.lingua.2015.05.008
- Hamers, J. F., Blanc, M., Blanc, M. H., and Hamers, J. F. (2009). *Bilinguality and bilingualism*. Cambridge: Cambridge University Press.
- Hess, E. H., and Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science* 143, 1190–1192. doi: 10.1126/science.143.3611.1190
- Higby, E., Gámez, E., and Holguín Mendoza, C. (2023). Challenging deficit frameworks in research on heritage language bilingualism. *Appl. Psycholinguist.*, 1–14. doi: 10.1017/S0142716423000048
- Hoeks, B., and Levelt, W. (1993). Pupillary dilation as a measure of attention: a quantitative system analysis. *Behav. Res. Methods Instrum. Comput.* 25, 16–26. doi: 10.3758/BF03204445
- Hofmeister, P., Casasanto, L. S., and Sag, I. A. (2013). “Islands in the grammar? Standards of evidence” in *Experimental syntax and island effects*. eds. J. Sproule and N. Hornstein (Cambridge: Cambridge University Press), 42–63.
- Hopp, H., and Putnam, M. T. (2015). Syntactic restructuring in heritage grammars: word order variation in Moundridge Schweitzer German. *Linguist. Approaches Biling.* 5, 180–214. doi: 10.1075/Lab.5.2.02Hop
- Hopstaken, J. F., van der Linden, D., Bakker, A. B., and Kompier, M. A. J. (2015). A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology* 52, 305–315. doi: 10.1111/psyp.12339
- Hulsen, M. E. H. (2000). Language loss and language processing: Three generations of Dutch migrants in New Zealand. [dissertation]. [Nijmegen]: Radboud University

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1180989/full#supplementary-material>

- Johns, M. A., and Dussias, P. E. (2022). Comparing single-word insertions and multi-word alternations in bilingual speech: insights from pupillometry. *Languages* 7:267. doi: 10.3390/languages7040267
- Just, M. A., and Carpenter, P. A. (1993). The intensity dimension of thought: pupillometric indices of sentence processing. *Can. J. Exp. Psychol.* 47, 310–339. doi: 10.1037/h0078820
- Kahneman, D., and Beatty, J. (1966). Pupil diameter and load on memory. *Science* 154, 1583–1585. doi: 10.1126/science.154.3756.1583
- Kalamala, P., Senderecka, M., and Wodniecka, Z. (2022). On the multidimensionality of bilingualism and the unique role of language use. *Biling. Lang. Cogn.* 25, 471–483. doi: 10.1017/S1366728921001073
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/136700691665435
- Kush, D., and Dahl, A. (2022). L2 transfer of L1 island-insensitivity: the case of Norwegian. *Second. Lang. Res.* 38, 315–346. doi: 10.1177/02676583209567
- Li, P., Sepanski, S., and Zhao, X. (2006). Language history questionnaire: a web-based interface for bilingual research. *Behav. Res. Methods* 38, 202–210. doi: 10.3758/BF03192770
- Luk, G., and Bialystok, E. (2013). Bilingualism is not a categorical variable: interaction between language proficiency and usage. *J. Cogn. Psychol.* 25, 605–621. doi: 10.1080/20445911.2013.795574
- Lum, J. A., Youssef, G. J., and Clark, G. M. (2017). Using pupillometry to investigate sentence comprehension in children with and without specific language impairment. *J. Speech Lang. Hear. Res.* 60, 1648–1660. doi: 10.1044/2017_JSLHR-L-16-0158
- Madsen, C. N. (2018). *De-centering the monolingual: A psychophysiological study of heritage speaker language processing*. [dissertation]. [New York]: City University of New York
- Madsen, C. N., Stern, M. C., Stevens, L. S., Lowry, C., and Martohardjono, G. (2019). Eye-tracking investigation of relative clause processing in two groups of bilingual speakers. [poster]. In: *93rd annual meeting of the linguistic Society of America (LSA)*, New York, NY. Available at: https://slal.commons.gc.cuny.edu/wp-content/blogs.dir/1833/files/2019/01/RCeye_Poster_LSA2019_v2.pdf
- Martohardjono, G., Lowry, C., Johns, M. A., Phillips, I., and Madsen, C. N. (2021). “Bilingual judgments and processing of Spanish wh-gap constructions: an exploratory study of cross-linguistic influence and island strength” in *Syntax Processing*. ed. V. Torrens (Newcastle upon Tyne, UK: Cambridge Scholars Publishing), 127–150.
- McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., and Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology* 54, 193–203. doi: 10.1111/psyp.12772
- McLaughlin, J., Tanner, D., Pitkanen, I., Frenck-Mestre, C., Inoue, K., Valentine, G., et al. (2010). Brain potentials reveal discrete stages of L2 grammatical learning. *Lang. Learn.* 60, 123–150. doi: 10.1111/j.1467-9922.2010.00604.x
- Montrul, S. (2008). *Incomplete Acquisition in Bilingualism: Re-examining the age factor incomplete acquisition in bilingualism*. Amsterdam: John Wiley & Sons.
- Montrul, S. (2016a). “Age of onset of bilingualism effects and availability of input in first language attrition” in *Bilingualism across the lifespan: Factors moderating language proficiency*. eds. E. Nicoladis and S. Montanari (Washington, DC: American Psychological Association), 141–161.
- Montrul, S. (2016b). *The Acquisition of Heritage Languages*. Cambridge, UK: Cambridge: Cambridge University Press.
- Montrul, S. (2022). *Native speakers, interrupted: Differential object marking and language change in heritage languages*. Cambridge, UK: Cambridge University Press.
- Montrul, S., and Bowles, M. (2009). Back to basics: differential object marking under incomplete acquisition in Spanish heritage speakers. *Bilingualism* 12, 363–383. doi: 10.1017/CBO9781139030502
- Morgan-Short, K., Finger, I., Grey, S., and Ullman, M. T. (2012a). Second language processing shows increased native-like neural responses after months of no exposure. *PLoS One* 7:e32974. doi: 10.1371/journal.pone.0032974
- Morgan-Short, K., Steinhauer, K., Sanz, C., and Ullman, M. T. (2012b). Explicit and implicit second language training differentially affect the achievement of native-like brain activation patterns. *J. Cogn. Neurosci.* 24, 933–947. doi: 10.1162/jocn_a_00119
- Otheguy, R., and Zentella, A. C. (2011). *Spanish in New York: Language contact, dialectal leveling, and structural continuity*. New York: Oxford University Press.
- Peltola, M. S., Tamminen, H., Toivonen, H., Kujala, T., and Näätänen, R. (2012). Different kinds of bilinguals—different kinds of brains: the neural organisation of two languages in one brain. *Brain Lang.* 121, 261–266. doi: 10.1016/j.bandl.2012.03.007
- Phillips, I., Martohardjono, G., Madsen, C. N., and Schwartz, R. G. (2021). “Bilingual processing of the first-acquired language: are heritage speakers and late bilinguals really that different?” in *Language in development: A cross-linguistic perspective*. eds. G. Martohardjono and S. Flynn (Cambridge, MA: MIT Press), 219–317.
- Piquado, T., Isaacowitz, D., and Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology* 47, 560–569. doi: 10.1111/j.1469-8986.2009.00947.x
- Platsikas, C., DeLuca, V., Moschopoulou, E., and Saddy, J. D. (2017). Immersive bilingualism reshapes the core of the brain. *Brain Struct. Funct.* 222, 1785–1795. doi: 10.1007/s00429-016-1307-9
- Platsikas, C., DeLuca, V., and Voits, T. (2020). The many shades of bilingualism: language experiences modulate adaptations in brain structure. *Lang. Learn.* 70, 133–149. doi: 10.1111/lang.12386
- Polinsky, M. (2006). Incomplete acquisition: American Russian. *J. Slav. Linguist.* 191–262.
- Polinsky, M. (2016). Structure vs. use in heritage language. *Linguist. Vanguard* 16, 1–14. doi: 10.1515/lingvan-2015-0036
- Polinsky, M. (2018). Bilingual children and adult heritage speakers: the range of comparison. *Int. J. Bilingual.* 22, 547–563. doi: 10.1177/1367006916656048
- Polinsky, M., and Scontras, G. (2020). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245
- Proverbio, A. M., Roberta, A., and Alberto, Z. (2007). The organization of multiple languages in polyglots: interference or independence? *J. Neurolinguistics* 20, 25–49. doi: 10.1016/j.jneuroling.2006.01.003
- Putnam, M. T., and Sánchez, L. (2013). What’s so incomplete about incomplete acquisition? A prolegomenon to modeling heritage language grammars. *Linguist. Approaches Biling.* 3, 478–508. doi: 10.1075/lab.3.4.04put
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J. A., Gharibi, K., et al. (2023). Monolingual comparative normativity in bilingualism research is out of “control”: arguments and alternatives. *Appl. Psycholinguist.* 44, 316–329. doi: 10.1017/S0142716422000315
- Scherger, A.-L. (2022). Rethinking bilingual language assessment: considering implicit language acquisition mechanisms by means of pupillometry. *Res. Methods App. Linguist.* 1:100014. doi: 10.1016/j.rmal.2022.100014
- Scherger, A. L., Urbanczik, G., Ludwigs, T., and Kizilirmak, J. M. (2021). The bilingual native speaker competence: evidence from explicit and implicit language knowledge using elicited production, sentence-picture matching, and Pupillometry. *Front. Psychol.* 12:717379. doi: 10.3389/fpsyg.2021.717379
- Schmid, M. S. (2011). *Language attrition*. Cambridge: Cambridge University Press.
- Schmidtke, J. (2014). Second language experience modulates word retrieval effort in bilinguals: evidence from pupillometry. *Front. Psychol.* 5, 1–16. doi: 10.3389/fpsyg.2014.00137
- Schmidtke, J. (2018). Pupillometry in linguistic research: an introduction and review for second language researchers. *Stud. Second. Lang. Acquis.* 40, 529–549. doi: 10.1017/S0272263117000195
- Schmitz, K., Di Venanzio, L., and Scherger, A. L. (2016). Null and overt subjects in Italian and Spanish heritage speakers in Germany. *Lingua* 180, 101–123. doi: 10.1016/j.lingua.2016.04.004
- Schmitz, K., and Scherger, A.-L. (2019). Effects of age and education on variable but native heritage grammars: theoretical and empirical implications for the null subject parameter. *Applied Linguistics Review*, 10, 443–467. doi: 10.1515/applrev-2017-0093
- Seropian, L., Ferschneider, M., Cholvy, F., Michey, C., Bidet-Caulet, A., and Moulin, A. (2022). Comparing methods of analysis in pupillometry: application to the assessment of listening effort in hearing-impaired patients. *Heliyon* 8:e09631. doi: 10.1016/j.heliyon.2022.e09631
- Sprouse, J., and Hornstein, N. (2013). *Experimental syntax and island effects*. Cambridge: Cambridge University Press.
- Tanner, D., and Van Hell, J. G. (2014). ERPs reveal individual differences in morphosyntactic processing. *Neuropsychologia* 56, 289–301. doi: 10.1016/j.neuropsychologia.2014.02.002
- Torrego, E. (1984). On inversion in Spanish and some of its effects. *Linguist Inquiry* 15, 103–129.
- Van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., and Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends Hear.* 23:2331216519832483. doi: 10.1177/2331216519832483
- Van Rij, J., Wieling, M., Baayen, R. H., and van Rijn, H. (2020). *Itsadug: interpreting time series and autocorrelated data using GAMMs. R package version, 2–4*.
- Weiss, D., and Dempsey, J. J. (2008). Performance of bilingual speakers on the English and Spanish versions of the hearing in noise test (HINT). *J. Am. Acad. Audiol.* 19, 005–017. doi: 10.3766/jaaa.19.1.2
- Wieling, M. (2018). Analyzing dynamic phonetic data using generalized additive mixed modeling: a tutorial focusing on articulatory differences between L1 and L2 speakers of English. *J. Phon.* 70, 86–116. doi: 10.1016/j.wocn.2018.03.002
- Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends Hear* 20, 1–17. doi: 10.1177/2331216516669723
- Winn, M., Edwards, J., and Litovsky, R. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear Hear.* 36, e153–e165. doi: 10.1097/AUD.0000000000000145

Winn, M. B., Wendt, D., Koelewijn, T., and Kuchinsky, S. E. (2018). Best practices and advice for using Pupillometry to measure listening effort: an introduction for those who want to get started. *Trends Hear* 22, 1–32. doi: 10.1177/2331216518800869

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73, 3–36. doi: 10.1111/j.1467-9868.2010.00749.x

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *J. Am. Stat. Assoc.* 111, 1548–1563. doi: 10.1080/01621459.2016.1180986

Yao, Y., Connell, K., and Politzer-Ahles, S. (2023). Hearing emotion in two languages: a pupillometry study of Cantonese–mandarin bilinguals' perception of affective cognates in L1 and L2. *Biling. Lang. Cogn.* 1–14. doi: 10.1017/S1366728922000931

Zekveld, A. A., Koelewijn, T., and Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: current state of knowledge. *Trends Hear* 22:7174. doi: 10.1177/2331216518777174

Zhao, S., Bury, G., Milne, A., and Chait, M. (2019). Pupillometry as an objective measure of sustained attention in young and older listeners. *Trends Hear* 23:7815. doi: 10.1177/2331216519887815



OPEN ACCESS

EDITED BY

Sergio Miguel Pereira Soares,
Max Planck Institute for Psycholinguistics,
Netherlands

REVIEWED BY

Seçkin Arslan,
Centre National de la Recherche Scientifique
(CNRS), France
Angelika Golegos,
University of Konstanz, Germany

*CORRESPONDENCE

Onur Özsoy
✉ onur.oezsoy96@gmail.com

RECEIVED 31 January 2023

ACCEPTED 12 June 2023

PUBLISHED 19 July 2023

CITATION

Özsoy O, Çiçek B, Özal Z, Gagarina N and
Sekerina IA (2023) Turkish-German heritage
speakers' predictive use of case:
webcam-based vs. in-lab eye-tracking.
Front. Psychol. 14:1155585.
doi: 10.3389/fpsyg.2023.1155585

COPYRIGHT

© 2023 Özsoy, Çiçek, Özal, Gagarina and
Sekerina. This is an open-access article
distributed under the terms of the [Creative
Commons Attribution License \(CC BY\)](#). The use,
distribution or reproduction in other forums is
permitted, provided the original author(s) and
the copyright owner(s) are credited and that
the original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Turkish-German heritage speakers' predictive use of case: webcam-based vs. in-lab eye-tracking

Onur Özsoy^{1*}, Büsra Çiçek², Zeynep Özal¹, Natalia Gagarina¹ and
Irina A. Sekerina³

¹Leibniz Center for General Linguistics (ZAS), Berlin, Germany, ²Berlin School of Mind and Brain,
Humboldt University of Berlin, Berlin, Germany, ³College of Staten Island, New York, NY, United States

Recently, Özge et al. have argued that Turkish and German monolingual 4-year-old children can interpret case-marking predictively disregarding word order. Heritage speakers (HSs) acquire a heritage language at home and a majority societal language which usually becomes dominant after school enrollment. Our study directly compares two elicitation modes: in-lab and (remote) webcam-based eye-tracking data collection. We test the extent to which in-lab effects can be replicated in webcam-based eye-tracking using the exact same design. Previous research indicates that Turkish HSs vary more in the comprehension and production of case-marking compared to monolinguals. Data from 49 participants—22 Turkish monolinguals and 27 HSs—were analyzed using a binomial generalized linear mixed-effects regression model. In the Accusative condition, participants looked for the suitable Agent before it is appeared in speech. In the Nominative condition, participants looked for the suitable Patient before it is appeared in speech. HSs were able to use morphosyntactic cues on NP1 to predict the thematic role of NP2. This study supports views in which core grammatical features of languages, such as case, remain robust in HSs, in line with the Interface Hypothesis. We were able to replicate the effect of the predictive use of case in monolinguals using webcam-based eye-tracking, but the replication with heritage speakers was not successful due to variability in data collection contexts. A by-participant analysis of the results revealed individual variation in that there were some speakers who do not use case-marking predictively in the same way as most monolinguals and most HSs do. These findings suggest that the predictive use of case in heritage speakers is influenced by different factors, which may differ across individuals and affect their language abilities. We argue that HSs should be placed on a native-speaker continuum to explain variability in language outcomes.

KEYWORDS

sentence processing, bilingualism, predictive processing, eye-tracking, visual word paradigm, heritage language, Turkish

1. Introduction

In languages with flexible word order grammatical case on noun phrases (NPs) is a predictive feature that allows comprehenders to anticipate thematic roles of upcoming referents. Prediction in spoken language comprehension by monolingual adults has been firmly established in the sentence processing research. However, whether children acquiring two L1s and second language (L2) learners can anticipate the thematic roles of NPs based on their grammatical case from the context of the sentence remains open (Pickering and Gambi, 2018; Felser and Arslan, 2019; Karaca et al., 2021b; Kunduz and Montrul, 2022). The inspiration for this line of research comes from the seminal eye-tracking study of Kamide et al. (2003) in which German-speaking adults rapidly used the accusative case on the NP1 (patient *den Hasen*, the_{ACC} rabbit) to predict the NP2 (agent *der Fuchs*, the_{NOM} fox) in the OVS sentences (1) before the latter appeared in spoken input:

- (1) *Den Hasen frisst gleich der Fuchs.*
the_{ACC} rabbit eats shortly the_{NOM} fox
“The fox will shortly eat the rabbit.”
- (2) *Der Hase frisst gleich den Kohl.*
the_{NOM} rabbit eats shortly the_{ACC} cabbage
“The rabbit will shortly eat the cabbage.”

In that experiment, participants viewed the pictures of four referents (rabbit, fox, cabbage, tree) as they listened to the spoken sentences (1)–(2) while their eye movements were recorded. Kamide and colleagues found that during the adverb region (shortly) in (1), the listeners looked significantly more to the agent NP (fox) whereas in (2), they looked more to the patient NP (cabbage). Thus, the second referent was anticipated prior to the onset of its name in the spoken input. This shows that speakers can process case-marking cues predictively to incrementally anticipate the upcoming words.

Recently, Özge et al. (2019, 2022) employed the same Visual World design developed by Kamide et al. (2003) and expanded the scope of their investigations to monolingual German- and Turkish-speaking children. Specifically, their research focus in Özge et al. (2019) was on Turkish-speaking children; adult participants as a control group. The study entailed two experimental conditions, with Experiment 1 involving the presentation of sentences in the verb-middle order and Experiment 2, sentences in the verb-final order. The initial finding of the study suggests that children can predictively use the case in their respective languages just like the monolingual adults do, as early as at age of four. The second finding indicates that both monolingual Turkish-speaking adults and children can anticipate the thematic role of the subsequent argument using only NP1 and its case marker, regardless of verb order.

The anticipatory processing of the grammatical case on NPs in sentences with non-canonical OVS word order in bilingual speakers, be it L2 learners or heritage language speakers (HSs), is also debated (Kaan and Grüter, 2021; Soares et al., 2022). As bilinguals often have difficulties with correct interpretation of morphosyntactic information, including the grammatical case (Gor et al., 2019; Ivanova-Sullivan and Sekerina, in press), it is possible that they are less likely to use such information

predictively. The findings so far range from no evidence of the prediction (Hopp, 2015; Mitsugi and Macwhinney, 2016) to native-like prediction (Dijkgraaf et al., 2017; Ito et al., 2018b). Moreover, the type of bilingualism, i.e., L2 vs. HSs, that is reflected in differences in proficiency, manner, and timing of acquisition, affects their predictive ability (Karaca et al., 2021b). The influence of demographic and language background factors, such as literacy, age of onset, and language exposure that can affect a speaker's ability to process grammatical cues predictively, is also largely unknown.

Our study builds on Özge et al. (2019) findings that monolingual Turkish-speaking preschool children and adults have predictive abilities in thematic role assignment and test it with bilingual heritage Turkish-German adults. The study has three key purposes: (1) to conceptually replicate Özge and colleagues' hypothesis by extending it to a new population; (2) to compare whether predictive abilities in HSs can be successfully investigated in the Visual World eye-tracking Paradigm (VWP) remotely using a web-based camera on a participant's laptop (Slim and Hartsuiker, 2021; Vos et al., 2021); and (3) explore individual differences in predictive abilities of HSs.

As far as the first purpose is concerned, our study could be thought of as a conceptual replication of Experiment 2 with case-marking cues on NP1 and verb-final order from Özge et al. (2019) because we test the same hypothesis and use experimental design, materials and measures reproduced from Özge and colleagues (Marsden et al., 2018; Grieve, 2021). At this point, the psycholinguistic research community considers testing the generalizability of the prediction hypothesis essential for the theories of psycholinguistics and language acquisition (Huettig and Mani, 2016; DeLong et al., 2017). Because our participants all started as child HSs in families where Turkish was spoken as a home language in Germany, we expected them to be quite similar to monolingual Turkish-speaking children. Later, at school start, they switched to German, the societal language. That is, the school entry is also the start of speaking German mostly in everyday communication (e.g., at school, in the society and public). While Turkish remains a part of everyday communication, it is limited to certain social groups, such as family and friends, who are also Turkish speakers. Many HSs preserve high proficiency, and strong Turkish identity that are characteristic of Turkish HSs residing in Germany (Küppers et al., 2015; Bayram and Wright, 2018). Thus, because the grammatical case acquisition in L1 Turkish is completed way before the age of four (Aksu-Koç and Slobin, 2017), one could expect that Turkish HSs in our sample should anticipate the thematic role of NP2 as soon as they hear NP1, just like monolingual children do.

However, our replication is only conceptual because adult Turkish HSs constitute a new population. In case we find that their predictive abilities differ from those of monolingual children, there may be a number of alternative explanations, including the fact that heritage language grammars can undergo restructuring and/or that HSs can show attrition in their HL with passing time. In heritage language bilingualism, different areas of grammar (Polinsky, 2018), such as syntax and morphology, seem to present difficulties for HSs (Sorace, 2011). This is embedded in an extension of the *Interface Hypothesis* which predicts preservation at the internal interfaces/core grammar (e.g., between morphology and syntax) and problems at the external interfaces (e.g., between syntax and

pragmatics; Sorace and Serratrice, 2009). As case-marking is a core grammatical feature of Turkish, we expect to find predictive processing in bilingual heritage Turkish-German speakers too.

Our second key purpose is to examine whether predictive abilities in HSs can be reliably tested without having access to a high-end in-lab expensive equipment, such as stationary eye-trackers (i.e., *EyeLink*, *Tobii*, and *SMI*). If this is the case, a simple set-up with a webcam-based laptop connected to the Internet will allow us to record eye movements online. It opens up a possibility to vastly expand our current modest efforts to investigate heritage languages that are understudied (or not studied at all). To achieve this purpose, we conducted our eye-tracking experiment twice: first, using a in-lab high-end *Tobii Fusion 120 Hz* eye-tracker (Experiment 3) and then replicating it with *PCIBex*, an open full service platform for online behavioral experiments (Schwarz and Zehr, 2021; Experiments 1 and 2).

Finally, individual variation in demographic and language history background is an important modulator of HSs' ability to process the grammatical case predictively. Parental input, language use, literacy levels, and processing strategies can affect HSs' language processing all the way down to neural signatures in the brain (Soares et al., 2022). Individual variation in HSs is a relatively novel line of research in heritage language bilingualism. Regarding the predictive use of case, it is possible that HSs with higher proficiency and frequent language use of Turkish show this effect while less proficient speakers do not. Therefore, individual variation is part and parcel of the present study as it suggests an alternative (or extension) to the commonly used approach of looking at the participants through the lens of group means.

2. Background

2.1. Conceptual replication: processing of grammatical case in heritage Turkish

Turkish is a language with very flexible word order even though (S)OV sentences are most common (Göksel and Kerslake, 2004). The present study started as a replication of Experiment 2 by Özge et al. (2019) that compared the predictive abilities of monolingual Turkish-speaking adults and 4-year-old children in verb-final SOV and OSV sentences with overt case-marked subject and direct object. The reason why Özge and colleagues used verb-final sentences was to see whether children could predict the thematic role on the NP2 from just the grammatical case on the NP1, without any additional information from the verb. Indeed, the authors demonstrated that children, like adults, made use of the grammatical case on the NP1 and successfully inferred the thematic role of the NP2. Thus, the case-marking alone, regardless of verb order, could be sufficient for prediction of the upcoming arguments in Turkish. We expect to replicate this finding in our monolingual Turkish-speaking adults using the web-based camera eye-tracking (Experiment 1).

Testing the prediction effect in monolingual Turkish and HSs is important because of the special cross-linguistic contribution that Turkish can make to investigations of predictive language processing. The previous studies of verb-medial languages with the strict SVO word order demonstrated that early grammatical cues

from the verb that is located between the NPs produce a strong anticipatory effect on subsequent argument processing (Mani and Huettig, 2012; Gambi et al., 2016). But what happens when some of the cues are late, such as when the verb is in the sentence-final position? There is some evidence that comes from Dutch (Brouwer et al., 2019) and German (Özge et al., 2022), but these languages exhibit less flexible word order, limited case marking, and obligatory overt arguments. Turkish allows us to disentangle the timing effects of the cues that come later, i.e., when the case-markings are at the end of the nouns and the verb is sentence-final.

For Turkish, recent work by Karaca et al. (2022) has been exploring the timing of the cues with HL Turkish-Dutch adults. The preliminary results reveal that HSs process the grammatical case predictively only when lexical and grammatical cues appear early and together, which happens in verb-medial sentences. In contrast, they found no prediction in verb-final sentences. It is possible that it could be due to the difference in the types of cue, in that lexical and semantic cues are stronger whereas grammatical (or morphosyntactic) ones are weaker. In our study, we utilized both types, namely, the early cue in the form of the grammatical case on NP1 and the late lexical cue on the verb in the sentence-final position.

2.2. Methodological advancement: comparing in-lab and webcam-based web-based eye-tracking

The few published VWP studies with HSs have employed the stationary high-end in-lab eye-trackers, such as *Tobii* (Karaca et al., 2022), *SMI* (Fuchs, 2019), and *EyeLink* (Sekerina and Sauermann, 2015; Jegerski and Sekerina, 2020; Fuchs, 2022). These eye-tracking studies have reliably measured the timing of cue processing of different phenomena in heritage languages in real-time. However, the progress in studying predictive processing in HLs is slow because stationary in-lab eye-trackers are expensive, require an experienced researcher to control the experiment, and have a long learning curve, which makes them less accessible for researchers in heritage language bilingualism. But every cloud has a silver lining; the recent COVID-19 pandemic has precipitated a potential solution to the prohibitive costs of in-lab eye-tracking, namely, switching to webcam-based eye-tracking with web-based cameras that these days come on most desktop and laptop computers.

The first methodological study assessing the pros and cons of webcam-based eye-tracking in cognitive research was published by Semmelmann and Weigelt (2018). Extending an experimental design used in the in-lab environment to a JavaScript-based eye-tracking algorithm implemented in online environment allowed the authors to compare the accuracy of the two methods in three different tasks: simple fixation, pursuit, and free viewing. Semmelmann and Weigelt, however, reported a greater rate of temporal error when eye movements were collected remotely via participants' web-based cameras on their personal computers because specifications, such as *frames per second (fps)* rates and inter-sampling interval, varied much more than in the stationary in-lab setting.

Recently, the first psycholinguistic VWP experiments conducted remotely using the web-based cameras on participants' computers have appeared. Vos et al. (2021) assessed the predictive processing of verb aspect (simple past vs. progressive) in English-speaking adults. Using WebGazer.js (Papoutsaki, 2015) with an average fps rate of 20.73, the authors replicated their in-lab results obtained with the SMI Red500 eye-tracker with 64 participants. The looks of 124 online participants to the picture that matched the verb aspect condition were earlier than in the mismatched condition, just like in the in-lab set-up. The authors argued that the web-based cameras were appropriate for investigating fine-grained temporal characteristics of predictive processing despite some minor issues. The latter included (a) the necessity to increase the subject power by at least 30% as 63 online participants did not pass the stringent hardware and calibration control requirements, (b) frequent re-calibration, i.e., every 12 trials, and (c) a 50-ms delay in the onset of the verb aspect effect.

In another recent study, Slim and Hartsuiker (2022) replicated the results of their in-lab VWP experiment (*EyeLink 1000*) of the effect of verb semantics on selection of a referent out of 4 referents presented in quadrants. They used the web-based eye-tracking method (average fps of 18.1) and the module for webcam-based eye-tracking from *PCIBex* (Schwarz and Zehr, 2021). The same issues as in Vos et al. (2021) occurred again, and they were even more substantial. To obtain a sample size of 90 participants, the authors had to (a) recruit 360 people on *Prolific*, (b) were only able to keep participants who obtained a higher calibration score of 50, and (c) found a consistent time lag of 300 ms on average in comparison to the original in-lab timing of the effect of verb semantics.

These studies clearly demonstrate that while web-based eye-tracking delivers good approximation of the location of fixations, it still not sensitive enough to accurately record the timing of eye movements. This is because the typical sampling rate of the consumer-grade web-based cameras, i.e., 24, 30, and 60 fps, is not sufficient to measure rapid eye movements, as opposed to stationary high-end (also known as *infrared*) eye-trackers, which range from 30 to 1,200 Hz (Dalmaijer, 2014; Vos et al., 2022).

Despite the drawbacks of the web-based eye-tracking, its flexibility, low cost, and scalability still present indisputable advantages for research in heritage language bilingualism. Our study is a first rigorous comparison of the (remote) web-based eye-tracking (Experiments 1 and 2) with the stationary in-lab Tobii eye-tracker (Experiment 3) in a VWP study with HSs. We used the same design to ascertain whether the timing of grammatical and lexical cue effects would be comparable in comprehension of SOV and OSV Turkish sentences. The second novelty has to do with the fact that we studied predictive processing with (remote) webcam-based eye-tracking with Turkish HSs. We hope to show what researchers in HL bilingualism need to take into the account when adopting remote web-based eye-tracking to HSs so that it can be established as a widespread, reliable, and accessible research method. Thus, our study addresses an emergent need outlined as necessary for HL bilingualism in the future (Bayram et al., 2021).

2.3. Individual differences in predictive abilities

The traditional group mean-based approach to cue predictive processing is expected to confirm that monolingual Turkish-speaking adults can successfully use the grammatical case information on the NP1 to anticipate the thematic role of the NP2. However, because HSs are characterized by large individual variation in their demographic and language history experience in Turkish, averaging their eye-movement patterns may mask the differential predictive abilities of HSs who, we argue, fall into three types—predictors, partial predictors, and non-predictors. We define in detail how we calculated these types in Section 4.4.

The driving force behind these types is what underlies an individual's ability to process sentences predictively or not. Previous literature has suggested several factors that might modulate individual's predictive abilities. The first and most prominent one is proficiency (e.g., Mani and Huettig, 2012; Brouwer et al., 2017; Hopp and Lemmerth, 2018). Heritage speakers are a very heterogeneous group as far as language proficiency is concerned (Wiese et al., 2022). The second factor has to do with typological similarity between the relevant grammatical features in a bilingual's two languages (e.g., Dussias et al., 2013; Foucart et al., 2014). In our study, Turkish and German are similar as both use case marking, which indicates the thematic role of the arguments, i.e., agent (NOM case) or patient (ACC case). However, Turkish is much more consistent in marking the case directly as a suffix on the noun. In contrast, in German, the case-marking system is less transparent. Morphemes that mark case overlap with other grammatical categories such as number and gender. Thus, in the present study, we focused on the category of masculine nouns for NP1 in the items, as this is the grammatical gender in German where accusative and nominative case always unambiguously contrast on the article which is the element in the study designated to allow predictive processing.

Finally, a speaker's cognitive resources is also another indicator of their predictive abilities in real-time processing (Ito et al., 2018a). For example, Huettig and Janse (2016) highlighted the role of working memory in predictive processing of grammatical gender in Dutch participants. Their results showed that faster processing speed and higher working memory capacity facilitated predictive looks. While we have not assessed participants' working memory, we assume that this might be one of the driving factors behind predictive abilities and encourage further work with heritage speakers to explore these aspects.

3. Method

The current study consisted of three experiments, i.e., Experiment 1, Experiment 2, and Experiment 3. All of them share the same design but differ either in terms of the group (i.e., monolingual Turkish vs. HSs) or method (i.e., in-lab *Tobii Fusion* 120 Hz vs. webcam-based eye-tracking). Experiment 1 was conducted with monolingual Turkish adults using the web-based camera eye-tracking. Experiments 2 (webcam-based eye-tracking) and 3 (in-lab *Tobii* eye-tracker) investigated two separate groups

of bilingual HL Turkish-German adults with the same linguistic background profile. The community we have worked with is Turkish HSs who live in Berlin, Germany. It is a highly cohesive and vital speech community where Turkish is used on an everyday basis in many informal settings (Özsoy et al., 2022). However, many HSs are the third and fourth generation and they often do not use a strictly monolingual mode when speaking Turkish. More often, they prefer to engage in code-switching and rely on lexical and grammatical borrowings as German is most likely their dominant language as well as the language exclusively used in education (Küppers et al., 2015).

3.1. Participants

3.1.1. Experiment 1: monolingual Turkish adults (webcam-based eye-tracking)

Twenty-two monolingual Turkish-speaking participants (59% females, $M_{age} = 33.5$, range 19–63, $median_{age} = 25$) were recruited from Anadolu University in Eskişehir (Turkey) who participated in the webcam-based eye-tracking experiment. They all were raised monolingually, and their first encounter with another language was in primary school. The data from all 22 participants were included in the analysis.

3.1.2. Experiments 2 and 3: heritage Turkish adults (webcam-based or in-lab Tobii eye-tracking)

Forty Turkish-speaking HSs living in Berlin participated in the study¹. The first half ($n = 20$, 61% female, $M_{age} = 24.8$, range 18–33, $median_{age} = 28$) participated in Experiment 2 (webcam-based eye-tracking). The second half ($n = 20$, 66% female, $M_{age} = 26.3$, range 18–35, $median_{age} = 31$) participated in Experiment 3 (in-lab Tobii eye-tracking). All HSs were recruited from the wider network of acquaintances of the first author and from those who replied to our recruitment flyers.

The participants were all born and raised in Berlin, Germany, and acquired Turkish from birth in their family (age of onset for Turkish was zero). They were the second or third generation of Turkish immigrants, because their (grand)parents moved to Berlin as part of the worker's recruitment agreement between Germany and Turkey in 1961–1973. It is estimated that more than 5% of people in Berlin are Turkish-speaking and in some areas (e.g., Kreuzberg) Turkish can serve as a language of everyday communication in business and shops. This leads to a high level of vitality of the Turkish language among the bilingual Turkish-German speakers. However, only two of our 40 participants have received some level of formal Turkish education at school. The overwhelming majority (i.e., the remaining 38 participants) have received mostly received no education in Turkish, e.g., only 1 year in primary school for 1 hour a week, or none at all. All participants can be assumed to be dominant in German due to its relevance in education, career and overall communication with the mostly

TABLE 1 Summary of possible issues and corresponding recommendations.

	Webcam-based eye-tracking	Lab-based high-end eye-tracking
Calibration	Prone to issues because of many varying conditions such as lighting, facial features, webcam-quality • can be immensely improved by careful instructions and in-person or videocall supervision to correct participants posture mistakes or help participant to set up background conditions correctly	Usually very robust and needs minimal instructions that tell the participant to look at the moving dot and not move their head much
Accuracy	Moderate, especially improved, when only participants that calibrate (>50%) well throughout the whole experiment are kept in the sample (for example, Slim and Hartsuiker, 2022 had to exclude 240 out of 330 participants because of insufficient calibration), but importantly it is good enough for quadrant based VWP eye-tracking	High and easy to reach accuracy over 90%
Error-proneness	High (technical usage, hardware, and software variability)	Low, as the experimenter is in the room and can control the devices and surroundings
Lighting conditions	Very important since it is based on visible light, crucial that light is stable and ideally the whole face is well illuminated; avoid distracting light sources from the side or back of the head	Important and needs to be controlled too, but less sensitive since it is based on infrared light
Supervision	Strongly suggested as this improved overall calibration rate immensely (comparing Experiment 1 with supervision and Experiment 2 with only partial supervision); the experimenter can give helpful feedback to help participant calibrate well and keep posture and attention up throughout the whole experiment	Suggested and required to begin the experiment; after a few successful trials, the participant can complete the experiment on their own and the experimenter can retreat to another location in the lab

German-speaking population (the mean age of onset for German was 6 months, range 0–3 years).

To ensure comparability among the groups, we sampled speakers from the same population who live in similar environments. For example, several of the participants are colleagues at the same workplaces, with certain established language practices. Many of the participants were also recruited directly from the first author's private networks and acquaintances which ensures a certain level of control of the environment. In the recruitment process, participants were required to speak and hear Turkish at home with their families, and they all confirmed that it was the case.

For both experiments, only 27 HSs in total were included in the analysis. In Experiment 2, seven of the 20 webcam-based participants were excluded because of (a) failure to calibrate successfully until the end of the experiment ($n = 6$) and (b) low accuracy score ($n = 1$). In Experiment 3, six

¹ The testing occurred between March and June 2022 which was a time-period when COVID-regulations were still in place and special hygiene and mask regulations were part as part of the laboratory testing protocols.

of the 20 in-lab participants were removed from the analysis because of (a) technical errors during the recording ($n = 2$), (b) failure to comply with the instructions ($n = 2$), and (c) low accuracy score (below 80%, $n = 2$). In Table 1, we present a summary of common issues with webcam-based eye-tracking and corresponding recommendations. We also compare these issues to our experiences with lab-based high-end eye-tracking.

3.2. Design and materials

All three experiments used 20 experimental, 10 filler, and 2 practice items that were adapted from Özge et al. (2019)'s study (see our OSF repository for the complete set of materials). Each item consisted of two visual displays presented in sequence. The first display contained three referent objects (fox, rabbit, carrot) and was projected on the screen (Figure 1). The participant heard a spoken sentence (3) or (4) that described a transitive event that connected the two of the referents (e.g., eating, biting, etc.). After that, the second visual display appeared that depicted the event which either matched or did not match the sentence, e.g., the fox getting ready to eat the rabbit or the rabbit getting ready to eat the carrot (Figure 1). The design was 2-factorial and crossed the independent variable Word Order (SOV vs. OSV)/Case (NOM vs. ACC) as illustrated in (3) and (4); they were manipulated within-participants.

- (3) *Hızlı tavşan şuradaki havuc-u birazdan yiyecek.*
fast rabbit.NOM over-there carrot-ACC soon eat
“The fast rabbit will soon eat the carrot over there.”
- (4) *Hızlı tavşan-ı şuradaki tilki birazdan yiyecek.*
fast rabbit-ACC over-there fox.NOM soon eat
“The fox over there will soon eat the fast rabbit.”

The participants' task was picture-matching in choosing whether the depicted event in the second video display matched the sentence that they had heard by pressing the F or J keys to indicate YES or NO answers, respectively. Among the total 30 items (critical and filler), 22 required the YES-answer and 8 required the NO-answer. Their eye movements were recorded during the presentation of both displays, but only the eye-movement patterns during the viewing of the first one (Figure 1) were analyzed as only these are informative regarding predictive processing of case-marking cues.

The spoken sentences were recorded by a monolingually raised female native Turkish speaker² with a focus accent on the verb. The NP1 was followed by 300 ms prosodic break that was judged as natural by a small pilot group of five native speakers. The pictures were color drawings of the referent objects and events taken from Özge et al. (2019) with the permission of the authors. In the experimental items, referents had three possible thematic roles, namely, a topic (i.e., the expressed noun), plausible agent (i.e., instigator of an event), and plausible patient (i.e., the referent that is affected by the instigated event). The referents included animate objects as plausible agents, such as people (e.g., *grandpa*, *baby*) and animals (e.g., *bear*, *monkey*), and inanimate objects as plausible

patients (e.g., *honey*, *ice-cream*). There were nine different transitive verbs (e.g., *hit*, *eat*). Three referents were placed in the visual display (Figure 1) in a triangle, with two in the top row and the third one in the middle of the bottom row. The location of each referent was pseudorandomized, with each thematic role appearing equally in three different locations (upper right, upper left, and lower middle). The video displays and the spoken sentences were combined in the script prepared in the PCIBex. Each video display started with 750 ms of silence and ended with 1,500 ms of silence.

The 10 filler items looked like the experimental items, with three referents in the first video display, and an event in the second display which also required the picture-matching task. Each participant saw the same filler items. The fillers were composed of intransitive sentences that started with a complex head-final NP which was preceded by a modifier that was either marked in the GEN case, as in (5), or formed a complex phrase with a non-finite verb, as in (6).

- (5) *Dikkatsiz çocuğ-un balon-u birazdan patlayacak*
careless child-GEN balloon-POSS soon explode
“The careless child's balloon will soon explode.”
- (6) *Genç polisin bindiği gemi birazdan batacak.*
young police-officer enter_{NMZ} boat soon sink
“The boat that the young police officer entered will soon sink.”

Four versions of the experiment were created. Experimental items were rotated through the two conditions (Word Order/Case), with five items per condition, in a Latin Square design. Participants in each experiment were randomly assigned to one of the four versions and responded to 20 items in total, including 10 filler items.

In addition to this experimental task, there was also a participant background questionnaire. The online version was directly implemented in PCIBex and was the second to last display that participants saw. The final display was a thank you screen with contact information of the experimenter. The offline version of the questionnaire was handed out in paper form. It contained six questions about the participants' gender, place of birth, place of residence, age of onset for both their languages and cumulative years in formal education (starting from primary school onward).

3.3. Procedure

3.3.1. Experiments 1 and 2: webcam-based eye-tracking

All experiments as well as the procedures were approved by the ethics committee of the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft) with the votum #2022-02-220202. The study was implemented on the PennController for Internet-Based Experiments (PCIBex) platform (Schwarz and Zehr, 2021). PCIBex uses the WebGazer.js eye-tracking library which can track participants eye movements using standard computer webcams (Papoutsaki et al., 2016). The script of the experiment was programmed using PCIBex's own simple language in a main JavaScript document. Modifications in the script were made offline and the updated script was then uploaded into

² We thank Yagmur Baydar for recording and sharing the audio files with us.

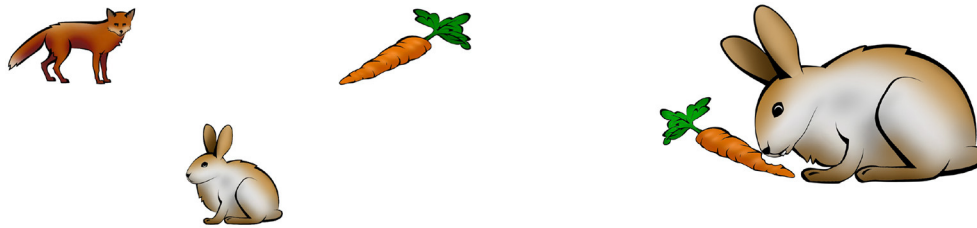


FIGURE 1

A sample 1st display with the three referents and a sample 2nd display for the picture-sentence-matching for (3).

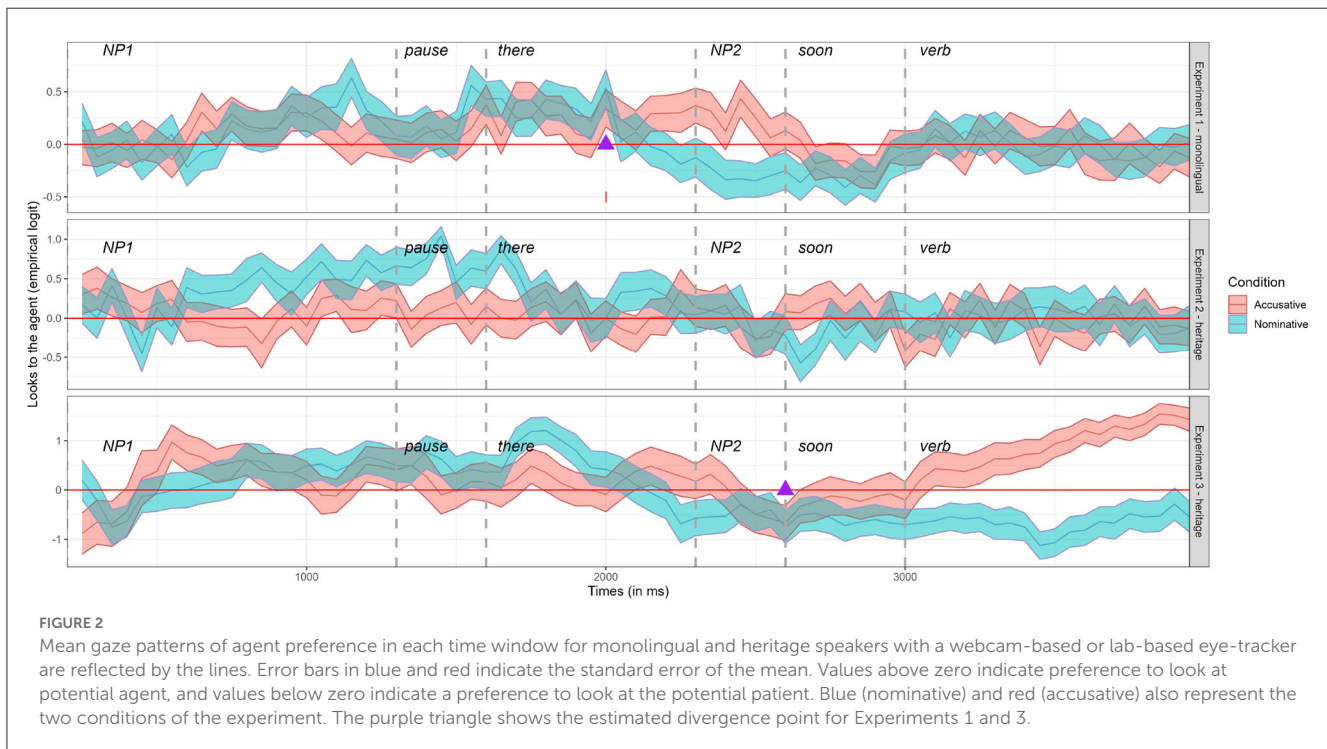


FIGURE 2

Mean gaze patterns of agent preference in each time window for monolingual and heritage speakers with a webcam-based or lab-based eye-tracker are reflected by the lines. Error bars in blue and red indicate the standard error of the mean. Values above zero indicate a preference to look at potential agent, and values below zero indicate a preference to look at the potential patient. Blue (nominative) and red (accusative) also represent the two conditions of the experiment. The purple triangle shows the estimated divergence point for Experiments 1 and 3.

the respective section of the PCIBex project overwriting the previous version. This ensured that all changes were saved and is recommended since there is a bug in PCIBex's autosave function. We uploaded all our experimental image and audio files directly into PCIBex "Ressources" section. The detailed documentation at <https://doc.pcibex.net/> outlines how different elements and whole experiments can be set up in PCIBex. The webcam video is converted into eye-tracking data in the participants browser. The eye-tracking data run through a PHP script that renders them into a standard data spreadsheet. This script needs to be stored externally and it also stored the resulting eye-tracking data there which is why it requires write-access on the server. In line with the European General Data Protection Regulation, we used our own server at the Humboldt-Universität zu Berlin for this purpose. Once the script was ready, PCIBex generated a web link that we provided to the participants (see the demo of [the full script of the experiment](#)). A 12.66" Dell Inspiron 7,400 laptop with a 30-fps web camera and the Internet connection was used. During Experiment 1 (monolingual Turkish speakers), the laptop was housed in a soundproof eye-tracking laboratory of Anadolu University in Eskisehir (Turkey). The appropriate

lighting, noise-proof environment, and reliable Internet connection in this lab were ideal for eye-tracking experiments. All 22 monolingual participants were tested on this laptop in the same location, with the experimenter present. Experiment 2 (Turkish HSs) was conducted in the field in Berlin, and the conditions varied much more due to changing testing environments. Twenty HSs participated at their homes or their workplaces in a quiet location. Among them, four HSs completed the experiment on their own personal computers and 16 HSs completed the experiment from the experimenter's Dell Inspiron 7,400 that was used with the monolingual speakers in Experiment 1. Variability in hardware and field conditions explains why the data from only 13 HSs were usable and included in the analysis.

At the beginning of the experiment, participants read the description of the experiment, electronically signed the consent form, and filled out the online demographic and language background questionnaire. Then they were asked to self-calibrate by following the instruction on the screen. Calibration was better when the experimenter was present and could assist participants by adjusting the laptop screen to the appropriate angle and optimizing the lighting conditions and background colors. Still,

some participants failed to calibrate either due to one of the aforementioned variables or due to other factors such as facial features or webcam quality. Following successful calibration, participants started the experiment with two practice trials followed by 20 experimental trials interspersed with 10 fillers. It took participants on average 10 min to complete the task itself.

3.3.2. Experiment 3: in-lab stationary Tobii eye-tracking

The experiment was conducted in the psycholinguistics laboratory of the Leibniz-ZAS in Berlin (Germany). Individual 30-min appointments were scheduled with each participant based on their availability. The participant was seated in front of the stimuli computer of the high-end stationary *Tobii Pro Fusion* 120 Hz eye-tracker. Calibration was controlled by the *Tobii Pro Lab* software and was validated by the experimenter. When participants looked away from the stimuli computer (e.g., toward the experimenter when asking questions), re-calibration was performed. Just like in Experiments 1 and 2, following the calibration, participants started the experiment with two practice trials followed by 20 experimental trials interspersed with 10 fillers. Participants completed the task itself in around 5 min which is faster than in Experiments 1 and 2 because in-lab stationary eye-tracking required fewer recalibrations and adjustments of the experimental set-up.

3.4. Data analysis

A data analysis plan and accompanying predictions were registered in advance of carrying out this study on the [AsPredicted web site: https://aspredicted.org/8B7_565](https://aspredicted.org/8B7_565). In addition to the registered analysis, we also conducted a divergence point analysis by closely following the procedure and script described by Stone et al. (2021). The eye-movement data were preprocessed and analyzed using R (R Core Team, 2022). We used the following packages: *tidyverse* (Wickham et al., 2019), *lme4* (Bates et al. (2015), *boot* (Davison and Hinkley, 1997), *mgcv* (Wood, 2003), and *polycor* (Fox, 2022).

4. Results

4.1. Accuracy in the picture-matching task

Both groups of participants performed the picture-matching task at ceiling: on average, the 22 monolingual speakers' accuracy was 94% (range 80–100%), the 27 HSS' accuracy was 95% (range 90–100%). This implies that the participants were highly attentive during the experiment.

4.2. Analysis of eye movements 1: agent Preference

We start by presenting the analysis of fixation data using a binary dependent variable called *Agent Preference* that we computed following Özge et al. (2019) and Özge et al. (2022). It

included only the looks to the plausible Agent or Patient of an item. All other looks were excluded from this variable as they are not relevant for the prediction effect under investigation. Different eye-movement patterns in Agent Preference allow us to directly compare looks in the two Case conditions, ACC and NOM, to test whether in the ACC condition, there was a statistically significant increase in looks to the plausible Agent during the NP2 (2,300–2,600 ms) but before it ends. Such an increase would indicate predictive processing based on the ACC case marker on the NP1 in the OSV sentences.

The Agent preference results are shown in Figure 2. The top panel represents Experiment 1 (monolingual speakers, the webcam-based eye-tracking), the middle panel is Experiment 2 (HSSs, webcam-based eye-tracking), and the bottom panel, Experiment 3 (HSSs, in-lab Tobii eye-tracking). The region of interest for the effect of predictive case marking begins after the 300-ms prosodic break that follows the end of NP1 (1,600 ms). The region of interest ends as soon as the NP2 is encountered in speech (2,600 ms). Any looks following the region of interest are no longer purely predictive because they are based on lexical or prosodic information from the NP2.

In the top panel for Experiment 1, we see that monolingual speakers show an effect of Agent preference around 2,000 ms. After the end of NP2 this effect fades and looks to the Agent and the Patient become roughly equal again. The middle panel for Experiment 2 shows no clear pattern of Agent preference. Throughout most of the time windows, the proportion of looks to the Agent and the Patient do not diverge in a meaningful way. This could partly be due to the low resolution and variable experimental set-up in the webcam-based eye-tracking with heritage speakers, and partly due to a missing predictive processing effect. Experiment 3 (HSSs, in-lab eye-tracking, bottom panel, Figure 2) yielded the clearest Agent preference in the ACC condition due to the high resolution and better quality of the in-lab stationary eye-tracker. The HSSs' looks to the Agent (above the zero line) and the Patient (below the zero line) clearly increase in the ACC condition OSV (red line for ACC, blue for NOM) after 300 ms from the onset of the NP2, which happens before the end of the NP2 (2,600 ms).

4.3. Analysis of eye movements 2: generalized linear mixed model

The first goal of this study was a conceptual replication of Özge et al. (2019). We analyzed eye movements using binomial generalized linear mixed effects regression models (GLMMs). We limited this analysis to the region of interest between the onset of the NP1 and the offset of the NP2 where predictive looks were expected to occur. Because looks to the Agent and the Patient were equal in the critical region in Experiment 2 (HSSs, webcam-based eye-tracking) and this clearly indicating that there is no effect, we only compared the results of Experiment 1 (monolingual webcam-based eye-tracking) and Experiment 3 (HSSs, in-lab eye-tracking) to estimate meaningful effects in two regression models. Table 2 presents the first, baseline, set of models with the exact same structure (1):

TABLE 2 Experiment 1 (monolinguals, webcam-based eye-tracking) and Experiment 3 (HSs, in-lab eye-tracking): agent preference in the predictive region of interest.

	Dependent variable	
	Agent preference	(AgentPrefScore)
	Monolinguals	HSs
Condition_AvN	−0.583 (0.382)	0.190 (0.813)
No. of observations	3,167	2,802

The binary variable Condition_AvN encodes the Accusative and Nominative conditions in the experiment.

(1) `glmer(data=dat, AgentPrefScore ~ Condition_AvN + [1|Participant)+(1|Item), family=binomial, control=glmerControl(optimizer="bobyqa")]`

There was no significant effect in the Agent preference looks between the two groups of participants in this baseline model.

In the second set of models, we incorporated Time as a variable in the form of 100-ms bins that were used to split the region of interest between the onset of the NP1 and the offset of NP2 (This region spanning 2,000–2,600 ms is set out in lavender in Table 3). Table 3 presents the results for these omnibus models. We added TimeWindows as an independent variable and tested the interaction between TimeWindows and Condition on Agent preference (2):

(2) `glmer[data=dat, AgentPrefScore ~ Condition_AvN * TimeWindows + [1|Participant)+(1|Item), family=binomial, control=glmerControl(optimizer="bobyqa")]`

The results showed that there were several significant relationships between TimeWindows and Agent preference between Experiment 1 (monolingual, webcam-based eye-tracking) and Experiment 3 (HSs, in-lab eye-tracking) (marked in boldface in Table 3). In Experiment 1, there were significant negative relationships between four Time bins (i.e., Time2600, Time2700, Time2800, and Time2900) and Agent preference, such that 1-unit increase in the Time bin windows was associated with a 0.553-, 1.002-, 1.035-, and 0.952-unit decrease in Agent preference, respectively (all $p < 0.01$). Significant interactions between Condition and TimeWindows were found at Time2200, Time2300, Time2400, and Time2500, with a negative relationship with Agent Preference corresponding to decreases ranging between 0.628 and 0.784 units.

In Experiment 3 (HSs, in-lab eye-tracking), there were also significant negative relationships between almost the same four Time bins (i.e., Time2500, Time2600, Time2800, and Time2900) and Agent preference, such that 1-unit increase in the Time bins was associated with a 1.014-, 0.876-, 0.607-, and 0.545-unit decrease in Agent preference, respectively (all $ps < 0.05$). Also, significant interactions between Condition and TimeWindows were found at Time2100, Time2200, Time2300, Time2400, that continued at Time2600, Time2700, Time2800, and Time2900, with a negative relationship to Agent preference corresponding to decreases ranging between 0.674 and 1.769 units.

As far as our first goal was concerned, Experiment 1 replicated Özge et al. (2019)'s findings where monolingual Turkish-speaking adults and children as young as 4 years of age made use of the

TABLE 3 Two omnibus regression models, one for monolinguals and one for heritage speakers, in a table with the interaction of condition and time (in 100 ms bins).

	Dependent variable	
	AgentPrefScore	
	Monolinguals	Heritage
Condition_AvN	−0.151 (0.475)	0.644 (0.595)
Time2100	−0.103 (0.263)	0.299 (0.286)
Time2200	−0.030 (0.259)	0.188 (0.292)
Time2300	0.064 (0.259)	0.208 (0.306)
Time2400	−0.121 (0.259)	−0.318 (0.300)
Time2500	−0.280 (0.257)	−1.014*** (0.289)
Time2600	−0.553** (0.251)	−0.876*** (0.294)
Time2700	−1.002*** (0.246)	−0.443 (0.291)
Time2800	−1.035*** (0.247)	−0.607** (0.283)
Time2900	−0.952*** (0.246)	−0.545* (0.289)
Condition_AvN:Time2100	−0.262 (0.380)	−0.776* (0.398)
Condition_AvN:Time2200	−0.673* (0.374)	−1.491*** (0.419)
Condition_AvN:Time2300	−0.784** (0.380)	−1.769*** (0.424)
Condition_AvN:Time2400	−0.763** (0.375)	−1.204*** (0.414)
Condition_AvN:Time2500	−0.628* (0.367)	−0.347 (0.400)
Condition_AvN:Time2600	−0.357 (0.367)	−0.674* (0.404)
Condition_AvN:Time2700	0.120 (0.362)	−0.883** (0.398)
Condition_AvN:Time2800	0.093 (0.363)	−1.015*** (0.394)
Condition_AvN:Time2900	0.077 (0.359)	−1.024*** (0.396)
Observations	3,167	2,802

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. Bold values indicate significant effects.

grammatical case on the NP1 and successfully inferred the thematic role of the NP2. We also extended these findings to adult Turkish

HSS (our first goal) in Experiment 3. In heritage speakers and monolinguals (Experiment 1 and 3), the significant interactions at the Time2100 and Time2200 windows point to the Case effect before the onset of the NP2 at 2300 ms. Thus, the case-marking alone, regardless of verb position, was sufficient for prediction of the upcoming arguments in monolingual Turkish adults and children, and in HSS.

For our second goal, webcam vs. in-lab replication, have two observations. Firstly, our results from Experiment 1 represent a conceptual replication of Özge et al. (2019), although the low resolution webcam-eye-tracking data shows a smaller effect of predictive use of case. Concerning the differences for HSS between Experiments 2 and 3, we do not find any significant effects in the HSS webcam-based data. The reason most likely lies in the small sample size and the variable experimental set-up for Experiment 2. When the experimental set-up is more stable as in Experiment 1, webcam-based eye-tracking is more feasible and can replicate previous findings.

To establish the precise point where the looks to the Agent diverge from the looks to the Patient in the OSV condition compared to those in the SOV condition in the two groups of participants, we followed Stone et al. (2021) and conducted divergence point estimation using corrected and uncorrected multiple comparisons. Surprisingly, the results in this analysis differed from the GLMM analysis. In this case, the group in Experiment 1, i.e., monolinguals showed a significant divergence point that indicated predictive use of case-marking cues. The group in Experiment 3, i.e., heritage speakers, showed a divergence point that lied behind the prediction region. We present this analysis in detail in Appendix and discuss differences in methodologies.

4.4. Individual variation: predictor categories

Our third goal was to investigate individual ability of HSS to process the grammatical case in the OSV sentences predictively. In line with previous psycholinguistic research (Hopp and Lemmerth, 2018; Brouwer et al., 2019; Karaca et al., 2021a), we hypothesized that HSS' participant background variables might have an effect on whether they can engage in predictive processing. Based on individual speakers growth-curve figures, we found three types of predictive processing behavior—i.e., predictor, partial predictor, and non-predictor—that are reflected in the Figure A1.

We also found a different and possibly more reliable way, to characterize individual speakers into predictors vs. non-predictors. Since the 2,200–2,300 ms time window is crucial for any predictive looks before the onset of the second NP at 2,300 ms, we focused on this time window. We then calculated the mean AgentPreferenceScore, i.e., whether the participant looked more to the potential agent or the possible patient, in this time window and this was limited to the Accusative condition as it included the first case-marked NP which could serve as a cue. If a person had a score above 0.5, we classified them as *predictors* since they looked to the agent were above chance. If a person had a score below 0.5, this indicated that they were not looking at the agent predictively, so we classified them as *non-predictors*. Table 4 shows the results of this classification. Most of the participants who

we classified as predictors are actually well above chance ranging from 75 to 100% which is a clear indicator that they process case predictively. In the monolingual as well as the heritage samples, we see participants who are classified as non-predictors and predictors. Compared to the group size, the proportion of non-predictors in the heritage speaker group is higher than in the monolingual group. However, in both groups there are also more predictors than non-predictors. Additionally, there is a limited number of partial predictors who seem to use case predictively at a chance level of 50% according to this threefold classification. This simple classification sheds light at structured individual variation that must be based on speakers' individual background factors. Future studies should carefully investigate especially sociolinguistic background variables (e.g., proficiency, language exposure) to be able to determine what drives predictive abilities in speakers.

5. Discussion

Our findings in multiple groups and using lab-based as well as webcam-based eye-tracking reveal several new insights regarding the predictive processing of case in Turkish heritage and monolingual speakers. Overall, we replicate Özge et al. (2019)'s findings for monolinguals using the webcam-based eye-tracking method. Our monolingual group was able to process case predictively before the onset of NP2. Our analysis located the divergence point at 2,000 ms which is 300 ms before the crucial onset and indicates the use of predictive processing. In contrast, for the heritage speaker group in the lab, our analysis located this divergence point at 2,600 ms which is 300 ms after the onset of NP2. Hence, we observe that heritage speakers, on a group-level, do not process case predictively.

However, our aim was to look further into this issue with more detailed analyses. The first step toward this came from conceptually replicating the same GLMM analysis as in Özge et al. (2019). We observed interactions between Condition and the predictive Time Windows of 2,100–2,300 ms in both groups. Contrary to the results from the divergence point estimates above, this indicates that there is predictive use of case in both groups and not just the monolingual group. This divergence in results points to the relevance of using appropriate methods when analyzing such large data sets across different groups (Vasishth, 2022). Relatedly, recently discussions about accepting uncertainty in experimental studies have emerged (e.g., Vasishth and Gelman, 2021) which we acknowledge by listing some limitations below. It also indicates that group-level analyses might not be the best option for such effects that may be guided by individual speakers abilities and backgrounds which is in line with recent proposals in heritage language research to consider the speaker more (Luk, 2022; Rothman et al., 2022).

To understand this individual variation better, we explored two new ways to categorize the predictive processing of case on a speaker-level. The first one is based on growth curves and the divergence between the conditions at three different time windows of interest as seen in the Figure A1. Based on this, we classified the use of predictive processing in heritage speakers lab-based data. The classification revealed that the majority of participants showed some predictive processing of case before the onset of NP2. However, this method was vague and harder to quantify. Therefore, we used our knowledge about the interaction effects from the

TABLE 4 A categorization of monolingual and heritage speakers into *predictors*, *partial predictors*, and *non-predictors*.

	Participant	Condition	Group	Mean	Type
1	1660552776	Accusative	MSwebcam	0.00	Non-predictor
2	1660555491	Accusative	MSwebcam	0.57	Predictor
3	1660562087	Accusative	MSwebcam	0.71	Predictor
4	1660565335	Accusative	MSwebcam	0.50	Partial-predictor
5	1660566543	Accusative	MSwebcam	0.25	Non-predictor
6	1660568100	Accusative	MSwebcam	0.57	Predictor
7	1660570668	Accusative	MSwebcam	0.50	Partial-predictor
8	1660571549	Accusative	MSwebcam	0.69	Predictor
9	1660572917	Accusative	MSwebcam	0.80	Predictor
10	1660573937	Accusative	MSwebcam	0.73	Predictor
11	1660579497	Accusative	MSwebcam	0.00	Non-predictor
12	1660651886	Accusative	MSwebcam	0.50	Partial-predictor
13	1660652919	Accusative	MSwebcam	0.50	Partial-predictor
14	1660653654	Accusative	MSwebcam	0.83	Predictor
15	1660655013	Accusative	MSwebcam	0.71	Predictor
16	1660655834	Accusative	MSwebcam	1.00	Predictor
17	1660656355	Accusative	MSwebcam	0.71	Predictor
18	1660732813	Accusative	MSwebcam	1.00	Predictor
19	1660734167	Accusative	MSwebcam	0.67	Predictor
20	1660826578	Accusative	MSwebcam	0.25	Non-predictor
21	1660830489	Accusative	MSwebcam	0.73	Predictor
22	1660831171	Accusative	MSwebcam	0.57	Predictor
23	2022-HT01T-A1	Accusative	HSlab	0.31	Non-predictor
24	2022-HT05T-A1	Accusative	HSlab	0.00	Non-predictor
25	2022-HT06T-A2	Accusative	HSlab	0.00	Non-predictor
26	2022-HT07T-B1	Accusative	HSlab	0.50	Partial-predictor
27	2022-HT08T-B2	Accusative	HSlab	1.00	Predictor
28	2022-HT09T-B2	Accusative	HSlab	0.79	Predictor
29	2022-HT10T-A1	Accusative	HSlab	1.00	Predictor
30	2022-HT11T-A2	Accusative	HSlab	1.00	Predictor
31	2022-HT13T-A1	Accusative	HSlab	1.00	Predictor
32	2022-HT14T-A2	Accusative	HSlab	0.60	Predictor
33	2022-HT15T-B1	Accusative	HSlab	0.20	Non-predictor
34	2022-HT16T-B2	Accusative	HSlab	1.00	Predictor
35	2022-HT17T-A1	Accusative	HSlab	1.00	Predictor
36	1646749940	Accusative	HSwebcam	1.00	Predictor
37	1647091947	Accusative	HSwebcam	0.75	Predictor
38	1647605908	Accusative	HSwebcam	1.00	Predictor
39	1647696601	Accusative	HSwebcam	1.00	Predictor
40	1647697953	Accusative	HSwebcam	0.00	Non-predictor
41	1649074797	Accusative	HSwebcam	0.67	Predictor

(Continued)

TABLE 4 (Continued)

	Participant	Condition	Group	Mean	Type
42	1651350251	Accusative	HSwebcam	0.33	Non-predictor
43	1651351209	Accusative	HSwebcam	0.83	Predictor
44	1651416084	Accusative	HSwebcam	0.33	Non-predictor
45	1652102291	Accusative	HSwebcam	0.00	Non-predictor
46	1652103834	Accusative	HSwebcam	0.67	Predictor
47	1652787358	Accusative	HSwebcam	0.43	Non-predictor
48	1653463987	Accusative	HSwebcam	0.25	Non-predictor

GLMM which informed us that there were predictive processing effects in the crucial last 2,200–2,300 ms time window before the onset of the second NP. Our analysis of AgentPreference looks on a speaker-level again showed us that most speakers in the heritage as well as the monolingual group used case predictively.

These detailed analyses allow us to add to group-level analyses to better understand how patterns of predictive processing of case are distributed among different speaker groups. Proportionally, more monolingual than heritage speakers process case predictively. The underlying factors of these results are most likely guided by individual cognitive capacities and other (linguistic) background variables. These expectations are based on previous literature that has shown an effect of these factors on predictive abilities such as Karaca et al. (2021a) who have shown effects of processing speed and language proficiency. To explore these factors further will be an important next step in predictive processing research in the future.

5.1. Do heritage speakers process grammatical case predictively?

The present study aimed to investigate the use of predictive case-marking in Turkish-German heritage speakers (HSs) using both in-lab and webcam-based eye-tracking methods, and to explore individual variation among HSs in their use of this grammatical feature. Our results showed that HSs were able to use morphosyntactic cues to predict the thematic role of NP2, supporting the idea that core grammatical features of languages remain robust in HSs. However, a by-participant analysis revealed individual variation in the use of predictive case-marking, with some speakers showing patterns similar to monolinguals and others showing divergent behavior.

These findings have several implications for our understanding of heritage language acquisition and processing. First, they support the view that HSs should be placed on a native-speaker continuum rather than being treated as a homogenous group. Previous research has demonstrated that HSs can show a range of proficiency levels in their heritage language, with some exhibiting near-native abilities and others exhibiting more limited proficiency (e.g., Bayram et al., 2021). Our results suggest that this individual variation may extend to the use of predictive case-marking, with some HSs exhibiting patterns similar to monolinguals and others showing differences. This highlights the importance of considering

individual differences when studying heritage language acquisition and processing.

Our results support the idea that core grammatical features of languages, such as case-marking, remain robust in HSs. This is in line with the Interface Hypothesis, which proposes that certain aspects of grammar, such as argument structure and the expression of agreement, are resistant to interference and erosion in bilingual speakers (e.g., Sorace, 2011). This suggests that heritage speakers may have a strong foundation in their heritage language, even if they are not fully proficient in it.

5.2. Is it possible to replicate in-lab findings with web-based eye-tracking?

Our study adds to the small but growing body of research on the use of webcam-based eye-tracking methods in psycholinguistic research. Webcam-based eye-tracking allows researchers to collect data from participants in their own naturalistic environments, rather than requiring them to come to a laboratory setting. It allowed us to recruit some heritage speakers, who may not have easy access to a laboratory or may be geographically dispersed. Additionally, we were able to collect data in Türkiye without requiring expensive high-end eye-tracking equipment. Our results from the monolingual group showed that the in-lab and webcam-based eye-tracking data were largely consistent, indicating that webcam-based eye-tracking may also be a viable method for studying heritage language processing. However, further research is needed to fully understand the potential effects, benefits and challenges of webcam-based data collection in psycholinguistic studies. Many of the aspects that have also been found to be crucial in the two other psycholinguistic webcam-based eye-tracking studies by Slim and Hartsuiker (2022) and Vos et al. (2022) turned out to be relevant for the present study too. In particular, researchers need to be aware of the critical conditions that affect data quality when applying webcam-based eye-tracking. To get the most out of this technology, an ideal lab-like setup with good lighting conditions, an undisturbed environment and a stable/consistent internet connection are minimum requirements. Additionally, participants should be closely guided any possibly monitored throughout the process of calibration and later stages of completing the experiment.

Furthermore, Steffan et al. (2023) in a much more large-scaled study have shown that the sampling rate varies between

participants due to different hardware conditions. Partially, these differences also stem from the different underlying techniques between webcam- and high-end lab eye-tracking. The former predicts the gaze based on the whole face focusing on the eye using visible light, and the latter tracks the movement of the eye focusing on the pupil using infrared light (Papoutsaki et al., 2016).

5.3. Can we better account for individual differences in eye movements?

In an attempt to move the field of heritage language research forward, much recent and some earlier discussion arose about moving away from dichotomous approaches to heritage grammars such as monolingual vs heritage, native vs nonnative, complete vs incomplete, baseline vs. divergence (Cabo and Rothman, 2012; Putnam and Sánchez, 2013; Rothman et al., 2022; Wiese et al., 2022). In line with this current stream, we observe that it comes short to just classify the monolingual group in Experiment 1 as showing a predictive effect, and the heritage groups in Experiments 2 and 3 as not showing it. We explored more nuanced ways to classify not just between groups but rather between speakers. This allowed us to see that we find different types of predictors in all our experimental groups: non-predictors, partial predictors and predictors. More extensive by-participant variables such as language proficiency or working memory scores would provide a better testing ground to be able to determine what influences individual's abilities to predictively process case, and hence be categorized into one of the three predictor types.

Future research can pick up this idea of more nuanced classifications that can also move in a gradient direction. For example, Kutlu et al. (2022) in this same special issue, demonstrates how a traditionally discretely categorized phenomenon such as speech perception can become more gradient to address bilingual speakers speech using possible more adequate methods and tools. In a similar way, we find that empirically more interesting patterns emerge when we move beyond the dichotomy of mono- vs. bilinguals and instead address the gradiency within these groups. We can now ask what common background variables characters mono- and bilingual speakers who use case predictively to different degrees. Having and including more extensive information about speakers like known parameters such as working memory, proficiency and literacy (Hopp, 2015; Huettig and Janse, 2016; Hopp and Lemmerth, 2018) will help us to understand and explain in different ways how bilingual sentence processing works in the mind.

5.4. Limitations

There are several limitations to consider in the present study. First, four participants completed the experiment using their own PCs in different locations, which may have introduced variations in monitor settings that could have affected the results of the webcam-based experiments. Additionally, internet connection quality may have varied across the different locations where participants completed the webcam-based experiments, which could also have influenced the results.

Second, we were not able to collect as much data as we had originally planned, and some data had to be eliminated due to technical issues or participant errors. This may have limited the power of our statistical analyses and could have introduced bias in the results.

Third, we did not collect in-lab eye-tracking data from monolinguals in this study, which means that it is not possible to directly compare the performance of monolinguals and HSs in the same experimental conditions. This is an important direction for future research, as it would provide more insight into the relationship between heritage language proficiency and the use of predictive case-marking.

Finally, it is important to note that our sample was relatively small and may not be representative of all Turkish-German HSs who represent an extremely diverse group (Küppers et al., 2015). Keeping in mind previous literature that has shown that webcam-based eye-tracking requires much larger sample sizes than in-lab eye-tracking, our study should be viewed as a starting point in using this method whose capacity to generalize is limited at this point. Further research with larger and more diverse samples would be needed to confirm and extend the findings of the present study. Alternatively, instead of recruiting different groups of people for each experiment, split-half procedures could have been used on each group (i.e., HSs and monolinguals) to minimize the individual differences in the HSs' profiles and to keep the computer settings constant for all participants. By increasing the number of items, applying split-half procedures could be an option for future studies with more accessible populations such as "monolingual" German speakers. Because heritage speakers of Turkish are relatively difficult to recruit due to a smaller community size among other factors, and because webcam-based eye-tracking requires much larger sample sizes to be exactly comparable to lab-based eye-tracking (Slim and Hartsuiker, 2022), this procedure was not feasible for this study.

6. Final remarks

In conclusion, the present study provides new insights into the use of predictive case-marking in Turkish-German HSs and the importance of considering individual differences in the study of heritage language acquisition and processing. Our results support the idea that core grammatical features of languages remain robust in HSs and suggest that webcam-based eye-tracking may be a useful method for studying heritage language processing. Future research could further explore the relationship between proficiency in the heritage language and the use of predictive case-marking in HSs, as well as the potential effects of webcam-based data collection on the results of eye-tracking studies.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: <https://osf.io/sehnf/>.

Ethics statement

The studies involving human participants were reviewed and approved by Ethikkommission der Deutschen Gesellschaft für Sprachwissenschaft (Ethics Committee of the German Linguistics Association). The patients/participants provided their written informed consent to participate in this study.

Author contributions

OÖ, IS, and NG contributed to conception and design of the study. OÖ and BÇ organized the database. OÖ performed the statistical analysis. OÖ, IS, BÇ, and ZÖ wrote the first draft of the manuscript. OÖ, IS, BÇ, ZÖ, and NG provided feedback on further versions and wrote sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This research was funded by the Deutsche Forschungsgemeinschaft grant 313607803 to GA 1424/10-1/. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

References

- Aksu-Koç, A., and Slobin, D. (2017). "The acquisition of Turkish," in *The Crosslinguistic Study of Language Acquisition*, ed Dan Slobin (London: Psychology Press), 839–878. doi: 10.4324/9781315802541-10
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bayram, F., Pisa, G., Rothman, J., and Slabakova, R. (2021). "Current trends and emerging methodologies in charting heritage language grammars," in *The Cambridge Handbook of Heritage Languages and Linguistics (Cambridge Handbooks in Language and Linguistics)*, eds S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 545–578. doi: 10.1017/9781108766340.025
- Bayram, F., Wright, C. (2018). "Turkish heritage language acquisition and maintenance in Germany," in *Handbook of Research and Practice in Heritage Language Education. Springer International Handbooks of Education*, eds P. Trifonas, and T. Aravossitas (Cham: Springer). doi: 10.1007/978-3-319-44694-3_49
- Brouwer, S., Özkan, D., and Küntay, A. C. (2019). Verb-based prediction during language processing: the case of Dutch and Turkish. *J. Child Lang.* 46, 80–97. doi: 10.1017/S0305000918000375
- Brouwer, S., Sprenger, S., and Unsworth, S. (2017). Processing grammatical gender in Dutch: evidence from eye movements. *J. Exp. Child Psychol.* 159, 50–65. doi: 10.1016/j.jecp.2017.01.007
- Cabo, D. P. Y., and Rothman, J. (2012). The (il)logical problem of heritage speaker bilingualism and incomplete acquisition. *Appl. Linguist.* 33, 450–455. doi: 10.1093/applin/ams037
- Dalmajier, E. (2014). *Is the Low-Cost Eyetracker Eye Tracker Any Good for Research?* Technical report, PeerJ. doi: 10.7287/peerj.preprints.585v1
- Davison, A. C., and Hinkley, D. V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511802843
- DeLong, K. A., Urbach, T. P., and Kutas, M. (2017). Is there a replication crisis? Perhaps. Is this an example? No: a commentary on Ito, Martin, and Nieuwland (2016). *Lang. Cogn. Neurosci.* 32, 966–973. doi: 10.1080/23273798.2017.1279339
- Dijkgraaf, A., Hartsuiker, R. J., and Duyck, W. (2017). Predicting upcoming information in native-language and non-native-language auditory word recognition. *Bilingualism* 20, 917–930. doi: 10.1017/S1366728916000547
- Dussias, P. E., Kroff, J. R. V., Tamargo, R. E. G., and Gerfen, C. (2013). When gender and looking go hand in hand: grammatical gender processing in L2 Spanish. *Stud. Sec. Lang. Acquis.* 35, 353–387. doi: 10.1017/S0272263112000915
- Felser, C., and Arslan, S. (2019). Inappropriate choice of definites in Turkish heritage speakers of German. *Heritage Lang. J.* 16, 22–43. doi: 10.46538/hlj.16.1.2
- Foucort, A., Martin, C. D., Moreno, E. M., and Costa, A. (2014). Can bilinguals see it coming? Word anticipation in L2 sentence reading. *J. Exp. Psychol.* 40:1461. doi: 10.1037/a0036756
- Fox, J. (2022). *Polycor: Polychoric and Polyserial Correlations. R Package Version 0.8-1*. Available online at: <https://CRAN.R-project.org/package=polycor>
- Fuchs, Z. (2019). *Gender in the nominal domain: evidence from bilingualism and eye-tracking* (Ph.D. thesis). Harvard University, Cambridge, MA, United States.
- Fuchs, Z. (2022). Eyetracking evidence for heritage speakers' access to abstract syntactic agreement features in real-time processing. *Front. Psychol.* 13:960376. doi: 10.3389/fpsyg.2022.960376
- Gambi, C., Pickering, M. J., and Rabagliati, H. (2016). Beyond associations: Sensitivity to structure in pre-schoolers' linguistic predictions. *Cognition* 157, 340–351. doi: 10.1016/j.cognition.2016.10.003
- Göksel, A., and Kerslake, C. (2004). *Turkish: A Comprehensive Grammar*. New York, NY: Routledge. doi: 10.4324/9780203340769
- Gor, K., Chrabaszc, A., and Cook, S. (2019). A case for agreement: processing of case inflection by early and late learners. *Linguist. Approach. Bilingual.* 9, 6–41. doi: 10.1075/lab.16017.gor
- Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics* 59, 1343–1356. doi: 10.1515/ling-2021-0094
- Hopp, H. (2015). Semantics and morphosyntax in predictive L2 sentence processing. *Int. Rev. Appl. Linguist. Lang. Teach.* 53, 277–306. doi: 10.1515/iral-2015-0014
- Hopp, H., and Lemmerth, N. (2018). Lexical and syntactic congruency in L2 predictive gender processing. *Stud. Sec. Lang. Acquis.* 40, 171–199. doi: 10.1017/S0272263116000437
- Huetting, F., and Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Lang. Cogn. Neurosci.* 31, 80–93. doi: 10.1080/23273798.2015.1047459

Acknowledgments

We are thankful to Anadolu University in Eskisehir, Türkiye and our colleague SE for opening their laboratory to us and assisting us in participant recruitment. We also thank audience at the Third Birmingham Statistics for Linguists Summer School, AMLaP 2022, and the Workshop on L2 sentence and discourse processing for their constructive feedback.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- Huetting, F., and Mani, N. (2016). Is prediction necessary to understand language? Probably not. *Lang. Cogn. Neurosci.* 31, 19–31. doi: 10.1080/23273798.2015.1072223
- Ito, A., Corley, M., and Pickering, M. J. (2018a). A cognitive load delays predictive eye movements similarly during L1 and L2 comprehension. *Bilingualism* 21, 251–264. doi: 10.1017/S1366728917000050
- Ito, A., Pickering, M. J., and Corley, M. (2018b). Investigating the time-course of phonological prediction in native and non-native speakers of English: a visual world eye-tracking study. *J. Mem. Lang.* 98, 1–11. doi: 10.1016/j.jml.2017.09.002
- Ivanova-Sullivan, T., and Sekerina, I. (in press). “The psycholinguistics of heritage languages,” in *The Cambridge Handbook of Slavic Linguistics*, eds D. Shipko and W. Browne (Cambridge: Cambridge University Press).
- Jegerski, J., and Sekerina, I. A. (2020). The processing of input with differential object marking by heritage Spanish speakers. *Bilingualism* 23, 274–282. doi: 10.1017/S1366728919000087
- Kaan, E., and Grüter, T. (2021). *Prediction in Second Language Processing and Learning*. Vol. 12. Amsterdam: John Benjamins Publishing Company. doi: 10.1075/bpa.12.01kaa
- Kamide, Y., Scheepers, C., and Altmann, G. (2003). Integration of syntactic and semantic information in predictive processing: cross-linguistic evidence from German and English. *J. Psycholinguist. Res.* 32, 37–55. doi: 10.1023/A:1021933015362
- Karaca, F., Brouwer, S., Unsworth, S., and Huetting, F. (2021a). “Individual differences in predictive processing: Evidence from Turkish-speaking monolingual adults,” in *The 27th Architectures and Mechanisms for Language Processing Conference (AMLaP 2021)*.
- Karaca, F., Brouwer, S., Unsworth, S., and Huetting, F. (2021b). Prediction in bilingual children, vol. 12,” in *Prediction in Second Language Processing and Learning*, eds E. Kaan and T. Grüter (Amsterdam: John Benjamins Publishing Company). doi: 10.1075/bpa.12.06kar
- Karaca, F., Brouwer, S., Unsworth, S., and Huetting, F. (2022). “Role of language experience in predictive processing,” in *Poster Presented at the 4th International Symposium on Bilingual and L2 Processing in Adults and Children (ISBPAC 2022)* (Tromsø).
- Kunduz, A. C., and Montrul, S. (2022). Sources of variability in the acquisition of [differential object marking] by Turkish heritage language children in the United States. *Bilingualism* 25, 603–616. doi: 10.1017/S1366728921001000
- Küppers, A., Şimşek, Y., and Schroeder, C. (2015). Turkish as a minority language in Germany: aspects of language development and language instruction. *Zeitsch. Fremdsprachenforsch.* 26, 29–51.
- Kutlu, E., Chiu, S., and McMurray, B. (2022). Moving away from deficiency models: gradience in bilingual speech categorization. *Front. Psychol.* 13:1033825. doi: 10.3389/fpsyg.2022.1033825
- Luk, G. (2022). Justice and equity for whom? Reframing research on the “bilingual (dis) advantage”. *Appl. Psycholinguist.* 44, 301–315. doi: 10.1017/S0142716422000339
- Mani, N., and Huetting, F. (2012). Prediction during language processing is a piece of cake-but only for skilled producers. *J. Exp. Psychol.* 88:843. doi: 10.1037/a0029284
- Marsden, E., Morgan-Short, K., Thompson, S., and Abugaber, D. (2018). Replication in second language research: narrative and systematic reviews and recommendations for the field. *Lang. Learn.* 68, 321–391. doi: 10.1111/lang.12286
- Mitsugi, S., and Macwhinney, B. (2016). The use of case marking for predictive processing in second language Japanese. *Bilingualism* 19, 19–35. doi: 10.1017/S1366728914000881
- Özge, D., Kornfilt, J., Maquate, K., Küntay, A. C., and Snedeker, J. (2022). German-speaking children use sentence-initial case marking for predictive language processing at age four. *Cognition* 221:104988. doi: 10.1016/j.cognition.2021.104988
- Özge, D., Küntay, A., and Snedeker, J. (2019). Why wait for the verb? Turkish speaking children use case markers for incremental language comprehension. *Cognition* 183, 152–180. doi: 10.1016/j.cognition.2018.10.026
- Özsoy, O., Iefremenko, K., and Schroeder, C. (2022). Shifting and expanding clause combining strategies in heritage Turkish varieties. *Languages* 7:242. doi: 10.3390/languages7030242
- Papoutsaki, A. (2015). “Scalable webcam eye tracking by learning from user interactions,” in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 219–222. doi: 10.1145/2702613.2702627.
- Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., and Hays, J. (2016). “Webgazer: scalable webcam eye tracking using user interactions,” in *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, 3839–3845.
- Pickering, M. J., and Gambi, C. (2018). Predicting while comprehending language: a theory and review. *Psychol. Bull.* 144:1002. doi: 10.1037/bul0000158
- Polinsky, M. (2018). *Heritage Languages and Their Speakers*, Vol. 159. Cambridge: Cambridge University Press. doi: 10.1017/9781107252349
- Putnam, M. T., and Sánchez, L. (2013). What’s so incomplete about incomplete acquisition?: a prolegomenon to modeling heritage language grammars. *Linguist. Approach. Bilingual.* 3, 478–508. doi: 10.1075/lab.3.4.04put
- R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Dunabeitia, J. A., Gharibi, K., et al. (2022). Monolingual comparative normativity in bilingualism research is out of “control”: arguments and alternatives. *Appl. Psycholinguist.* 44, 316–329. doi: 10.1017/S0142716422000315
- Schwarz, F., and Zehr, J. (2021). “Tutorial: introduction to PCIBex-an open-science platform for online experiments: design, data-collection and code-sharing,” in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 43.
- Sekerina, I. A., and Sauermann, A. (2015). Visual attention and quantifier-spreading in heritage Russian bilinguals. *Sec. Lang. Res.* 31, 75–104. doi: 10.1177/0267658314537292
- Semmelmann, K., and Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: a first look. *Behav. Res. Methods* 50, 451–465. doi: 10.3758/s13428-017-0913-7
- Slim, M. S., and Hartsuiker, R. J. (2021). “Online visual world eye-tracking using webcams,” in *Poster Presented at the 27th Annual Mechanisms and Language Processing Conference (AMLaP 2021)*.
- Slim, M. S., and Hartsuiker, R. J. (2022). Moving visual world experiments online? A web-based replication of Dijkgraaf, Hartsuiker, and Duyck (2017) using PCIBex and webgazer.js. *Behav. Res. Methods*. doi: 10.3758/s13428-022-01989-z
- Soares, S. M. P., Prystauka, Y., DeLuca, V., and Rothman, J. (2022). Type of bilingualism conditions individual differences in the oscillatory dynamics of inhibitory control. *Front. Hum. Neurosci.* 16:910910. doi: 10.3389/fnhum.2022.910910
- Sorace, A. (2011). Pinning down the concept of “interface” in bilingualism. *Linguist. Approach. Bilingual.* 1, 1–33. doi: 10.1075/lab.1.1.01sor
- Sorace, A., and Serratrice, L. (2009). Internal and external interfaces in bilingual language development: beyond structural overlap. *Int. J. Bilingual.* 13, 195–210. doi: 10.1177/1367006909339810
- Steffan, A., Zimmer, L., Arias-Trejo, N., Bohn, M., Ben, R. D., Flores-Coronado, M. A., et al. (2023). Validation of an open source, remote web-based eye-tracking method (WebGazer) for research in early childhood. *PsyArXiv [Preprint]*. doi: 10.31234/osf.io/7924h
- Stone, K., Lago, S., and Schad, D. J. (2021). Divergence point analyses of visual world data: applications to bilingual research. *Bilingualism* 24, 833–841. doi: 10.1017/S1366728920000607
- Vasishth, S. (2022). Some right ways to analyze (psycho) linguistic data. *Annu. Rev. Linguist.* 9, 273–291. doi: 10.31234/osf.io/y54va
- Vasishth, S., and Gelman, A. (2021). How to embrace variation and accept uncertainty in linguistic and psycholinguistic data analysis. *Linguistics* 59, 1311–1342. doi: 10.1515/ling-2019-0051
- Vos, M., Minor, S., and Ramchand, G. (2021). “Comparing infrared and webcam-based eye tracking in the visual world paradigm,” in *Poster Presented at the 27th Annual Mechanisms and Language Processing Conference (AMLaP 2021)*. doi: 10.31234/osf.io/36skd
- Vos, M., Minor, S., and Ramchand, G. C. (2022). Comparing infrared and webcam eye tracking in the visual world paradigm. *Glossa Psycholinguist.* 1. doi: 10.5070/G6011131
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686
- Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., et al. (2022). Heritage speakers as part of the native language continuum. *Front. Psychol.* 12:717973. doi: 10.3389/fpsyg.2021.717973
- Wood, S. N. (2003). Thin-plate regression splines. *J. R. Stat. Soc. B* 65, 95–114. doi: 10.1111/1467-9868.00374

Appendix

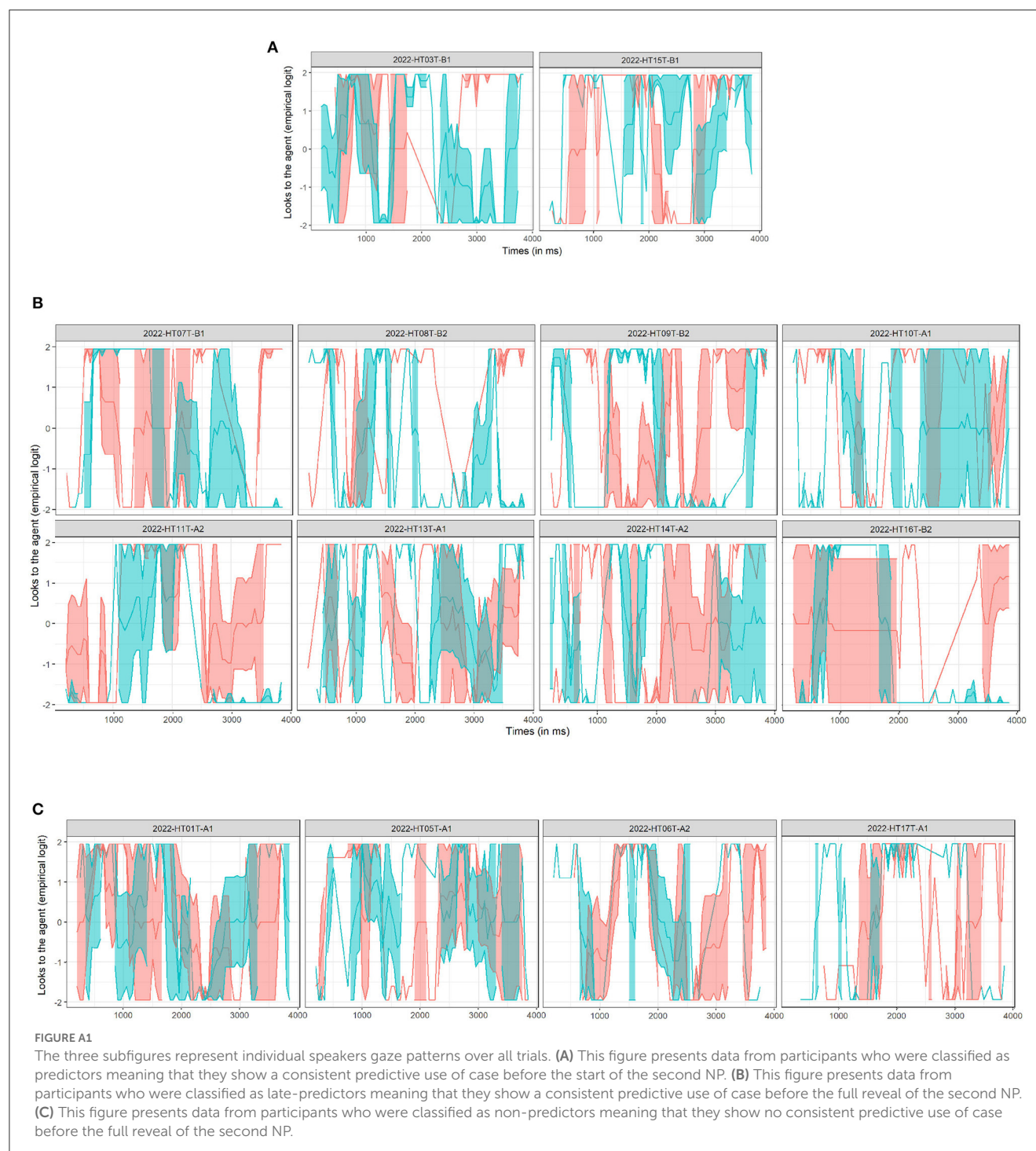
1. Divergence point estimation

In [Figure 2](#), the (uncorrected) divergence point estimates are represented by the purple triangle-shaped symbols in each panel. Recall that the NP2 onset began at 2,300 ms in all experiments. The divergence point in Experiment 1 (monolingual speakers, webcam-based eye-tracking) fell within the region of interest for the predictive effect of the case, at 2,000 ms ($z = 2.77$, $p = 0.00554$). In contrast, in Experiment 3 (HSs, in-lab eye-tracking), the divergence point was outside of the region of interest, at 2,600 ms ($z = 2.56$, $p = 0.0106$). There was no divergence point in Experiment 2 (HSs, webcam-based eye-tracking) at all in the region

of interest. Thus, at the group level, we locate the effect of predictive processing of case-marking cues in the monolingual speakers, but not in the HSs, regardless of the eye-tracking set-up.

It is important to consider the corrected multiple comparisons which is why we have calculated also the Bonferroni-corrected and FDR-controlled divergence point estimates ([Stone et al., 2021](#)). The aim here is to limit the rate of false positives (Type I error) which might arise given the big number of time-points that are statistically compared to each other using this method. However, none of the estimated divergence points in our analysis survived the correction which is why we do not report them here. Plausible explanations for this are the low data resolution from webcam-based eye-tracking in Experiment 1 and 2 and the limited sample size in Experiments 2 and 3.

2. Individual predictor categories based on growth curve analyses





OPEN ACCESS

EDITED BY

Fatih Bayram,
UiT The Arctic University of Norway, Norway

REVIEWED BY

Mike Putnam,
The Pennsylvania State University (PSU),
United States
Javier Pérez-Guerra,
University of Vigo, Spain
Cristina Guardiano,
University of Modena and Reggio Emilia, Italy

*CORRESPONDENCE

Wintai Tsehaye
✉ wintai.tsehaye@uni-mannheim.de

RECEIVED 23 January 2023

ACCEPTED 15 June 2023

PUBLISHED 24 August 2023

CITATION

Tsehaye W (2023) Light-weights placed right:
post-field constituents in heritage German.
Front. Psychol. 14:1122129.
doi: 10.3389/fpsyg.2023.1122129

COPYRIGHT

© 2023 Tsehaye. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](#). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Light-weights placed right: post-field constituents in heritage German

Wintai Tsehaye*

Department of English Linguistics, University of Mannheim, Mannheim, Germany

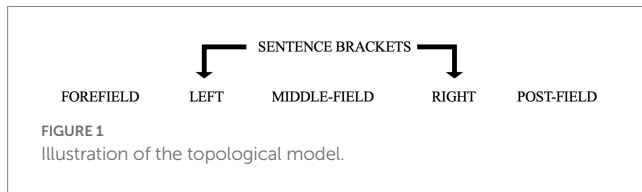
This study focuses on the linearization of constituents at the right sentence periphery in German, specifically on non-clausal light-weight constituents (LWCs) in the post-field. Spoken and written productions of German heritage speakers (HSs) with English as their majority language (ML) and of monolingually-raised speakers (MSs) of German are analyzed in different registers. The right sentence periphery is an area comprising a lot of variation and it is therefore intriguing to see how the two speaker groups deal with the options available if faced with the same communicative tasks. The overall goal is to answer the question whether the production of post-field LWCs in German HSs and MSs can provide us with evidence for ongoing internal language change and for the role of language contact with English. The analyses show a similar variational spectrum of LWC types and frequencies across speaker groups but a different distributional variation. The results show effects of register-levelling in the HS group, as they do not differentiate between the formal and informal setting unlike the MS group. Therefore, rather than transfer from the ML, the source of differing distributional variation of LWCs lies in the diverging adherence to register norms due to different exposure conditions across speaker groups.

KEYWORDS

heritage German, right sentence periphery, post-field, light-weight constituents, German–English language contact, register

1. Introduction

Heritage speakers (HSs) are a theoretically most relevant speaker group for linguistic research across subdomains of their grammars. Their often very heterogeneous acquisition context and outcome makes them an excellent learner type to investigate bilingualism, interface phenomena, as well as synchronic and diachronic effects of language contact. Heritage speaker's linguistic competence and performance show considerable inter- and intraindividual variation and they often rate themselves better in spoken than in written productions (Montrul, 2016, p. 44ff.), especially where their heritage language (HL) is not supported within the educational system. Furthermore, specific linguistic areas are more prone to variation (e.g., morphology, discourse) than others (e.g., phonology, syntax). An explanation for variation across linguistic subdomains is found in the interface hypothesis (Sorace, 2011; Tsimpli, 2014), which states that “language structures involving an interface between syntax and other cognitive domains are less likely to be acquired completely than structures that do not involve this interface” (Sorace, 2011, p. 1).



Adopting a topological framework (see below), this paper focuses on the linearization of constituents at the right sentence periphery of German, specifically on post-field constituents in spoken and written productions of German HSs with English as their majority language (ML) and of monolingually-raised speakers (MSs) of German. I investigate the production of light-weight constituents (LWCs), i.e., non-clausal constituents which appear after the clause-final predicate, in the post-field (see Figure 1 and example 1). These particular clausal patterns diverge from the canonical pattern of German word order, and their status as more or less “marked” involves the interface of syntax and discourse-pragmatic factors.

In the topological model, constituents appear in different “fields” from which they can be moved either to the forefield, via topicalization, scrambled in the middle-field or extraposed into the post-field¹ (Drach, 1963; Zifonun et al., 1997; Wöllstein, 2014; Zifonun, 2015). While the forefield and the left clausal edge have received considerable attention (Müller, 2003; Freywald et al., 2015; Wiese and Müller, 2018; Bunk, 2020; Rocker, 2022; Wiese et al., 2022, among others), less attention has been given to the post-field and the right clausal edge. Researchers who have however worked on the right sentence periphery have identified it as a very heterogeneous domain and called for a more differentiated analysis with conceptually separable subdivisions (see Vinckel-Roisin, 2015 for an overview).

In example (1), the LWC in the post-field is realized as the adverbial phrase (ADVP) *ganz schnell* (very quickly) which appears after the participle *gestoppt* (stopped).

(1) das erste Auto hat gestoppt (**ganz schnell**_{ADVP}) (RUEG corpus informal spoken²)

“The first car had stopped very quickly.”

The post-field, broadly defined as the area following the right sentence bracket,³ is typically considered an area reserved for heavy constituents such as subordinate clauses extraposed from the middle-field in order to reduce cognitive load⁴ (Haider, 2010;

Proske, 2015; Imo, 2016, p. 207). The realization of LWCs in the post-field as shown in example (1), while not ungrammatical, is often considered marked (e.g., Andersen, 2008; Vinckel-Roisin, 2012; Frey, 2015 among others). However, when we take into account different registers in speaking and writing, the situation is not as straightforward. Depending on the formality and mode of a production, we find a considerable range of constituents, like those in example (1), in the post-field not only of HSs but also of MSs of German. Therefore, the role and the effects of register variation need to be included in the analysis of LWCs in the post-field.

Previous research has shown that prepositional phrases (PPs) are particularly frequent in the post-field (Haider, 2010, p. 191; Zifonun, 2015; Imo, 2016). In German, PPs can occur before the verb, in the middle-field (example 2a), or after the verb, in the post-field⁵ (example 2b).

(2a) weil das Auto (**wegen dem Hund**_{PP}) stoppen musste.

(2b) weil das Auto stoppen musste (**wegen dem Hund**_{PP}).

“Because the car had to brake on account of the dog.”

In English, comparable PPs must follow the verb but cannot appear between the subject and the verb. Therefore, within the analysis of LWCs undertaken here, special emphasis is placed on the extraposition of PPs across speaker groups as it can provide us with information on the influence of language contact.

Even though English and German are both Germanic languages, they exhibit considerable typological differences in terms of word order. These differences make English and German an intriguing language pair to investigate the influence of language contact, language dominance and transfer potential. German is among one of the better-researched languages in the field of HL research. There is a long-standing history of investigations on Germanic varieties in English dominant environments, such as Australian German, Texas German, Pennsylvania German, and Moundridge Schweitzer German and existing research on these varieties, indeed, finds trends of increased frequencies of LWCs in the post-field attributable to language contact with English (e.g., Clyne, 2003; Westphal Fitch, 2011). However, there is so far little work on the type of HSs discussed here, namely second-generation immigrants born in the U.S. or early-childhood arrivers who are not part of a bigger German speaking Language Island community.

Overall, the phenomena investigated here have until recently been neglected in German linguistics, under-researched for different acquisition types, and, to the best of my knowledge, not pursued in research on German as a HL in second-generation immigrants under intense language contact with English as a ML. Section 2 provides the theoretical background and anchors the present analysis in previous studies. Section 3 introduces the participants, the corpus, and the applied methodology. Section 4 illustrates the results, followed by a discussion in Section 5. Section 6 summarizes the results and

¹ Current research calls for further distinctions and additional fields, such as the pre-forefield, the extended post-field, and the right outer field (Zifonun, 2015), which only play a marginal role in the later discussion of this article.

² This refers to one of the four narrations (formal spoken, formal written, informal spoken, informal written) which the participants were asked to produce. Section 3.2 provides a detailed explanation of the herein applied method for data collection.

³ The right sentence bracket can be realized or realizable (see Vinckel-Roisin, 2012, p. 144).

⁴ Especially long relative clauses, which are placed in the middle-field create a considerable distance between the subject and the finite verb, which makes them hard to process.

⁵ With the exception of resultative or directional predicates (e.g., Er hat es geschnitten [in kleine Stücke] (He has cut it into small pieces), Haider, 2010, p. 191).

addresses limitations of the current analysis as well as perspectives for follow-up research.

2. Theoretical background

2.1. Heritage speakers

One finds a plethora of HS definitions in the literature, depending on the theoretical research focus. According to the definition adopted here, HSs are bilinguals who grow up acquiring their HL within the family but are raised in an environment where another language has majority status (Rothman, 2007; Montrul, 2016; Polinsky, 2018). They can be considered either simultaneous bilinguals, exposed to two languages (the HL and the ML) from birth, or early sequential bilinguals who first acquire the HL and are then exposed to the ML of their country of residence. Intensive exposure to an early second language often results in a dominance shift from the HL to the ML (Pascual Y Cabo and Rothman, 2012; Kupisch and Rothman, 2018; Ortega, 2020 among others). Consequently, HSs usually use their ML in a wider range of communicative situations than their HL. In some cases, they may only be addressed in their HL by one other family member, in other cases, there may be an actual HL speaker community outside the family.⁶

Past research on HSs reveals a deficit-oriented view on their linguistic competence and performance, which resulted in labels such as semi-speakers or incomplete acquirers. However, this view has shifted due to a surge of interest in divergent attainment or differential acquisition (cf. Kupisch and Rothman, 2018) and led to extensive discussions of a suitable baseline, i.e., the actual input that HSs receive in the HL and not the variety spoken by MSs they are not exposed to (Polinsky, 2018, p. 3ff.; Rothman et al., 2022). Accordingly, recent studies argue that HSs are native speakers of their HL (Rothman and Treffers-Daller, 2014; Montrul, 2016; Kupisch and Rothman, 2018; Tsehaye et al., 2021; Wiese et al., 2022). In the current study, the data collected from German MSs is not used as a baseline, but as comparative data enabling us to identify contact-independent internal dynamics as well.

2.2. Syntactic linearization in German

The topological model, first conceptualized by Drach (1963), uses the metaphors of *sentence brackets* and *topological fields* to describe and investigate German sentences. It should be emphasized that using the topological model results in a purely linear analysis and not in hierarchical, binary-branching structures.⁷ Table 1 illustrates the placement of constituents across topological fields with unmarked post-field constituents.

TABLE 1 Example sentences with unmarked post-field constituents.

3	Forefield	Left sentence bracket	Middle-field	Right sentence bracket	Post-field
a	Ich	habe	heute einen ziemlich heftigen Unfall	erlebt. ¹	
<i>'I have experienced a rather severe accident today.'</i>					
b	Ich	wollte	gerne über einen Unfall	berichten	den ich gesehen habe.
<i>'I would like to report about an accident which I have seen.'</i>					
c	den		ich	gesehen habe.	
<i>'which I have seen'</i>					
d	Ich	wollte	gerne über einen Unfall, den ich heute gesehen habe,	berichten.	
<i>'I would like to report about an accident which I have seen today.'</i>					
e	Dann	fingen	die beiden Autofahrer	an,	den Unfall zu begutachten.
<i>'Then both drivers started to assess the accident.'</i>					

¹Most of the examples throughout this article have been taken from the RUEG corpus and were indicated as such (<https://korpling.german.hu-berlin.de/annis3/#c=rueg>). Some of the examples have been adapted to illustrate the variational spectrum of German sentences. They do, however mirror the syntactic patterns identified in the corpus.

In main and declarative clauses (examples 3a/b/d/e) the finite verb occurs in the left sentence bracket (LSB) while the rest of the verbal complex occurs in the right sentence bracket (RSB). In subordinate clauses (example 3c), complementizers⁸ occupy the LSB while the finite predicate occurs in the RSB. The area in front of the LSB is called the forefield. It holds constituents that are pre-posed or topicalized from the middle-field, which is the field encompassed by the sentence brackets. The area after the RSB is labeled the post-field. The post-field can hold constituents that have been extraposed from the middle-field, including clausal adjuncts such as relative or complement clauses (see examples 3b/e).⁹ While Table 1 showed the canonical, unmarked linearization of constituents in German sentences, Table 2 illustrates

⁶ For the participants presented in this research, no larger HL speaker community outside the family is present. Some participants, however, report regular visits to relatives in Germany.

⁷ See Haider (2010) for hierarchical approaches in post-field analyses.

⁸ Even though relative pronouns and relative adverbs also lead to VL clauses, they are not placed in the LSB. One line of argumentation is that relative pronouns and relative adverbs, unlike complementizers, function as constituents and are, thus, placed in the forefield (Wöllstein, 2014, p. 27ff.; Imo, 2016, p. 214).

⁹ From a generativist perspective, researchers still discuss the source of constituents appearing in the post-field (extraposition vs. base-generation). Some argue that movement as the source of extraposition is lacking in its explicatory nature (Haider, 2010), while others even go as far as saying that there is no movement to the right in German (Frey, 2015).

a different set of cases, thereby shifting the attention to the spectrum of constituents found in the post-field.

Although the clauses in Table 2 show canonical verb placement, we also see deviations from what are assumed to be orthodox—or stylistically “desirable”—constituent candidates in the respective fields. Example (4a) illustrates the extraposition of the PP *auf einem Parkplatz* (in a parking lot). Example (4b) exhibits the placement of the adverbial *heute* (today) in the post-field while example (4c) shows the extraposition of the DP *einen ziemlich heftigen* (a rather severe one).

TABLE 2 Example sentences with marked post-field constituents.

4	Forefield	LSB	Middle-field	RSB	Post-field
a	Ich	habe	heute einen Unfall	beobachtet	auf einem Parkplatz.
<i>‘I have observed an accident in a parking lot today.’</i>					
b	Ich	habe	einen ziemlich heftigen Unfall	beobachtet	heute.
<i>‘I have observed a rather severe accident today.’</i>					
c	Ich	habe	heute einen Unfall auf einem Parkplatz	beobachtet	einen ziemlich heftigen.
<i>‘I have observed a rather severe accident in a parking lot today.’</i>					

All post-field constituents in Table 2 can be categorized as LWCs which, as in the case of (4a/b) could have easily “stayed” in the middle-field. Example (4c) functions as the specification of the DP antecedent *einen Unfall* (an accident) in the middle-field and, thus, could not have been realized in the middle-field. However, the DP could have been modified as *einen ziemlich heftigen Unfall* (a rather severe accident) within the middle-field, i.e., there is no syntactic demand to extrapose this information. Such occurrences show the existence of a variational spectrum that holds especially for spoken productions of German (cf. Zifonun et al., 1997; Imo, 2015; Zifonun, 2015). A greater variational spectrum in spoken or conceptually spoken¹⁰ productions compared to written or conceptually written productions has been shown for other syntactic phenomena as well, suggesting that some linearization patterns might occur exclusively or more frequently in the spoken mode (Andersen, 2008, p. 2). However, variation is also found in written productions. Previous studies have attested considerable variation in the frequency of post-field productions in the written mode, with the least occurrences in scientific texts and most occurrences in informal productions (Roelcke, 1997, p. 158). This strengthens the fact that register differentiations need to be taken into account in investigations of post-field variation.

The availability of large synchronic and diachronic corpora of spoken and written German shows that even across MSs of German, the right sentence periphery is an area of considerable variation, with

fluctuating degrees of markedness across registers. It is therefore intriguing to ask how both speaker groups, HSs and MSs, when faced with the same communicative challenge, deal with post-field options, given the fact that HSs of German have less contact with different registers than MSs and experience extensive language contact.

The existence of a post-field and its availability for various constituents in it is ultimately dependent on the formation of the sentence brackets. Only after the distinction of finite and non-finite verbs, and the asymmetric placement of finite and non-finite verbs in main and subordinate clauses is mastered, are we able to assess whether and with which constituents the post-field is filled. Head directionality within the verb phrase (VP), and hence, the RSB, are acquired early in L1, quickly followed by the discovery of the LSB and its canonical occupant, finite verbs (Tracy, 2011; Schulz and Tracy, 2018). The head parameters relevant for German main and subordinate clauses can be considered fixed around age three (Fritzenschaft et al., 1990; Rothweiler, 2006; Tracy, 2011; Müller et al., 2018). Once the post-field “exists,” learners still need to figure out which constituents can access it. A study which looked at the emergence of the topological fields and the occurrence of constituents in the right sentence periphery in children around age two found instances of complements, i.e., direct objects in form of DPs, in the post-field, which is highly non-canonical in contemporary German. With time, children’s productions converged on those of adults and became canonical (Elsner, 2015). The results of this study illustrate that even in monolingual L1 acquisition without contact with another language, one finds (non-) canonical variation in the linearization at the right sentence periphery.

After head directionality and finiteness are acquired, the placement of constituents in the post-field is furthermore influenced by register norms and discourse-pragmatic requirements of the communicative situation which will be outlined in the following. According to Biber and Conrad (2001, p. 175), a register is a variety which can be defined by specific communicative and contextual parameters, such as interlocutors involved, purpose, as well as mode and formality of the interaction. Previous research (Polinsky, 2018, pp. 323–324; Aalberse et al., 2019, p. 148 to name but a few) has shown that HSs, who often do not learn to read and write in the HL, cannot be expected to have available the register spectrum, genres, or styles accessible to age-matched ML speakers of the same language in the country of origin. Dominance shift, the unavailability of a HL community, the greater social prestige of their ML, as well as the absence of formal education in the HL contribute to diverging levels of adherence to register norms between HSs and MSs as well as between the HL and the ML in individual speakers.

Discourse-pragmatic reasons for placing constituents in the post-field are manifold, and arguments for differentiating various subfields and ways for filling them (movement, free adjunction) are controversial, as shown in previous research (Zifonun et al., 1997; Frey, 2015; Vinckel-Roisin, 2015; Zifonun, 2015; Imo, 2016, among others). It has been argued that (a) the post-field cannot be a single undifferentiated field¹¹ and (b) not all constituents that appear in this

¹⁰ Spoken and written productions can be seen as part of a conceptual continuum. This means that, depending on the situation and the context, written productions can become conceptually spoken (e.g., a diary entry) and spoken productions can become conceptually written (e.g., a sermon, cf. Koch and Oesterreicher, 2012).

¹¹ Due to scarceness of datapoints in this corpus, no distinction between the narrow and extended post-field (or post-field and right outer field) is applied in the quantitative analysis.

area seem to be extraposed from the middle-field but could also be more or less freely adjoined and base-generated (Vinckel-Roisin, 2012; Frey, 2015). Zifonun et al. (1997) propose subdividing the right sentence periphery into two fields: the post-field and the right outer field. The post-field contains syntactically integrated as well as non-integrated constituents such as subordinate clauses. The right outer field can be distinguished from the post-field insofar as its constituents are not syntactically integrated units of the preceding clause (Vinckel-Roisin, 2012). The right outer field can be occupied, regardless of whether or not the post-field is filled, and constituents in this position are typically prosodically or orthographically highlighted. The right outer field is usually reserved for constituents with discourse-pragmatic functions such as comments, verification of the audience's attention or requests for reactions (cf. Imo, 2016, p. 223 ff.). Example (5) illustrates this distinction with the relative clause *der ziemlich heftig war* (which was rather severe) in the post-field and the discourse marker *nicht wahr* (isn't that right) in the right outer field.

(5) Wir haben heute einen Unfall auf einem Parkplatz gesehen, der ziemlich heftig war, **nicht wahr**?

"We saw an accident in a parking lot today, which was rather severe, isn't that right?"

Depending on their placement within the overall area of the post-field (narrow vs. extended post-field), their clausal status, and the degree of phonetic integration,¹² functions addressed in the literature on MSs of German are the addition of detail to previously mentioned content, repairs, and evaluative afterthoughts in the service of discourse coherence.¹³

2.3. The influence of language contact

As already mentioned, the HSs in this study have English as their ML. For the phenomena under discussion in this paper, the most crucial difference between German and English consists in verb placement, with German being head-last within the VP, while English is head-first. German further exhibits an asymmetry in finite verb placement, with V2 structures in main clauses and VE structures in subordinate clauses, whereas English has an SVO structure across clauses apart from subject-auxiliary-inversion and highly restricted subject-main-verb-inversion with intransitive verbs (see Table 3).

TABLE 3 German and English word order.

	Contrasts	German	English
I	VP (across clauses)	[O...V _(-fin)]	[V _(-fin) O ...]
II	main clauses	(X) V2 _(+fin) ...V _(-fin)	(X) SV _(+fin) O + residual V2
III	subordinate clauses	COMP..... V _(+fin)	COMP SV _(+fin) O

¹² Discourse structuring devices, i.e., hesitations, pauses, and intonational breaks (or punctuation in written productions) can provide relevant cues to the degree of connectedness to the previous clause and can be used to distinguish between functional differences of constituents in the right sentence periphery (e.g., Altmann, 1981; Frey, 2015 and the references therein; Imo, 2015).

¹³ The functional exploitation of the postverbal position is already visible in German-speaking children's early multiword utterances (Tracy, 1991, p.187).

One relevant question to ask, then, is the following: Given intensive language contact between German and English, to what extent do HSs observe these contrasts? Do we see an increase in extrapositions which could be due to cross-linguistic influence from English? Such trends have been observed in previous studies on speakers of German Language Islands. Westphal Fitch (2011) found increased numbers of extrapositions in spoken productions in speakers of Palatinate and Pennsylvania German in comparison to speakers of Standard German due to language contact with English.

Despite the variational spectrum documented especially in spoken German, a crucial restriction, as already mentioned, is that contemporary German, does not allow the placement of direct objects in the post-field¹⁴ (Zifonun, 2015, p. 30), as in example (6).

(6) *Wir haben gesehen **einen Hund**.

"We have seen a dog."

The translation of example (6) demonstrates that English calls exactly for this linearization, with the verbal head immediately adjacent to its complement. Previous studies on heritage German in Australia also attested increased extrapositions of LWCs, including the extraposition of direct objects, which Clyne (2003), attributes to intense contact with English, see example (7).

(7) Mummy hat gesagt **die Wörter für mich**.

"Mummy told me what to say" (Clyne, 2003, p. 137).

Productions like the one in example (7) legitimize the question whether language contact with English enhances the non-canonical placement of direct objects in the post-field of HSs of German.

The typological differences between English and German also become apparent when looking at the linearization of PPs. In English for instance, PPs usually appear after the verb due to the strict VO serialization across clauses.¹⁵ In German, due to the sentence brackets, the PP can occur in the middle-field (i.e., before the finite verb) or in the post-field (i.e., after the finite verb). Therefore, HSs have an additional option for PP placement in German in comparison to English. Choosing to extrapose the PP into the post-field results in clauses which are, in their surface syntactic realization, more parallel to the unmarked English linearization contrary to producing the PP in the middle-field, which is not possible in English. Research on German Language Islands in the USA has shown that if parallelism between structures exists, these structures may appear more frequently than

¹⁴ Diachronic analyses of the post-field show that (direct) objects are found in the post-field without jeopardizing the grammaticality of the sentence up until the era of New High German (Hinterhölzl, 2004; Coniglio and Schlachter, 2015).

¹⁵ PPs can also be topicalized in English, thus occurring before the subject (e.g., on the table, she placed a vase). In German, topicalization of PPs is also possible. The PP would then, however, be placed in the forefield (e.g., auf den Tisch platzierte sie eine Vase). This serialization would be ungrammatical in English (i.e., *on the table placed she a vase). Similar surface syntactic patterns in English are residual and restricted to transitive verbs (e.g., on the table stood a vase) and presentational there-constructions, both highly dependent on the preceding context.

non-parallel ones (Westphal Fitch, 2011, p. 374; Hopp and Putnam, 2015 and the references therein).¹⁶

Examples (8a/b) were produced by the same participant, once in the HL, German and once in the ML, English and illustrate this surface parallelism with the PP following the verb in both cases.

(8a) der Hund an der anderen Seite von der Straße ist vorgerannt (zum Ball_{PP}) RUEG corpus formal written.

“The dog on the other side of the street ran towards the ball.”

(8b) and the dog leaped forward (to the ball_{PP}) RUEG corpus formal written.

In the light of this typological difference between German and English, the question arises whether language contact with English facilitates the production of PPs in the post-field of German HSs, resulting in an overlapping surface structure across their languages—a question that explores the interplay of surface parallelism on the one hand and transfer or avoidance on the other hand.

An additional point—and analytical problem—paramount to the question of cross-linguistic influence and transfer phenomena due to surface parallelism is the fact that whenever we have a clause with an empty RSB (9a) or a clause with an empty RSB and a filled post-field (9b), the surface structure between German and English clauses becomes identical (see Table 4).

TABLE 4 Example clauses with empty RSB illustrating surface parallelism.

9	Forefield	LSB	Middle-field	RSB	Post-field
a	Ich	sah	einen Autounfall.	-	
	<i>‘I saw a car accident.’</i>				
b	Ich	sah	einen Autounfall	-	gestern.
	<i>‘I saw a car accident yesterday.’</i>				

In the face of these partial overlaps and cross-linguistic parallels in surface structure, the question of whether contact with English boosts LWCs (including direct objects) in the post-field in HSs in comparison with MSs becomes particularly relevant.

2.4. The present study

The data presented in this article was not specifically elicited to investigate post-field productions. Nevertheless, it is highly suitable to investigate the variational spectrum at the right sentence periphery in different registers and the role of language contact: It contains the productions of MSs and HSs of German who were faced with the same communicative tasks, therefore allowing for adequate comparisons. The following research questions and hypotheses could therefore be formulated:

RQ1: Which types of LWCs can be found in the post-field of HSs and MSs of German, and with which frequency?

H1: Due to typological differences in the syntactic realization of constituents in German and English, HSs will show more various

LWCs and increased frequencies of LWCs in their post-field productions.

RQ2: Does register influence the type and frequency of constituents in the post-field of HSs and MSs of German?

H2: Register will have an influence on the frequency of LWCs in the post-field across speaker groups with more constituents produced in the informal setting and the spoken mode.

RQ3: Do HSs of German produce more PPs in the post-field than MSs of German?

H3: HSs of German will have higher frequencies of PPs in their post-field than MSs of German due to extensive contact with English.

3. Method

3.1. Participants

The present study included 61 adolescent participants aged 13 to 19 years (mean age = 16.1, SD = 1.35, 32 females). The overall number of participants can be subdivided into 29 HSs of German with ML English (mean age = 15.6, SD = 1.57, 12 females),¹⁷ and 32 MSs of German (mean age = 16.6, SD = 0.91, 20 females). All HSs grew up in the USA in a majority English environment, speaking German with at least one native German-speaking parent in the household.¹⁸ The participants in the MS group were defined as individuals whose L1, German, was the only language spoken at home, but who might have acquired further languages through foreign language instruction. The German and English productions of the HSs were elicited in the U.S., the productions of the German MSs in Germany. The data was retrieved from the openly accessible RUEG 0.4.0 corpus (Wiese et al., 2021).

3.2. Materials and procedure

The controlled and standardized data elicitation followed the language situations methodology (Wiese, 2020). Participants watched a short non-verbal video of a rear-ending car accident and recounted what they saw, imagining themselves witnesses to the accident in four different narrations, which we operationalized as productions in different registers. Data collection took place in two differently arranged rooms: a formal and an informal one with

¹⁷ One adolescent HS did not enter their birthdate, therefore, the mean and standard deviation for the HS group was calculated for 28 participants only.

¹⁸ Participation requirements were that the HSs were either born in the U.S., or moved there before age two. The HS participants should not have received bilingual education but may have participated in German “Saturday schools” or other German-speaking activities. Speakers of established German Language Islands were excluded from the study.

¹⁶ As we also know from code-switching research, parallel surface structures may ease language mixing (Poplack, 1980; Muysken, 2000).

different elicitors in each room. The elicitation of the formal productions took place in an office-like room, whereas the informal productions were elicited in a casual setting with snacks and beverages offered and following a 10–15 minute-long informal, task-unrelated conversation in the target language in order to create a more relaxed atmosphere. During one session, all participants watched the video three times in total (twice in the first setting, once in the second setting) and were asked to recount it in two different modes: spoken and written.

In the formal recounting, the participants were asked to send a voice message to a police hotline (spoken) and a witness report to the police (written). In the informal setting, they had to send a voice message (spoken) and a text message (written) to a friend via an instant messenger. The order of settings (formal/informal) and modes (spoken/written) was balanced across participants. The MSs completed all tasks in one session. The HSs completed the tasks in two sessions – one for each language – with an interval of three to five days in between to minimize priming effects and the order of languages counterbalanced across participants. Upon completion of all tasks, participants filled out an online questionnaire¹⁹ about their language background as well as a self-assessment of their abilities in each language on a five-point Likert scale. Self-assessment showed that, in line with previous research, HSs rated their speaking skills higher than their writing skills in their heritage German (speaking mean = 3.71, SD = 0.79; writing mean = 3.03, SD = 1.29). German MSs rated their speaking skills at ceiling and their writing skills almost at ceiling (speaking mean = 4.96, SD = 0.17; writing mean = 4.6, SD = 0.64).

3.3. Data analysis

The spoken and written productions of both speaker groups (HSs and MSs) were annotated according to the topological model based on the KiDKo annotation guidelines (Bunk et al., 2020). All post-field constituents were exported from the RUEG corpus and additionally annotated for their constituent type. Table 5 shows examples for each constituent type produced in the post-field. A total of 708 post-field constituents were annotated.

The corpus includes a total of eight different constituent types: finite subordinate clause (SC), non-finite subordinate clause (INF), prepositional phrase (PP), adverbial phrase (ADVP), determiner phrase (DP), adjectival phrase (ADJP), discourse marker (DM), and DP realized as non-canonical direct object (NONC) of which we found a total of two in the corpus, both produced by the same speaker.

As has already been established, the occurrence of (non-)finite subordinate clauses in the right sentence periphery is canonical and unmarked as it serves to avoid “overloading” the middle-field. Therefore, the focus of the current analysis lies on constituents that are not subordinations, i.e., LWCs. Due to scarceness of data points (a total of 140 LWCs) and, therefore, small numbers in certain categories, the eight constituent types were collapsed into

TABLE 5 List of constituents in the right sentence periphery with examples.

Constituent type	Example
SC: subordinate clause (finite)	hat den mann nicht gesehen [<i>weil ein auto in sein sichtfeld war_{SC}</i>] ¹ 'didn't see the man because a car was in his field of view'
INF: subordinate clause (non-finite)	und ein hund hat versucht [<i>ihn zu fangen_{INF}</i>] 'and a dog tried to catch it'
PP: prepositional phrase	die haben die Straße runtergelaufen [<i>mit einem Ball_{PP}</i>] 'they walked down the street with a ball'
ADVP: adverbial phrase	das auto vorne hat angehalten [<i>plötzlich_{ADVP}</i>] 'the car in front had stopped suddenly'
DP: determiner phrase	die haben irgendwelche Sachen fallen gelassen [<i>Lebensmittel_{DP}</i>] 'they have dropped some things, groceries'
ADJP: adjectival phrase	und die Frau war sehr schockiert [<i>also bisschen perplex_{ADJP}</i>] 'and the woman was very shocked so a bit perplexed'
DM: discourse marker	und die autofahrer sind dann auch gleich ausgestiegen [<i>und so_{DM}</i>] 'and the drivers immediately exited and so on'
NONC: non-canonical direct object	die Mann geht zu helfen [<i>die Mädchen [die essen aufzuholen]_{NONC}</i>] 'the man goes to help the girl pick up the food'

¹All productions in this table have been kept in their original orthography, if written, and in their original structure, if spoken, while canonical morphosyntax and choice of auxiliary have been ignored.

subordinations and LWCs. This resulted in a dependent variable “constituent type” with two levels (1 for LWCs and 0 for SCs²⁰). Generalized binomial linear mixed effects models in R (R Core Team, 2021) and the lme4 package (Bates et al., 2015) were used to analyze the distribution and frequency of LWCs in the right sentence periphery. I specified the fixed effects by including the following dependent variables and their potential interactions: speaker group (HS/MS), setting (formal/informal), and mode (spoken/written) and I used treatment contrast and maximally specified the random effect of participants. To avoid overfitting, I performed backward ANOVAs to deduce the most suitable model. For each model, the *z*- and *p*-values are reported.

In order to answer the third research question, I additionally performed an analysis on the distribution of PPs across narratives and speaker groups. The dependent variable for this analysis was “PP” with two levels (1 for PP and 0 for no PP). Again, I maximally specified the fixed and random effects, used generalized binomial linear mixed effects models, and performed backward ANOVAs for model fitting.

¹⁹ Questionnaire for adolescent participants of Research Unit Emerging Grammars: <https://osf.io/qhupg/>.

²⁰ The variable SC now includes both, finite and non-finite subordinations in the quantitative analysis.

The language situations method and the included task of recounting an accident, especially where a police report is called for, creates a bias in favor of a specific functional kind of extrapositions, namely providing expansions or specifications. Therefore, the post-field constituents can be categorized as:

- i. constituents that can be placed in the middle-field or the post-field resulting in different degrees of markedness: less marked for extraposed heavy constituents such as subordinations with the function of decreasing cognitive load, and more marked for LWCs functioning as afterthoughts or specifications (except for direct objects),
- ii. constituents which can only appear in the post-field as they have an antecedent in the middle-field which they semantically specify or elaborate, or
- iii. syntactically non-integrated constituents that function as metacommentaries.

4. Results

4.1. Descriptives

Descriptive statistics show the mean percentages of LWC types in the post-field across speaker groups (Table 6), the absolute frequencies of LWC types in the post-field across speaker groups and narratives (Table 7) and the mean percentages of LWCs in the post-field across speaker groups and narratives (Table 8).

TABLE 6 Mean percentages of LWC types in the post-field across speaker groups.

Constituent type	Mean percent in HSs	Mean percent in MSs
PP	13.81	9.84
DP	2.86	1.81
DM	0.92	5.02
ADVP	2.86	2.01
ADJP	2.38	0.40
NONC	0.95	0.00

TABLE 7 Absolute frequencies of LWCs in the post-field across speaker groups and narratives.

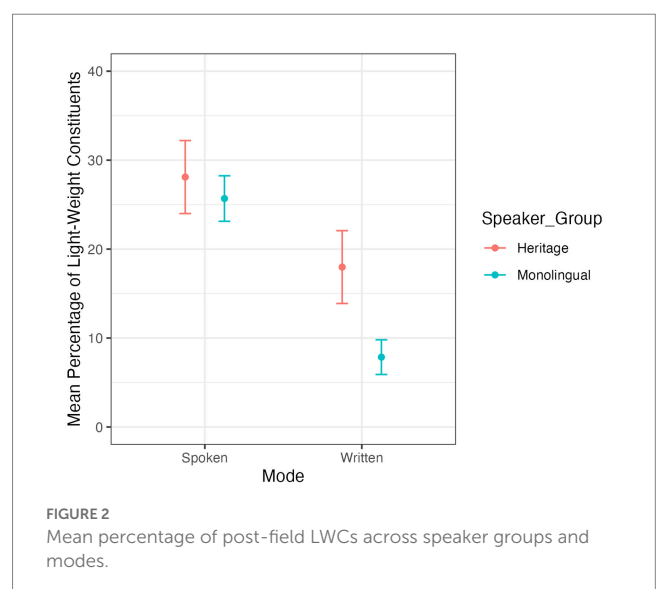
Narrative	Spoken formal		Spoken informal		Written formal		Written informal	
Speaker group	HS	MS	HS	MS	HS	MS	HS	MS
PP	16	25	5	11	8	4	0	5
DP	3	3	1	4	1	0	1	2
DM	0	2	2	21	0	0	0	2
ADVP	2	3	1	5	2	1	1	0
ADJ	0	1	2	0	3	1	0	0
NONC	1	0	1	0	0	0	0	0

TABLE 8 Mean percentages of LWCs in the post-field across speaker groups and narratives.

Narrative	Speaker group	Mean percent of LWCs
Spoken formal	HS	30.5
Spoken formal	MS	21.8
Spoken informal	HS	24.5
Spoken informal	MS	31.8
Written formal	HS	19.4
Written formal	MS	5.4
Written informal	HS	11.8
Written informal	MS	13.8

4.2. LWCs across speaker groups and narratives

For the frequency of post-field LWCs, the model output (Appendix A) shows no significant difference between the two speaker groups ($z = -1.173$, $p = 0.241$). For the distribution of LWCs in the post-field across registers (i.e., settings and modes), the model output (Appendix B) shows a main effect of mode ($z = -4.677$, $p < 0.01$, Figure 2) with both speaker groups producing more post-field LWCs in spoken productions than in written productions, independently of the setting. The model additionally shows an interaction between speaker group and setting ($z = 3.226$, $p = 0.001$, Figure 3). To interpret this interaction, I ran Tuckey's multiple comparison test using the *emmeans* package (Lenth, 2020). Tuckey's multiple comparison test (Appendix C) revealed a significant difference between speaker group in the formal setting (estimate = 0.976, SE = 0.345, $z = 2.831$, $p = 0.024$) but no such difference in the informal setting (estimate = -0.559 , SE = 0.429, $z = -1.305$, $p = 0.56$). This indicates that HSs and MSs overlap in their frequency and distribution of post-field LWCs in the informal setting but not in the formal setting. Furthermore, Tuckey's multiple comparison test (Appendix C) also revealed a



significant difference in the setting of the MSs (estimate = -0.769 , $SE = 0.257$, $z = -2.99$, $p = 0.0148$, Figure 3). MSs produced significantly more post-field LWCs in the informal setting than in the formal setting. In the HSs data, there is no significant difference in the production of post-field LWCs across settings. This shows that while mode plays a role in the production of post-field LWCs across speaker groups, setting only has an influence on the productions of MSs.

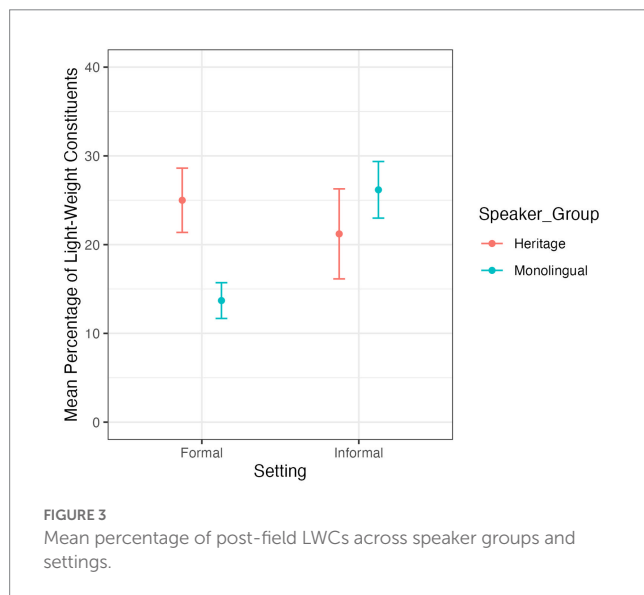


FIGURE 3
Mean percentage of post-field LWCs across speaker groups and settings.

4.3. PPs across speaker groups and narratives

For PPs in the post-field, the model output (Appendix D) shows no significant difference for the frequency of PPs between speaker groups ($z = -1.506$, $p = 0.132$, Figure 4). Hence, HSs and MSs do not differ significantly in their production of post-field PPs.

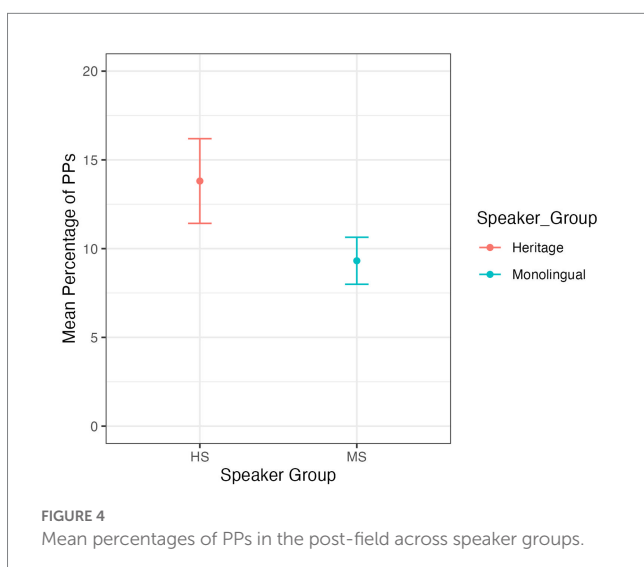


FIGURE 4
Mean percentages of PPs in the post-field across speaker groups.

4.4. Non-canonical placement of direct objects in the post-field

The corpus presents two instances of NONCs in the post-field which can be attributed to the influence of the ML, English on the HL, German. We find these two instances in both the formal spoken and the informal spoken productions of one HS (see example 10a/b²¹).

(10a) und die mann geht zu helfen²² [**die mädchen**_{NONC}] (–) **die essen** (–) **äh aufzuholen**²³ (RUEG corpus formal spoken)

“the man goes to help the girl pick up the food”

(10b) diese mann: geht zu helfen [**diese** (–) **de: de frau**_{NONC}] **die essen** (–) **au/(–) aufzuheben** (RUEG corpus informal spoken)

“this man goes to help this woman pick up the food”

The examples consist of two DPs and two infinitive clauses (INFs) each. In both cases, not only the direct object *die Mädchen* or *diese Frau* (the girl, this woman) but also the two infinitival constructions *zu helfen* (to help) and *die Essen aufzuholen/aufzuheben* (to pick up the food) are placed after the finite verb *geht* (goes). The extraposition of the second INF is not problematic and can be considered unmarked in German. Colloquially, the example sentences in (10a/b) could have been canonically produced as in example (10c).

(10c) der Mann geht der Frau helfen, **das Essen aufzuheben**.

“The man goes to help the woman pick up the food.”

What is problematic, and ungrammatical in German, however, is the switched position of the infinitive *zu helfen* and the direct object *die Mädchen* or *diese Frau*. As a consequence, the direct object surfaces post-verbally, where it would be expected in English. The influence of English is not only visible in the linearization of the constituents but also in how the infinitive is realized. In this case, due to the collocation *helfen gehen* (help go, go to help), the infinitival particle *zu* (to) must be left out.²⁴

It appears likely, then, that English provided the clausal matrix in these cases and that we are dealing with a calque. Support for this claim can be found in three corresponding English narrations of the very same speaker (see examples 11a–c).

²¹ The spoken and written productions in examples (10) and (11) were not corrected or normalized and the original orthography of the written productions was kept.

²² The undisrupted productions of the first infinitive construction *zu helfen* can be interpreted as a sign that the speaker does not question the fact that the matrix verb “help” needs to be produced with the particle *zu*. The second infinitive constructions *aufzuholen/aufzuheben* is accompanied by an increased number of non-verbal elements. Determining whether this is due to word finding issues or the production of the infinitive goes beyond the scope of this paper.

²³ These are transcriptions of the spoken data that include non-verbal discourse elements, such as pauses “(–)”, ruptures “/”, prolongations “:”, and hesitations “äh”.

²⁴ The German infinitive, *sui generis*, depends on the matrix verb. It can be realized as an infinitive without the particle *zu*, an infinitive with the particle *zu*, or an infinitive with the particle *um zu*. In examples (10a/b), the matrix verb “help” does not require the addition of the particle *zu* in German. An alternative canonical option would be *der Mann geht, um der Frau zu helfen, das Essen aufzuheben*. I am aware that this is a radically reduced explanation of the German infinitive, but it is merely to show the three options of infinitive-formation in German.

(11a) the man went to go help the lady pick up his food (RUEG corpus formal spoken)

(11b) the: guy he went to go help th(e)la(d)y pick (–) pick up the food (RUEG corpus informal spoken)

(11c) When he try to help the lady pick up her food (RUEG corpus informal written)

One further case of a seemingly highly marked LWC in the post-field is found in the formal written production of another HS (see example 12).

(12) Nichts ist passiert **zu die zwei Autofahrer**. (RUEG corpus formal written)

“Nothing happened to the two drivers.”

In German, *passieren* (happen) can be used with a dative complement with or without a PP (*etwas passiert (mit) jemandem_{DAT}*, something happens with to somebody/something happens to somebody). What makes the pattern in (12) look like a calque from English, at first sight, may just be due to the choice of *zu* instead of *mit* (with). Had the participant written *Nichts ist passiert mit den zwei Autofahrern*, one would simply consider it unusual in a written narrative.²⁵

5. Discussion

This study investigated the production of post-field LWCs in spoken and written productions of HSs and MSs of German, taking into account different registers. The goal was to determine how the two speaker groups deal with the options available to them under the same communicative tasks.

The first research question focused on types of LWCs produced in the post-field across speaker groups, and on their relative frequencies. The analysis of the data shows that, apart from two instances of clearly non-canonically placed direct objects in the post-field produced by one HS, all listed constituent types were found with overall similar frequencies in the post-field productions of both speaker groups. Hence, hypothesis 1, which stated that the productions of HSs will show a greater variety and a higher frequency of LWCs in the post-field, is not confirmed. HSs and MSs do not differ with respect to the frequency and variety of LWCs in the right sentence periphery. So, even though we are looking at an interface phenomenon, HSs adhere to German canonicity requirements: the head position in the VP and the placement of finite verbs in main and subordinate clauses, phenomena acquired early and relatively stable even under intensive language contact.²⁶

The second research question focused on the influence of register (i.e., different modes and settings) on the frequency of LWCs in the post-field. With respect to MSs, the data confirms hypothesis 2. Setting and mode had an influence on the production of post-field LWCs in the MS group. MSs produced significantly more post-field LWCs in the informal setting than in the formal setting and they produced significantly more post-field LWCs in the spoken mode than in the written mode. With respect to the HSs, the data just partly supports hypothesis 2. Only mode had an influence on the production of post-field LWCs in the HS group. HSs produced significantly more post-field LWCs in the spoken mode than in the written mode. However, the data shows no difference between post-field LWCs in the informal and the formal setting. Hence, while there is no group-specific difference in the overall frequency and variety of post-field LWCs, HSs and MSs show different distributions across registers, resulting in larger production differences between HSs and MSs in the written mode and in the formal setting. This result aligns with previous findings which observed register levelling across different phenomena in HSs (Polinsky, 2018, pp. 323–324; Tsehaye et al., 2021; Pashkova et al., 2022 among others) and can be traced back to HSs' limited exposure to communicative situations in their HL compared to their ML.

In order to test the influence of language contact and transfer more specifically, the third research question focused on the realization of PPs in the post-field. The goal was to investigate whether HSs of German produce more PPs in the post-field than MSs of German. The data does not confirm hypothesis 3, indicating that extensive contact with English does not lead to an increase in PP extraposition in HSs. This finding is not in line with the assumption that the availability of surface structure parallelism leads to an increase in converging patterns. Again, a possible explanation for this result might be that core syntactic features are acquired early both in monolingual children and simultaneous bilinguals (Müller and Hulk, 2000; Genesee, 2001; Gawlitzek-Maiwald and Tracy, 2005; Tracy, 2011 among others) and hence may prove to be particularly robust in HSs as well, even under increased contact with the ML and reduced contact with the HL. Another line of argumentation could be that we are witnessing language internal changes within German, with PPs being increasingly prone to extraposition among MSs.

The role of language contact and transfer was also addressed by a qualitative analysis of the two instances of NONCs in the post-field produced by a single speaker. The claim as to the influence of an English clausal pattern as the underlying matrix for these constructions has been corroborated by the English productions of this very speaker since they exhibit an identical pattern. These two instances, however, also indicate that even though a speaker produces non-canonical syntactic structures, these structures are systematic: they occur in two out of four German narrations and both times only in the spoken mode.

Concluding, we can say that the narrations produced by HSs and MSs exhibit different degrees of variation at the right sentence periphery. These differences, however, do not seem to be primarily due to bilingualism, language contact, or transfer, as we only find very marginal evidence (two cases in total) for NONCs in the post-field and no difference in PP productions. This finding is even more remarkable as we also find occasional non-canonically placed direct objects in the post-field productions of monolingually-raised

²⁵ The non-canonical preposition in this example changes the semantics of the verb *passieren*, which may result in different interpretations (happen to somebody vs. happen with somebody). In the present analysis, this constituent was categorized as a PP.

²⁶ Stability and retention of verb placement, but with considerable interindividual variation, have also been attested in research on German Language Islands, such as Pennsylvania German (Westphal Fitch, 2011), Moundridge Schweitzer German (Hopp and Putnam, 2015) or Texas German (Boas, 2009).

German children (Elsner, 2015). It is therefore the role of register variation or, rather, register-levelling that becomes apparent in the HSs data which leads to distributional differences between the two speaker groups.

Limitations of this study include the relatively small sample size of the different post-field constituents which did not allow for a more fine-grained quantitative analysis of the distribution of different types of LWCs. Moreover, the overall length of narrations per speaker and the constituents in the middle-field have not been taken into account. This could have influenced the results in two ways. Firstly, shorter, less detailed narratives provide less opportunity for the extraposition of constituents, plus the self-ratings of the HS group indicate lower proficiency in the written mode, which, in some cases, coincided with shorter written productions. Secondly, no conclusions about the overall number of constituents which have been placed in the post-field in proportion to those realized in the middle-field has been drawn. An additional limitation can be found in the research design. This study relied on the standardized elicitation of quasi-naturalistic productions and not on an experimental task geared to the elicitation of post-field items. Additionally, the elicitation task of recounting a car accident in as much detail as possible facilitated the production of LWCs in the post-field as participants tended to add further detail where they felt more information might be needed, like in the police report. Further research with different elicitation scenarios, including turn-taking, could enhance the production of a wider range of post-field LWCs and more diversified discourse functions.

6. Conclusion

This article investigated the linearization of constituents at the right sentence periphery in narrative productions of adolescent HSs of German and MSs of German. More specifically, the frequency of post-field LWCs in different registers was analyzed in order to shed further light on the variational spectrum found at the right clausal edge. Bilingualism, language contact, register variation, and internal dynamics were investigated as possible sources of variation. Analyses showed a similar variational spectrum of constituent types and their frequencies in HSs and MSs. Furthermore, HSs and MSs behaved similarly regarding the frequency and type of LWCs across modes, providing evidence that post-field LWCs are still more of a spoken phenomenon. The analyses for setting, however, showed effects of register-levelling in the HS group, as, unlike MSs, they did not differentiate between formal and informal settings. This suggests that diverging awareness of register norms due to different input conditions is the source of distributional differences observed rather than transfer from the dominant language.

Previous studies have considered PPs to be particularly affected by language contact and transfer. This, however, was not the case here, as the two speaker groups did not differ in their overall productions of PPs. But most importantly: While we find more variation in the right sentence periphery in different registers in the productions of HSs, the overall grammaticality of clausal syntax is not in jeopardy. Therefore, in the light of research on language change and language contact, we can say that the data discussed does not show evidence that heritage German is changing from an OV to a VO structure. Constituents placed right are still placed right.

Data availability statement

The data presented in this article is openly accessible via the RUEG corpus: <https://zenodo.org/record/5808870>.

Ethics statement

The studies involving human participants were reviewed and approved by the Deutsche Gesellschaft für Sprachwissenschaft ethics committee and the Institutional Review Board (IRB) of the University of Maryland at College-Park. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

The research results presented in this publication were funded by the German Research Foundation (DFG) as part of the research unit *Emerging grammars in language contact situations: a comparative approach* (FOR 2537) in project P5 (project no. 394995401, GZ TR 238/5-1). The publication of this article was funded by the University of Mannheim.

Acknowledgments

I thank project members and student assistants as well as everybody who took the time to support me with valuable feedback.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the author and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1122129/full#supplementary-material>

References

- Aalberse, S., Backus, A., and Muysken, P. (2019). *Heritage languages: a language contact approach*. Amsterdam: John Benjamins, vol. 58.
- Altmann, H. (1981). *Formen der "Herausstellung" im Deutschen: Rechtsversetzung, Linksversetzung, freies Thema und verwandte Konstruktionen*. Berlin, New York: Max Niemeyer Verlag.
- Andersen, C. (2008). *Topologische Felder in einem Korpus der gesprochenen Sprache. Probleme zwischen theoretischem Modell und Annotation*. 3, 1–15.
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Biber, D., and Conrad, S. (2001). "Register variation: a Corpus approach," in *The Handbook of Discourse Analysis*. eds. D. Schiffrin, D. Tannen and H. E. Hamilton (Hoboken: John Wiley & Sons, Inc.), 175–196.
- Boas, H.-C. (2009). *The life and death of Texas German*. Durham, NC: Duke University Press for the American Dialect Society.
- Bunk, O. (2020). "Aber immer alle sagen das" the status of V3 in German: use, processing, and syntactic representation. *PhD Dissertation*. Humboldt-Universität zu Berlin
- Bunk, O., Hamm, S., Kostka, J., Popova, G., Reinhold, N., Visser, E., et al. (n.d.): *KiDKo 2.0 Annotationsrichtlinien*.
- Clyne, M. (2003). *Dynamics of language contact*. Cambridge: Cambridge University Press.
- Coniglio, M., and Schlachter, E. (2015). Das Nachfeld im Deutschen zwischen Syntax, Informations- und Diskursstruktur: Eine diachrone korpusbasierte Untersuchung. *Das Nachfeld im Deutschen. Theorie und Empirie*.
- Drach, E. (1963). *Grundgedanken der deutschen Satzlehre*. Frankfurt: Moritz Disterweg.
- Elsner, D. (2015). "Das Nachfeld in der Kindersprache: Möglichkeiten und Grenzen einer konstruktionsgrammatischen Analyse," in *Das Nachfeld im Deutschen*. ed. H. Vinckel-Roisin (Berlin, München, Boston: De Gruyter), 345–361.
- Frey, W. (2015). "Zur Struktur des Nachfelds im Deutschen," in *Das Nachfeld im Deutschen: Theorie und Empirie*. ed. H. Vinckel-Roisin (Berlin, München, Boston: De Gruyter), 53–76.
- Freywald, U., Cornips, L., Ganuza, N., Nistov, I., and Opsahl, T. (2015). "Beyond verb second - a matter of novel information-structural effects? Evidence from Norwegian, Swedish, German and Dutch," in *Language, Youth and Identity in the 21st Century: Linguistic Practices across Urban Spaces*. eds. J. Nortier and B. Svendsen (Cambridge: Cambridge University Press), 73–92
- Fritzenschaft, A., Gawlitzek-Maiwald, I., Tracy, R., and Winkler, S. (1990). Wege zur komplexen syntax. *Z. Sprachwiss.* 9, 52–134. doi: 10.1515/zfsw.1990.9.1-2.52
- Gawlitzek-Maiwald, I., and Tracy, R. (2005). The multilingual potential in emerging grammars. *Int. J. Biling.* 9, 277–297. doi: 10.1177/13670069050090020801
- Genesee, F. (2001). Bilingual first language acquisition: exploring the limits of the language faculty. *Annu. Rev. Appl. Linguist.* 21, 153–168. doi: 10.1017/s0267190501000095
- Haider, H. (2010). *The Syntax of German (Cambridge Syntax Guides)*. Cambridge: Cambridge University Press.
- Hinterhölzl, R. (2004). "Language change versus grammar change," in *Diachronic Clues to Synchronic Grammar*. eds. E. Fuß and C. Trips (Amsterdam/Philadelphia: John Benjamins), 131–160.
- Hopp, H., and Putnam, M. T. (2015). Syntactic restructuring in heritage grammars. *Linguist. Approaches Biling.* 5, 180–214. doi: 10.1075/lab.5.2.02hop
- Imo, W. (2015). "Nachträge im Spannungsfeld von Medialität, Situation und interaktionaler Funktion," in *Das Nachfeld im Deutschen: Theorie und Empirie*. ed. H. Vinckel-Roisin (Berlin, München, Boston: De Gruyter), 231–254.
- Imo, W. (2016). *Grammatik: Eine Einführung*. Stuttgart: J.B. Metzler.
- Koch, P., and Oesterreicher, W. (2012). "Language of immediacy—language of distance: orality and literacy from the perspective of language theory and linguistic," in *Communicative Spaces: Variation, Contact, Change. Papers in Honour of Ursula Schaefer*. eds. C. Lange, B. Weber, and G. Wolf (Frankfurt: Peter Lang), 441–473.
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Lenth, R. (2020). Emmeans: estimated marginal means, aka least-squares means. R package version 1.4.7. (R package version 1.4.7). Available at: <https://cran.r-project.org/package=emmeans>
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.
- Müller, S. (2003). Mehrfache Vorfeldbesetzung. *Dtsch. Sprache* 31, 29–61. doi: 10.1515/9783111671956.177
- Müller, N., and Hulk, A. (2000). Bilingual first language acquisition at the interface between syntax and pragmatics. *Biling. Lang. Cogn.* 8, 52–78. doi: 10.1075/eurosla.8.06mul
- Müller, A., Schulz, P., and Tracy, R. (2018). "Spracherwerb," in *Konzepte zur Sprach- und Schriftsprachförderung entwickeln*. eds. C. Titz, S. Geyer, A. Ropeter, H. Wagner, S. Weber, and M. Hasselhorn (Stuttgart: Kohlhammer), 53–68.
- Muysken, P. (2000). Bilingual speech: a typology of code-mixing. *J. Linguist.* 39, 678–683. doi: 10.1017/s002226703272297
- Ortega, L. (2020). The study of heritage language development from a bilingualism and social justice perspective. *Lang. Learn.* 70, 15–53. doi: 10.1111/lang.12347
- Pascual Y Cabo, D., and Rothman, J. (2012). The (IL)logical problem of heritage speaker bilingualism 644 and incomplete acquisition. *Appl. Linguist.* 33, 450–455. doi: 10.1093/applin/ams037
- Pashkova, T., Tsehaye, W., Allen, S., and Tracy, R. (2022). Syntactic optionality in heritage language use: clause type preferences of German heritage speakers in a majority English context. *Herit. Lang. J.* 19. doi: 10.1163/15507076-12340022
- Polinsky, M. (2018). *Heritage languages and their speakers*. Cambridge: Cambridge University Press.
- Poplack, S. (1980). Sometimes ill start a sentence in Spanish y termino en español. *Linguistics* 18, 581–618. doi: 10.1515/ling.1980.18.7-8.581
- Prose, N. (2015). "Die Rolle komplexer Nachfeldbesetzungen bei der Einheitenbildung im gesprochenen Deutsch," in *Das Nachfeld im Deutschen. Theorie und Empirie*. Berlin: De Gruyter, 279–297.
- R Core Team (2021). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Rocker, M. (2022). Variation in finite verb placement in heritage Iowa low German: the role of prosodic integration and information structure. *PhD Dissertation*, Penn State University.
- Roelcke, T. (1997). *Sprachtypologie des Deutschen: historische, regionale und funktionale Variation*. Berlin, New York: Walter De Gruyter.
- Rothman, J. (2007). Heritage speaker competence differences, language change, and input type: inflected infinitives in heritage Brazilian Portuguese. *Int. J. Biling.* 11, 359–389. doi: 10.1177/13670069070110040201
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Duñabeitia, J. A., Gharibi, K., et al. (2022). Monolingual comparative normativity in bilingualism research is out of "control": arguments and alternatives. *Appl. Psycholinguist.* 44, 316–329. doi: 10.1017/s014716422000315
- Rothman, J., and Treffers-Daller, J. (2014). A prolegomenon to the construct of the native speaker: heritage speaker bilinguals are natives too! *Appl. Linguist.* 35, 93–98. doi: 10.1093/applin/amt049
- Rothweiler, M. (2006). "The acquisition of V2 and subordinate clauses in early successive acquisition of German," in *Interfaces in Multilingualism: Acquisition and Representation*. ed. C. Lleó (Amsterdam: John Benjamins), 91–113.
- Schulz, P., and Tracy, R. (2018). "Revisiting the tolerance of Universal Grammar," in *T.O.M. and Grammar. Thoughts on Mind and Grammar: A Festschrift in honor of Tom Roeper*. eds. B. Hollebrandse, J. Kim, J. A. T. Pérez-Leroux, and P. Schulz (University of Massachusetts, Graduate Linguistics Student Association. UMOF 41), 129–145.
- Sorace, A. (2011). Pinning down the concept of "interface" in bilingualism. *Linguist. Approaches Biling.* 1, 1–33. doi: 10.1075/lab.1.1.01sor
- Tracy, R. (1991). *Sprachliche Strukturentwicklung: Linguistische und kognitionspsychologische Aspekte einer Theorie des Erstspracherwerbs*. Tübingen: Narr.
- Tracy, R. (2011). "Konstruktion, Dekonstruktion und Rekonstruktion: Minimalistische und (trotzdem) konstruktivistische Überlegungen zum Spracherwerb" in *Sprachliches Wissen zwischen Lexikon und Grammatik*. eds. S. Engelberg, A. Holler and K. Proost (Berlin, Boston: De Gruyter), 397–428.
- Tsehaye, W., Pashkova, T., Tracy, R., and Allen, S. E. M. (2021). Deconstructing the native speaker: further evidence from heritage speakers for why this horse should be dead! *Front. Psychol.* 12. doi: 10.3389/fpsyg.2021.717352
- Tsimpli, I. M. (2014). Early, late or very late? *Linguist. Approaches Biling.* 4, 283–313. doi: 10.1075/lab.4.3.01tsi
- Vinckel-Roisin, H. (2012). "Das 'Nachfeld' im Deutschen: Rechte Satzperipherie und Diskurstopik-Auszeichnung," in *Studia Linguistica XXXI*. Wrocław: TOTEM, 143–163.
- Vinckel-Roisin, H. (2015). *Das Nachfeld im Deutschen: Theorie und Empirie*. Berlin, München, Boston: De Gruyter.
- Westphal Fitch, G. (2011). "Changes in frequency as a measure of language change," in *Studies on German-Language Islands*. ed. M. T. Putnam (Amsterdam: John Benjamins Publishing Company), 371–384
- Wiese, H. (2020). "Language situations: a method for capturing variation within speakers' repertoires," in *Methods in Dialectology XVI*. ed. A. Yoshiyuki (Frankfurt a. M.: Peter Lang), 105–117.
- Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., et al. (2021). *RUEG Corpus (0.4.0) [Data set]*. Zenodo.

Wiese, H., Alexiadou, A., Allen, S., Bunk, O., Gagarina, N., Iefremenko, K., et al. (2022). Heritage speakers as part of the native language continuum. *Front. Psychol.* 12. doi: 10.3389/fpsyg.2021.717973

Wiese, H., and Müller, H. G. (2018). "The hidden life of V3: an overlooked word order variant on verb-second" in *Non-canonical verb positioning in main clauses*. eds. M. Antomo and S. Müller (Hamburg: Helmut Buske), 201–224.

Wöllstein, A. (2014). *Topologisches Satzmodell*. Heidelberg: Winter.

Zifonun, G. (2015). "Der rechte Rand in der IDS-Grammatik: Evidenzen und Probleme," in *Das Nachfeld im Deutschen: Theorie und Empirie*. ed. H. Vinckel-Roisin (Berlin, München, Boston: De Gruyter), 25–52.

Zifonun, G., Hoffmann, L., and Strecker, B. (1997). *Grammatik der deutschen Sprache*. Berlin, New York: De Gruyter.



OPEN ACCESS

EDITED BY
Fatih Bayram,
UiT The Arctic University of Norway, Norway

REVIEWED BY
Anna Mikhaylova,
The University of Queensland, Australia
Ayhan Aksu-Koç,
Boğaziçi University, Türkiye

*CORRESPONDENCE
Suzan D. Tokaç-Scheffer
✉ sdtokac@gmail.com

RECEIVED 14 October 2022
ACCEPTED 26 July 2023
PUBLISHED 22 September 2023

CITATION
Tokaç-Scheffer SD, Arslan S and Nickels L
(2023) Insights into the time course of
evidentiality processing in Turkish heritage
speakers using a self-paced reading task.
Front. Commun. 8:1070510.
doi: 10.3389/fcomm.2023.1070510

COPYRIGHT
© 2023 Tokaç-Scheffer, Arslan and Nickels. This
is an open-access article distributed under the
terms of the [Creative Commons Attribution
License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or
reproduction in other forums is permitted,
provided the original author(s) and the
copyright owner(s) are credited and that the
original publication in this journal is cited, in
accordance with accepted academic practice.
No use, distribution or reproduction is
permitted which does not comply with these
terms.

Insights into the time course of evidentiality processing in Turkish heritage speakers using a self-paced reading task

Suzan D. Tokaç-Scheffer^{1,2,3*}, Seçkin Arslan⁴ and Lyndsey Nickels³

¹International Doctorate for Experimental Approaches to Language and Brain (IDEALAB), University of Groningen, Netherlands/University of Newcastle, United Kingdom/University of Potsdam, Germany and Macquarie University, Sydney, NSW, Australia, ²Center for Language and Cognition Groningen (CLCG), University of Groningen, Groningen, Netherlands, ³School of Psychological Sciences, Macquarie University, Sydney, NSW, Australia, ⁴Université Côte d'Azur & CNRS, Nice, France

Introduction: Studies with heritage language speakers (HLS) have often used offline measurements, investigating the post-interpretive effects which emerge after processing has been completed. Relatively few studies have investigated heritage language processing using time-sensitive methods that allow the collection of evidence regarding real-time language processing rather than post-interpretive judgments. Using a self-paced-reading paradigm, we aimed to expand our understanding of HLS language processing by investigating evidentiality-the linguistic marking of information source, which is grammatically expressed in Turkish, but not in English.

Method: Participants were 54 bilingual speakers of Turkish and English: 24 HLS (English onset: 0-5 yrs) and 30 emigrant Turkish speakers (ES) who grew up in Turkey before emigrating to Australia (English onset = 6-17 yrs). Participants read sentences with evidential-marked verb forms that either matched or mismatched to the information source context. Word-by-word reading times and end-of-sentence acceptability judgment speed and accuracy were measured.

Results: The results showed that although the HLS' responses were slower and less accurate than the ES in both reading times and end-of-sentence acceptability judgments, they showed similarities in online processing patterns. Both groups were faster at reading the mismatching sentences compared to the matching sentences; however, this pattern emerged during the time course of reading first for the indirect condition for the ES, and only later for the direct condition and for the HLS for both evidential conditions. Only HLS read faster in the target region with the direct evidential that is shown to be acquired earlier in childhood, than they did for the indirect evidential which is mastered later. In contrast, the end-of-sentence judgment data showed that while the ES group responded faster to matching direct sentences than matching indirect, this effect was missing for the HLS. Nevertheless, there were similar patterns for accuracy across evidential conditions: both groups were more accurate with the direct evidential.

Discussion: Overall, the use of the self-paced-reading paradigm allowed insights into HLS' evidentiality processing above and beyond their generally slower and less accurate processing compared to the reference group. This study provides further evidence for differences in the patterns observed using online vs. post interpretive measures in HLS, reinforcing the importance of combining these methodologies for further understanding of HLS competence and performance.

KEYWORDS

heritage language speakers, evidentiality, Turkish, sentence comprehension, Turkish-English bilinguals, self-paced reading (SPR)

Introduction

Most studies conducted with heritage language speakers report results from offline tasks including paradigms tapping sentence comprehension (sentence-picture matching, Montrul et al., 2008), oral production (structured elicitation, Bayram et al., 2017; picture naming, Hulsen et al., 2002; storey-telling (retelling), Montrul, 2002, 2004; Polinsky and Kagan, 2007; Polinsky, 2008; Montrul and Sánchez-Walker, 2013), written production (Montrul, 2002; Montrul et al., 2008), sentence judgement tasks (aural acceptability judgement task, Fuchs et al., 2015; sentence conjunction judgement task, Montrul, 2009; written acceptability judgement task, Montrul and Bowles, 2009; context/sentence-matching task, Rothman, 2007). These tasks measure the competence of heritage speakers in terms of whether they are aware of the grammatical rules in the language. In contrast, online methods (e.g., self-paced reading, eye-tracking, EEG) allow the measurement of real-time processing, which is not possible to investigate with discrete per-item responses (Felsner et al., 2009; Clackson et al., 2011; Felsner and Cummings, 2012; Lago et al., 2018). According to Keating and Jagerski, the real-time component of online methods helps “tap participants’ implicit knowledge of language” (2015, p. 2) since they measure processing and knowledge as it happens rather than giving speakers a chance to evaluate, deduct and make a decision based on their learned knowledge. In bilingualism research, online measures have been shown to have an advantage for disentangling sentence processing mechanisms (Keating and Jegerski, 2015).

Here we focus on self-paced reading. During a self-paced reading task, language users are presented with a single segment consisting of a word or a phrase, which disappears on command (usually a button press on the keyboard) allowing a new segment to appear. The time passing between each command/button press gives an indication of the processing load or cost of the segment on the language user. This includes enabling identification of the point in the sentence that participants may encounter processing difficulties and how long these difficulties persist. These indicators may reflect increased cognitive load, or mental effort required to process the sentence (Just and Carpenter, 1980). Consequently, as an online method, self-paced reading enables recording of segment-by-segment reading times when a reader is presented with each word or group of words in a sentence, and how they process and react to them (Chen et al., 2005; Keating and Jegerski, 2015).

There is extensive research using self-paced reading tasks with monolingual speakers (e.g., Carminati, 2002; Filiaci, 2011; Xu et al., 2018; Lee and Fraundorf, 2022) and second language learners (for review see Nicklin and Plonsky, 2020) examining the processing of various linguistic phenomena. Self-paced reading experiments with heritage language speakers are, on the other hand, relatively scarce. However, those studies that have been carried out found that although heritage language speakers were slower and/or less accurate than reference groups (monolinguals and/or late bilinguals and second language speakers), their processing patterns had qualitative similarities and showed differences to those measured with offline tasks (Keating et al., 2016; Jegerski, 2018a,b; Mikhaylova, 2018; Di Pisa et al., 2022). For example, Keating et al.

(2016) investigated monolingual and heritage language speakers’ antecedent choice for ambiguity resolution between null vs. overt pronouns in Spanish. They used an online self-paced reading task with sentences like *Cuando la diva visitó a la directora, Oella ofreció cantar un aria en italiano* (translation: When the diva visited the director, \emptyset /she offered to sing an aria in Italian). In Spanish it is more accurate to assign the null pronoun (\emptyset) to the word in the subject position (i.e., *diva*); whereas the overt pronoun *she* is preferentially attributed to the *director* which is the object of the preceding clause (Keating et al., 2016; Supplementary material). An earlier offline study found heritage language speakers showed no such processing preferences for null vs. overt pronouns (Keating et al., 2011). In contrast, in Keating et al. (2016) self-paced reading study, heritage language speakers did show a preference for attributing null pronouns to antecedents in the subject position that is shown by monolingual native speakers. However, they did not show a preference for overt pronouns. Nevertheless, the key point here is that they showed dissimilar processing of null vs. overt pronouns during this online task. In contrast, in the responses to end-of-sentence comprehension questions (Keating et al., 2016), heritage language speakers showed no preference between null and overt pronouns, just as they had not in the earlier offline study (Keating et al., 2011). Although the heritage language speakers’ processing was not completely parallel to that of monolinguals, this study clearly illustrates that tasks tapping online and offline processing can provide different insights and thereby underlines the importance of contrasting experimental methods.

Jegerski (2018b) also reports a study that supports the utility of self-paced reading tasks, for identifying which linguistic phenomena are challenging for all speakers of that language and which are only challenging for heritage language speakers. They tested heritage language speakers’ *Differential Object Marking* (DOM) in Spanish compared to a group of late Spanish-English bilinguals using an online self-paced reading task interspersed with an offline end-of-sentence acceptability judgement task. In the offline, end-of-sentence acceptability judgements, heritage language speakers were less accurate and slower than the late bilinguals and did not show any differences between the conditions. This result was similar to that of Montrul and Bowles (2009) finding, also from an offline acceptability judgement task, which showed that Spanish heritage language speakers could not distinguish sentences that were ungrammatical for DOM from grammatical sentences. However, Jegerski (2018b) found that during online self-paced reading, both groups showed slower RTs for the ungrammatical DOM of the inanimate direct object, but no sensitivity to the ungrammatical omission of DOM for animate direct objects (the only condition where a direct object can be marked with “a” in Spanish) (Jegerski, 2018b). Indeed, they report similar online sensitivity in both heritage language speakers and late bilinguals that highlighted the fact that the variability in DOM processing could not be attributed to incomplete attainment of DOM markers. This study particularly highlights the fact that differences between heritage language speakers and a reference group of bilingual speakers were no longer apparent during online processing. Consequently, these studies demonstrate how the self-paced reading task can provide additional information to

facilitate the unravelling of heritage language speakers' processing. It is important to note, that, in these studies, a phrase-by-phrase presentation was employed. However, it has been suggested that a word-by-word presentation would allow a more refined analysis in terms of time-course of processing as it does not collapse across several words, and therefore gives smaller time windows (Keating et al., 2016).

The studies of self-paced reading cited above found heritage language speakers to show some qualitative similarity to reference groups and/or provided more insights into heritage language speakers' processing. Consequently, we were interested in whether the same would hold for heritage language speakers' processing of evidentiality in Turkish, especially given that in the primarily offline measures used in earlier studies, heritage language speakers showed slower reaction times and lower accuracy compared to monolingual (Arslan et al., 2017) and bilingual (Schmid and Karayayla, 2019; Tokaç-Scheffer et al., to appear) reference groups.

Evidentiality is the specification of how a speaker received the information in their utterance (Aikhenvald, 2004). Information can be received through different sources (i.e., visually, aurally, from a third person, etc.) and the evidential markers available in a particular language are used to specify this source. In some languages, such as Indo-European languages like English, evidential meanings are conveyed by means of lexical elements, such as *I saw* for direct visual evidence, or *I have been told* or *I assume* for inferred or reported knowledge (e.g., *I saw John ate the apple yesterday* vs. *I was told/I assume that John ate the apple yesterday*). These forms are optional, yet they indicate the amount of evidence for a speaker's assertion (De Haan, 1999). However, in a quarter of the world's languages, evidentiality is a grammatical unit and specification of the evidence type is obligatory in one's utterance (Aikhenvald, 2004). Turkish, the heritage language under investigation in this paper, is one such language and, in Turkish, it is obligatory to use evidential markers when referring to the past. Evidentiality in Turkish is marked as a verb inflexion that indicates the source of a past event: the evidential marker specifies whether the speaker witnessed and/or personally carried out the action firsthand or received that information non-firsthand, as in hearsay or inference (Aikhenvald, 2004; De Haan, 2005). In the case of firsthand information, the direct evidential marker *-DI*¹ is used. For example, in "*Bahçivan çiçekleri suladı*," (I know/saw that) the gardener watered the plants. Information that is non-firsthand is marked with the indirect evidential marker *-mİ*: "*Bahçivan çiçekleri sulamış*," (I infer it from the wet plants or someone else told me that) the gardener watered the plants.

Acquisition studies conducted with monolingual Turkish children have shown that children start producing evidentiality in their utterances very early (e.g., 1,5 years in Aksu-Koç et al., 2009); and that the acquisition of the indirect evidential marker follows the direct evidential marker (Aksu-Koç, 1988; Aksu-Koç et al., 2009). It has also been shown that it may take up to the age of seven for children to fully master the distinctions

between the evidentiality markers (Öztürk and Papafragou, 2008; Ünal and Papafragou, 2016). This makes evidentiality a "late-mastered" language component and late-mastered linguistic items have been shown to be challenging for heritage language speakers (e.g., Montrul, 2002; Montrul et al., 2008; Polinsky, 2008).

Tokaç-Scheffer et al. (to appear) conducted an offline study that is of particular relevance to the study reported here. In this offline study, we compared the evidentiality processing of two groups of bilingual Turkish speakers, namely heritage language speakers of Turkish and bilingual Turkish-English speakers who had migrated from Turkey using an (offline) auditory sentence verification task (see also Arslan et al., 2017). The task was to listen to sentences presented in the heritage language and to respond as fast as possible whenever a word was detected that made the sentence unacceptable and/or semantically inappropriate. Tokaç-Scheffer et al. (to appear) found that the heritage language speakers were slower and less accurate in detecting these unacceptable sentences compared to emigrant speakers. Further analyses revealed that neither of the groups showed better processing for either evidential condition in their reaction times. However, emigrant speakers performed better in the indirect evidential condition. They rejected the sentences with the violation of a direct information source by the use of indirect evidential marker more accurately compared to when the indirect information source was followed by a mismatching use of direct evidential. This pattern was not shown by the heritage language speakers.

Karaca (2018), in an unpublished Master's thesis, also presented a relevant study using a self-paced listening task to compare three groups of Turkish speakers: a bilingual group of heritage language speakers of Turkish born in, or moved to, Canada before the age of 5, a bilingual group of first-generation migrants from Turkey to Canada and, a monolingual group of Turkish speakers residing in Turkey. Karaca administered a self-paced listening task, where participants listened to dialogues including congruent and incongruent evidentiality sentences and answered yes/no comprehension questions (for a quarter of the stimuli). While listening to the critical segment that included the verb marked with indirect evidential, monolinguals showed differences between incongruent and congruent conditions (and sustained this effect in the following segment): they were slower in the incongruent conditions for both evidentials. In contrast, the heritage language speakers showed no processing differences at any point. However, the migrant speakers mirrored the effect that the monolinguals showed for the indirect evidential sentences but only on the third segment—a delayed effect (Karaca, 2018), suggesting reduced processing speed for bilingual individuals who have experienced attrition in their first acquired language. In the third segment monolingual speakers also showed a difference between congruent and incongruent sentences for the direct evidential; such an effect was missing both for heritage language speakers and migrant speakers. However, Karaca argued that without the presentation of a fourth segment, we cannot exclude the possibility of heritage language speakers and/or migrant speakers showing a late effect. Karaca did not present the results for the comprehension questions. In sum, during this online study monolingual speakers showed processing differences between

¹ The capitalization of the letters indicates that the sounds represented follow harmonization rules in Turkish, i.e., vowel harmony and consonant assimilation rules. As an agglutinative language, in Turkish sounds may be modified when appended through suffixation.

congruent and incongruent evidential conditions and rejected particularly quickly the incongruent indirect evidential marker use early in processing. The migrant speakers showed some similarities to the monolingual speakers while heritage language speakers showed no evidence of evidentiality processing preferences.

In the present study, we aimed to further explore the online and temporal processing of evidentiality of heritage language speakers relative to a reference group of migrant bilingual speakers by using a self-paced reading task with longer stimuli presented with a word-by-word moving window paradigm.

Specifically, we addressed the following research question:

- (1) Does the online processing of Turkish heritage language speakers and Turkish speakers who are also late second-language speakers of English (emigrant Turkish speakers) differ during the time course of reading sentences with evidentiality marking?

Previous studies using self-paced reading have shown heritage language speakers can pattern with late bilinguals in their sensitivity during sentence reading, even when they differ in offline end-of-sentence judgement responses (Jegerski, 2018b). Therefore, although, in other studies with heritage language speakers of Turkish (e.g., Arslan et al., 2015, 2017; Karayayla and Schmid, 2019; Tokaç-Scheffer et al., to appear) heritage language speakers showed different patterns of evidentiality processing to bilingual emigrant Turkish speakers, we hypothesised that using an online task may reveal qualitative similarities between these speakers and thereby gain more information on the time course of evidentiality processing by heritage language speakers. We would nevertheless expect heritage language speakers to process the sentences more slowly than the emigrant Turkish speakers, given the differences in their Turkish exposure.

- (2) Does the online processing of evidentiality by Turkish heritage language speakers and emigrant speakers of Turkish differ depending on the evidentiality distinction -direct vs. indirect? Do the two groups differ in this regard?

Based on previous research showing that the direct evidentiality condition is mastered earlier in children's acquisition and that the indirect evidential has more semantic connotations, and is therefore cognitively more complex, we expected both groups to show better processing of direct evidential which would be manifested in shorter reading times for the verbs marked with the direct evidential compared to the indirect evidential (Aksu-Koç, 1988; Öztürk and Papafragou, 2008; Aksu-Koç et al., 2009; Ünal and Papafragou, 2016).

- (3) Are there differences between online and offline processing of evidentiality for Turkish heritage language speakers and emigrant speakers of Turkish? Do the two groups differ in this regard?

Similar to the previous studies described above (e.g., Keating et al., 2016), we expect online and offline tasks to show differences in the processing of evidentiality, with more similarities between heritage language speakers and emigrant speakers of Turkish during online processing.

Materials and methods

Participants

A total of 53 bilingual speakers of Turkish and English ($M_{AGE} = 34.4$ years; $SD = 9.1$; 28 Female; four left-handed) all of whom resided in Sydney, Australia participated in this study. All participants performed this self-paced reading task first. The participants were recruited via student organisations, Turkish cultural and language centres, the Turkish Consulate Sydney, schools that deliver Turkish education, and advertisements posted on social media and in neighbourhoods with large Turkish communities. Inclusion criteria were that participants were 20–54 years of age, had to have started acquiring Turkish from birth, used English actively in daily life, and had no previous psychological, neurological, or communication disorders. The participants were given a bilingual language background questionnaire constructed based on the Language and Social Background Questionnaire (LSBQ; Anderson et al., 2018) with adaptations tailored to our research questions. These adaptations included editing sub-sections such as the community language use behaviour section (now language use behaviour) and amending the scaling method for life stages to use average percentages of exposure/use. In addition, reading was expanded into a whole section (reading habits) to collect more detailed input on participants' reading behaviours given that we planned to use a reading task. We excluded questions from the language use section that were more detailed than needed for our study (i.e., language use for social, religious, extracurricular activities; shopping/restaurant/other commercial services; health care services/government/public offices/banks) and instead included only languages preferred at home, work, social life and in general). The final questionnaire consisted of four sections: social background (education, occupation, parents' language history, countries they had lived, etc.), language background, language use behaviour, and reading habits. Table 1 provides a summary of the outcomes of the questionnaire.

The participants were divided in two groups according to their heritage language situation and characteristics: heritage language speakers of Turkish and emigrant speakers.

Heritage language speakers ($n = 23$)

Twenty-three heritage language speakers of Turkish, speaking both Turkish and English from early childhood, were recruited. Their age ranged between 20 and 45 years. While the age onset of Turkish was always from birth in this group, the onset for English ranged between birth and 5 years of age, and these participants were either born in Australia or migrated there at a very young age (i.e. at or before the age of 5). Although they spoke Turkish as their "home" language and it was their first learned language, most of them learned to read and write in Turkish after they had acquired these skills in English. They began acquiring English in kindergarten in Australia and received all their education in English.² They were

² Nevertheless, many children from Turkish backgrounds have access to some Turkish schooling in Australia. Most of the participants reported here either received Turkish instruction at school or attended "Saturday Schools," at which they carried out activities in Turkish once a week for a couple of

TABLE 1 Summary of critical outcomes from bilingualism language background questionnaire for both groups; heritage language speakers and emigrant speakers.

	HLS		ES		Welch <i>t</i> -test (<i>p</i> -value)	95% CIs
	Mean	SD	Mean	SD		
Age (years)	30.17	8.73	37.70	8.16	3.20 (0.002)*	[2.79, 12.26]
Education (years)	16.17	1.75	16.67	2.90	0.76 (0.447)	[−0.80, 1.79]
Years of residence in Australia	28.04	9.31	10.46	8.59	−7.04 (<0.001)*	[−22.61, −12.56]
Age of bilingualism Onset	1.00	1.98	11.53	1.96	19.29 (<0.001)*	[9.43, 11.63]
Turkish proficiency (self-rated)	7.91	1.52	9.90	0.28	6.22 (<0.001)*	[1.33, 2.66]
English proficiency (self-rated)	9.88	0.41	8.24	1.60	−5.39 (<0.001)*	[−2.26, −1.02]
Turkish exposure and use (current %)	30.02	18.20	33.00	19.00	0.54 (0.590)	[−7.55, 13.11]
English exposure and use (current %)	69.78	18.18	67.00	19.00	−0.54 (0.590)	[−13.11, 7.55]
Turkish material—audio and video (hr/day)	1.72	1.26	1.75	2.07	0.07 (0.947)	[−0.89, 0.96]
English material—audio and video (hr/day)	4.26	5.87	2.92	1.66	−1.06 (0.296)	[−1.24, 3.74]
Turkish material—written (current %)	19.70	17.24	33.33	22.45	2.50 (0.016)*	[2.69, 24.58]
English material—written (current %)	80.30	17.24	66.66	22.45	−2.50 (0.016)*	[−24.58, −2.69]
Turkish material—written (hr/week)	3.00	2.62	10.98	9.09	4.57 (<0.001)*	[4.43, 11.53]
English material—written (hr/week)	22.73	19.07	25.30	19.40	0.48 (0.632)	[−8.15, 13.27]

SD, Standard Deviation.

Values of significant effects ($p < 0.05$) are printed in bold and asterisked.

exposed to Turkish at home and within the Turkish community they lived in and in social environments, but their use of English became dominant over time.

Emigrant speakers (n = 30)

The emigrant speakers (age range 23–54 years) comprised participants who were raised in Turkey during their childhood and emigrated to Australia due to professional or educational circumstances at or after adolescence. They were non-native speakers of English who had started learning this language between the ages of 6 and 17 (mean = 11.50; SD = 2.0), at school in Turkey as a second/foreign language. The duration of their residence in Australia was from 1 month up to 31 years (see Table 1 for details). As skilled migrants, most of them had an upper intermediate level of English and reported using English more than Turkish since they moved to Australia.

Stimuli

The stimuli comprised 134 sentences. There were 104 target evidentiality sentences (26 in each of the four conditions) and 30 filler sentences. Twenty-six unique verbs referring to different actions were selected and each verb was inserted into a sentence frame which was then adapted for each of the four conditions of the evidentiality manipulation as described below.

hours. One participant reported having been home schooled in Turkish by her/his parents, and another participant had received Turkish as a second language lessons during her regular schooling.

Evidentiality sentences were 12 words long. The first three words constituted the contextual support which included a statement of the information source, specifically whether it was firsthand or nonfirsthand. If the information source was *firsthand*, it indicated the event was witnessed by the speaker her/himself (using *ben gördüğüme eminim*; “I am sure that I saw”). A *nonfirsthand* information source, on the other hand, specified the event was witnessed by others and that they transferred this knowledge, meaning that the speaker had heard about the event from others (*başkaları gördüğünü söylüyor*; “others say they saw it”). See Table 2 for example sentences.

The statement of the information source (contextual clause) was followed by the critical clause, which included the target verb region (R-TV)—the verb inflected with the evidentiality marker. The evidential marker on the verb was either *direct* or *indirect* (condition) and either matched or did not match the information source (firsthand or nonfirsthand) in the contextual clause. In the *match* sentences, the evidentiality marker appended on the main verb of the critical clause matched the preceding information source: the direct evidential marker (-DI) was used in firsthand information source sentences (*firsthand—direct*; see Table 2) and the indirect evidential marker (-mİş) was used in the nonfirsthand information source conditions (*nonfirsthand—indirect*). In the *mismatch* sentences the evidential marker on the verb did not match with the information source: Following a nonfirsthand information source the direct evidential marker (-DI) was given (*nonfirsthand—direct**) and following a firsthand information source the indirect evidential marker (-mİş) was provided (*firsthand—indirect**).

The last part of the sentence was the padding phrase (last five words) which included the spillover region (divided into two

TABLE 2 Example of evidentiality sentences used in the self-paced reading task.

	Condition	Contextual support	Critical clause	Padding phrase
Direct	Firsthand-direct	Ben gördüğüme eminim,	Mehmet ceketinin düğmesini kopardı	bu sebepten terziye gitmesi gerekecek.
		I see.DIRECT EVID.1SG sure.1SG	Mehmet jacket.POSS.GEN button.POSS.DEF pull off.DIRECT EVID.3SG	this reason.ABL tailor.DAT go.MOD require.FUTURE
		‘I am sure I saw Mehmet pull off (witnessed) the button of his jacket, that’s why he will need to go to the tailor.’		
	Nonfirsthand-direct*	Başkaları gördüğünü söylüyor,	Mehmet ceketinin düğmesini *kopardı	bu sebepten terziye gitmesi gerekecek.
		Others see.DIRECT EVID.1SG say.3PL	Mehmet jacket.POSS.GEN button.POSS.DEF *pull off.DIRECT EVID.3SG	this reason.ABL tailor.DAT go.MOD require.FUTURE
		“Others say they saw Mehmet *pull off (witnessed) the button of his jacket, that’s why he will need to go to the tailor.”		
Indirect	Nonfirsthand-indirect	Başkaları gördüğünü söylüyor,	Mehmet ceketinin düğmesini koparmış	bu sebepten terziye gitmesi gerekecek.
		Others see.DIRECT EVID.1SG say.3PL	Mehmet jacket.POSS.GEN button.POSS.DEF pull off.DIRECT EVID.3SG	this reason.ABL tailor.DAT go.MOD require.FUTURE
		“Others say they saw Mehmet pull off (reportedly) the button of his jacket, that’s why he will need to go to the tailor.”		
	Firsthand-indirect*	Ben gördüğüme eminim,	Mehmet ceketinin düğmesini *koparmış	bu sebepten terziye gitmesi gerekecek.
		I see.DIRECT EVID.1SG sure.1SG	Mehmet jacket.POSS.GEN button.POSS.DEF *pull off.DIRECT EVID.3SG	this reason.ABL tailor.DAT go.MOD require.FUTURE
		“I am sure I saw Mehmet *pull off (reportedly) the button of his jacket, that’s why he will need to go to the tailor.”		

Examples 1 and 3 are Match conditions; Examples 2 and 4 are Mismatch conditions (indicated with *).

spillover regions R-SO1 and R-SO2, two words in each region) and the final word region (R-FW). This padding phrase (last five words) included extra information regarding the event to enhance the clarity and comprehension of the overall meaning conveyed in the text. This phrase also allowed us to observe potential delays in processing since in self-paced-reading, effects may carry over to next segments. These phrases were presented in the present continuous or simple future tense to avoid any confusion with the time of the event (past).

The 30 filler sentences which also included morphosyntactic mismatches were sentences without specification of an information source. Half of the sentences were ungrammatical, created either by person/number disagreements or semantically incorrect verb choice (for the full list of stimuli see [Supplementary material](#)). The number of fillers (30) was determined based on the number of unique verbs (26) selected for the evidentiality sentences. These 26 verbs were then manipulated across four conditions, resulting in a total of 104 experimental sentences. Given the relatively large number of experimental sentences, to prevent fatigue, we decided not to add additional filler sentences. Furthermore, this self-paced reading experiment was only one part of a longer study, which in total took 2 h for participants to complete.

The 26 critical action verbs used to construct the sentence stimuli described above were chosen from a larger set of stimuli, that were normed for surface frequency, cloze probability of the evidential verb (see [Tokaç-Scheffer, 2023](#) for further details).

Procedure

The sentence materials were programmed in a non-cumulative self-paced reading design with end-of-sentence acceptability judgement ([Just et al., 1982](#)) using the web platform Ixex Farm ([Drummond, 2013](#)). The sentences were presented in black font

(96px) on a white background. The stimulus sentence advanced segment-by-segment with each press of the SPACE button in a *moving-window paradigm*. The first segment always contained the contextual clause where the information source was presented (consistently including 3 words; see [Table 2](#)) followed by the critical sentence material presented per word *per segment*. The *uninformative mask* technique was employed to the sentences with the word boundaries shown on the screen. With the first press, the information source for each sentence was presented as a single chunk of three words (e.g. *başkaları gördüğünü söylüyor*; “others say they saw it”) at the beginning of the sentence. Following a space bar press, this first segment disappeared from view, and the next word appeared to the right, such that only one segment (comprising a single word) was visible at any one time. At the end of each sentence participants were required to judge whether the sentence was “grammatically coherent.” After the answer, the next sentence appeared automatically.

The experiment started with an explanation of the task and what was expected. Participants were shown the first practise item and instructions in Turkish were given with a demonstration of the first practise item “This is a sentence. You will read each word by pressing the space key on the keyboard. After each press, a word will appear and with the next press that word will disappear and the next one will appear. You will see each word and consequently each sentence only once and will not be able to go back. This long line here [pointing to a line that indicates where the information source phrase will appear] gives you [relevant] information and then the rest of the sentence will follow. At the end of each sentence, you will decide if this sentence was grammatically coherent or not by choosing the smiley face emoji for *yes* and sad face emoji for *no* that will appear on the screen which corresponds to the *f* and *j* keys, respectively, on the keyboard.” They were not given any instructions regarding the speed of reading or responding (to the judgement questions). Testing started after four example

trials. Each participant was presented with sentences in all four conditions together with the filler items, as described in Table 2. The presentation order of the sentences was randomised for each participant. Participants were given the opportunity to have breaks when needed.

Data pre-processing and analyses

All data pre-processing and analyses were conducted in R studio version 1.2.5 (R Core-Team, 2012). First, sub-datasets were created for the analysis of each region of interest and acceptability question answers. Accordingly, five separate datasets were created: four for the analyses of reading times included the following regions: Target Verb (R-TV; target verb inflected with the evidentiality marker), Spillover 1 (R-SO1; comprising the individual reading times to the two words following the target verb), Spillover 2 (R-SO2; comprising the individual reading times to the next two words following the R-SO1), and Final Word (R-FW; the final word of the sentence seen before the presentation of the sentence acceptability judgement). A separate dataset was created for the reaction time analyses for the Sentence Judgements (SJ-RT; the acceptability question after the presentation of each sentence), and for the analyses of the sentence judgement accuracy (SJ-Acc; accuracy for the sentence judgements). Each dataset consisted of 5512 data points initially (26 sentences * 4 conditions * 53 participants). For data cleaning procedures, we followed Nicklin and Plonsky (2020), which presents a comprehensive overview of data pre-processing practises in bilingual studies and adopted those that fitted our population and research objectives. Specifically, we started with a visual inspection of the data and looked at histograms, boxplots, and Q-Q plots (see Supplementary material) to determine appropriate cut-offs to exclude outliers: Trials in the reading time data sets that were faster than 100 ms (e.g., Luce, 1991; Jegerski, 2016; Litcofsky and Van Hell, 2017; Kim et al., 2018) and slower than 8,000 ms (R-TV and R-SO) and 15,000 ms (R-FW) were excluded.³ Eight trials at R-TV, 12 at R-SO1 (6th word – 8 trials; 7th word – 3 trials), 14 at R-SO2 (8th word – 8 trials; 9th word – 6 trials), and 40 at R-FW regions were excluded. Lower and upper boundaries for SJ reaction times were 100 and 15,000 ms respectively, which resulted in the exclusion of 46 data points. This pre-processing resulted in the exclusion of <1% of the data for each dataset. The accuracy analyses were computed on the same data set that was created for the SJ-RT analysis. We analysed response times for both accurately and inaccurately judged sentences as these responses have been shown to be informative (Jegerski, 2015).

Statistical analyses of the reading/reaction times were performed using mixed-effects models computed with the “lme4” package in R (Bates et al., 2015). We started by constructing a maximal model including random intercepts and slopes

for participants and items (Barr et al., 2013) and also used an optimizer in the models analysing the reading time data (optimizer = ‘bobyqa; Powell, 2009). As the maximal models failed to converge, we simplified them by removing the random slopes. Each final model included both participant and item as random intercepts. The two-level factorial interaction variables, evidentiality (direct, indirect), grammaticality (match, mismatch), and group (heritage speakers, emigrant speakers) were sumcoded.⁴ All reading times were log-transformed in the models to reduce the positive skew. As the participant groups differed significantly in age and verbal working memory, we controlled for these variables by adding them as fixed effects. Since it is known to affect reading speed, region length, that is the number of letters composing each word, was also included as a control variable (Jegerski, 2014). To explore the accuracy differences between groups, similar models to those outlined above were built and the scores were analysed using logit generalised mixed-effects models. *Post hoc* pairwise comparisons to explore the nature of the interactions were conducted using the “emmeans” package (Lenth, 2019) and adjusted using Holm correction for multiple comparisons.

Results

Offline processing: sentence judgement results

Table 3 shows the mixed-effect model output for the response time data and the generalised mixed-effects model output for accuracy to sentence judgement questions.

Sentence judgement accuracy

The generalised mixed effect model for sentence judgement accuracy revealed a three-way interaction between group, grammaticality and evidential (see Figure 1A). *Post-hoc* analyses showed that, in the *direct* condition, both groups were significantly more accurate in their judgements of the match sentences compared to mismatch sentences (HLS: $\beta = 1.22$, $SE = 0.184$, $z = 6.61$, $p < 0.001^*$; ES: $\beta = 1.55$, $SE = 0.183$, $z = 8.51$, $p < 0.001^*$). However, while the same held for the *indirect* condition for heritage language speakers ($\beta = 0.64$, $SE = 0.180$, $z = 3.57$, $p = 0.002^*$) it did not for the emigrant speakers who showed no significant difference between the match and mismatch sentences in this condition ($\beta = 0.23$, $SE = 0.173$, $z = 1.33$, $p = 0.369$). In terms of evidentiality, for match sentences, both groups were significantly more accurate in the *direct* compared to *indirect* condition (HLS $\beta = 0.58$, $SE = 0.186$, $z = 3.12$, $p = 0.007^*$; ES $\beta = 1.01$, $SE = 0.18$, $z = 5.46$, $p < 0.001^*$), but showed no such difference for the mismatch sentences (HLS $\beta < 0.001$, $SE = 0.177$, $z = 0.04$, $p = 0.970$; ES $\beta = -0.32$, $SE = 0.169$, $z = -1.87$, $p = 0.185$).

³ A variety of upper boundaries (1,000, 2,000, 2,500, 5,000 ms) have been mentioned in self-paced reading studies (e.g., Hofmeister, 2011; Vasishth and Drenhaus, 2011; Nicklin and Plonsky, 2020). Given the population studied in this study and their varying language competencies we aimed to not exclude any critical data and selected an upper boundary of 8,000 ms, taking into account the visual inspection of the data.

⁴ Although we labelled this variable “grammaticality” the mismatching sentences are not ungrammatical *per se*. It is possible in some contexts for these sentences to be plausible. Nevertheless, without the presentation of a full context (as is the case in this experiment) these mismatching sentences are not acceptable, as they present opposite sources of information.

TABLE 3 Mixed-effects estimates of accuracy and response times for the sentence judgement questions.

	Sentence judgement questions accuracy				Sentence judgement questions response times			
	β	SE	t	p	β	SE	z	p
(Intercept)	0.76	0.11	7.00	<0.001*	7.13	0.05	133.03	<0.001*
Age	> -0.01	0.01	-0.30	0.764	<0.001	0.01	0.28	0.779
Verbal Working Memory	0.11	0.06	1.90	0.057	0.04	0.03	1.08	0.285
Group	0.19	0.11	1.61	0.108	-0.08	0.06	-1.21	0.232
Evidential	-0.16	0.05	-2.91	0.004*	0.01	0.01	1.04	0.299
Grammaticality	-0.45	0.05	-8.25	<0.001*	> -0.01	0.01	-0.35	0.728
Evidential—Grammaticality	0.24	0.05	4.32	<0.001*	-0.01	0.01	-1.15	0.252
Evidential—Group	-0.01	0.03	-0.41	0.679	0.01	0.01	0.70	0.486
Grammaticality—Group	<0.001	0.03	0.29	0.769	-0.01	0.01	-0.88	0.379
Evidential—Grammaticality—Group	0.09	0.03	2.96	0.003*	-0.02	0.01	-2.62	0.009*
Observations	<5468				<5468			
Marginal R ² /Conditional R ²	0.092/0.245				0.011/0.254			
	glmer (Accuracy ~ Evidential * Grammaticality * Group + c. (VerbalWM) + c. (Age) + (1 ParticipantCode) + (1 Item), data=dataAR, family = binomial (link = "logit"))				lmer (log (ResponseRT) ~ Evidential * Grammaticality * Group + c. (RegionLength) + c. (VerbalWM) + c. (Age) + (1 Participant) + (1 Item), data = dataSJ, REML = FALSE, control = lmerControl (optimizer = "bobyqa"))			

Values of significant effects ($p < 0.05$) are printed in bold and asterisked.

Sentence judgement response time

The model output for the response times to the sentence judgement questions showed a three-way interaction between group, grammaticality and evidential (see Figure 1B). The emmeans analyses showed that in the *indirect* condition emigrant speakers were marginally, but not significantly, faster in their responses to the mismatch sentences ($\beta = 0.09$, $SE = 0.036$, $t = 2.60$, $p = 0.067$) compared to the match sentences. They were marginally faster when firsthand information was (incorrectly) followed by the indirect evidential marker compared to when nonfirsthand information was (correctly) followed by the indirect evidential marker. In the match sentences, emigrant speakers' responses were significantly *faster* in the direct condition, that is *firsthand-direct* sentences, compared to the indirect, *nonfirsthand-indirect* sentences ($\beta = -0.10$, $SE = 0.036$, $t = -2.90$, $p = 0.031^*$). The heritage language speakers showed no significant differences in sentence judgement response times across the sentence types.

Online processing: reading time results

The reading time analysis of each critical region revealed, as expected, slower reading times for heritage language speakers compared to the emigrant speakers at every time point (HLS overall mean RT = 907.1, SD = 834.0; ES overall mean RT = 683.6, SD = 690.9). Figure 2 shows by-region reading time averages.

Outputs of the mixed-effects models, computed at each region, are presented in Table 4. At all four regions, as expected, there was a significant effect of the Speaker Group on reading times: heritage

language speakers were slower in their reading times at the target verb compared to emigrant speakers. We will discuss the remaining results for each region in turn.

Target verb

Reading times at the target verb showed a significant two-way interaction between evidential and group (see Figure 3A). *Post hoc* analysis indicated that the source of this interaction was that, while the emigrant speaker group showed no significant difference between the direct and indirect evidentiality conditions ($\beta < 0.01$, $SE = 0.018$, $t = 0.23$, $p = 0.816$), the heritage language speaker group did show a significant difference ($\beta = -0.05$, $SE = 0.020$, $t = -2.60$, $p = 0.040^*$). Irrespective of grammaticality, the heritage language speakers read verbs marked with *direct* evidentiality faster than those with *indirect* evidentiality (see Supplementary material for the full pairwise comparisons of the emmeans analyses).

Spillover region 1

In the first spillover region, comprising the reading times of the first two words following the evidentiality-marked verb, there was a significant three-way interaction between group, grammaticality and evidential (see Figure 3B). This reflected that in the *indirect* condition (when the verb is marked with the indirect evidential) the emigrant speakers were faster when the evidentiality marker did not match with the information source than when it did ($\beta = 0.07$, $SE = 0.016$, $t = 4.47$, $p < 0.001^*$), but this was not the case in the *direct* condition ($\beta = 0.01$, $SE = 0.016$, $t = 0.60$, $p = 1.000$). In addition, emigrant speakers were faster in the *indirect* condition

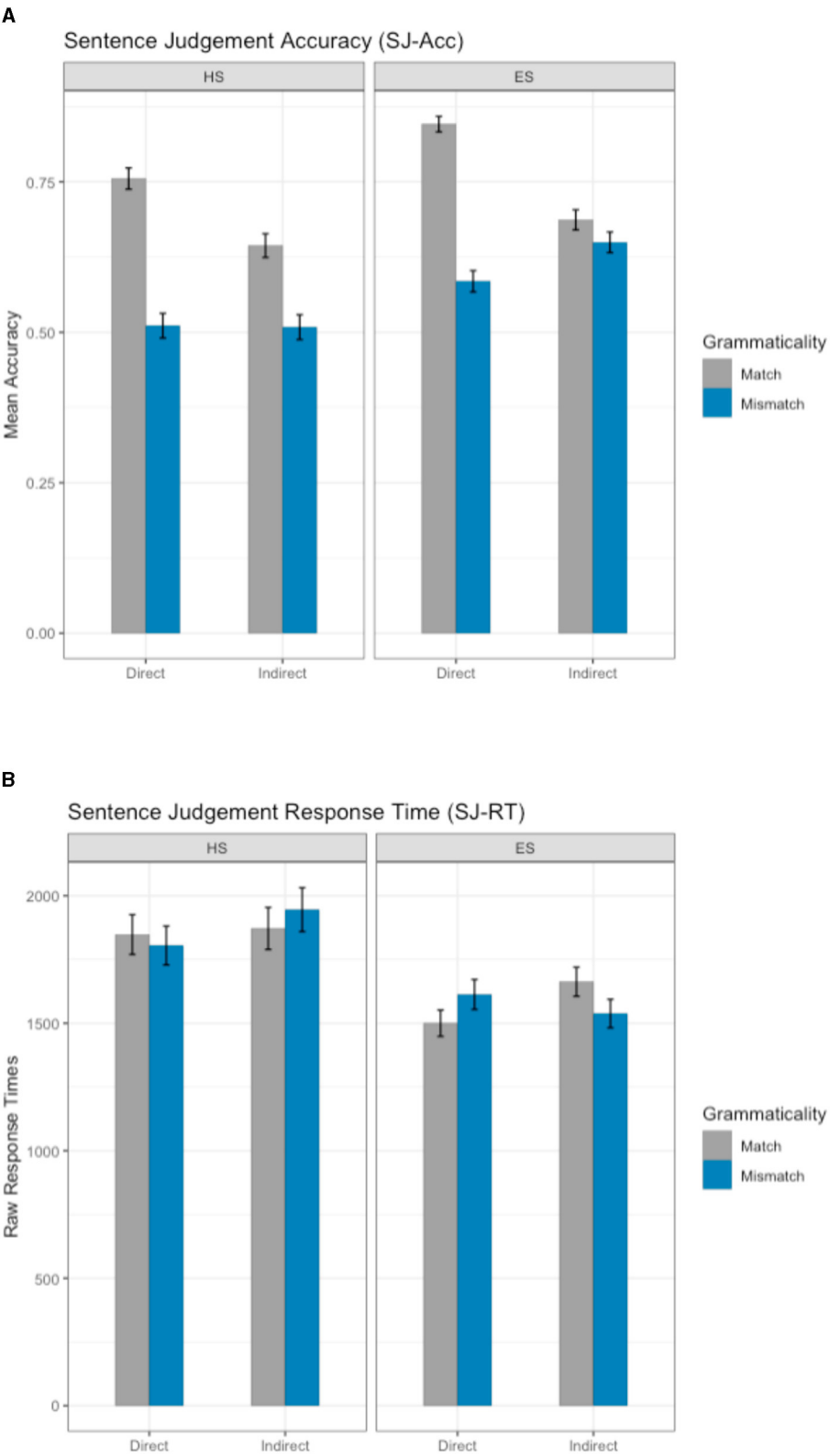


FIGURE 1 Sentence judgement accuracy (A) and response times (B). Error bars represent the standard error of the mean. The means and standart deviations are provided in [Table A1](#).

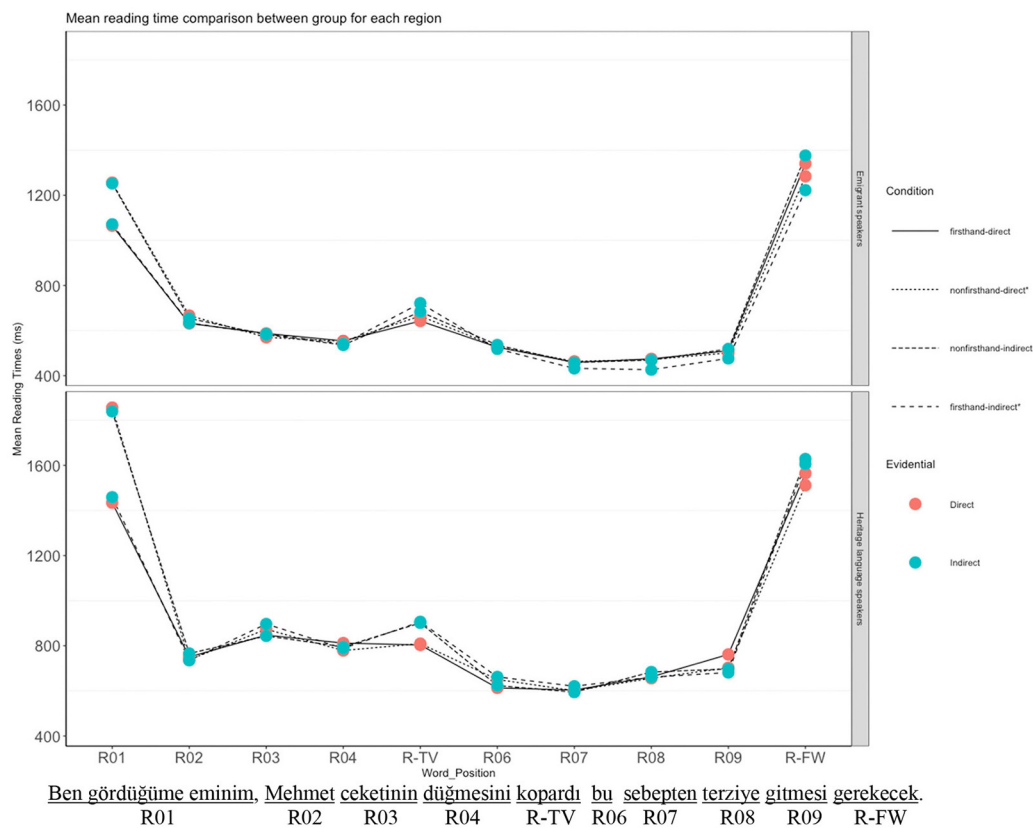


FIGURE 2

Word reading time comparisons across the sentence, for each evidential condition for emigrant speakers and heritage language speakers.

compared to the *direct* condition for mismatch sentences ($\beta = 0.06$, $SE = 0.016$, $t = 3.70$, $p < 0.001^*$) but in the match sentences there was no significant condition difference ($\beta = -0.01$, $SE = 0.016$, $t = -0.17$, $p = 1.000$). However, for the heritage language speakers, there were no significant differences in any of the comparisons.

Spillover region 2

In the second spillover region, comprising the responses to the third and fourth words following the evidentiality-marked verb, there was also a significant three-way interaction between group, grammaticality and evidential (see Figure 3C). In the *indirect* condition, similar to R-SO1, the emigrant speakers were faster when the evidentiality marker did not match with the information source than when it did ($\beta = 0.11$, $SE = 0.018$, $t = 6.26$, $p < 0.001^*$), and also showed a significant but much smaller difference between match and mismatch sentences for the *direct* evidential (ES $\beta = 0.04$, $SE = 0.018$, $t = 2.67$, $p = 0.032^*$).

The heritage language speakers also showed significantly faster reading times for mismatch than match sentences for both the *indirect* ($\beta = 0.06$, $SE = 0.020$, $t = 3.17$, $p = 0.008^*$), and *direct* conditions ($\beta = 0.08$, $SE = 0.020$, $t = 3.98$, $p < 0.001^*$).

When comparing the direct and indirect conditions, the patterns were the same as for R-SO1, in the mismatch conditions emigrant speakers were, once again faster in the *indirect* condition

compared to the *direct* condition ($\beta = 0.06$, $SE = 0.018$, $t = 3.61$, $p = 0.002^*$), but not in the match condition ($\beta < 0.001$, $SE = 0.018$, $t = 0.02$, $p = 1.000$). Similar to R-SO1, the heritage language speakers did not show any significant differences between the direct and indirect conditions in R-SO2 (match: $\beta = 0.01$, $SE = 0.020$, $t = 0.51$, $p = 1.000$; mismatch: $\beta = 0.02$, $SE = 0.020$, $t = 1.32$, $p = 0.563$).

Final word

The model for word reading time at the Final Word region of interest revealed main effects only for grammaticality, but no significant interactions (see Table 4; Figure 3D). Reading times for *mismatch* sentences were significantly shorter than for the *match* sentences in this region.

Summary of results

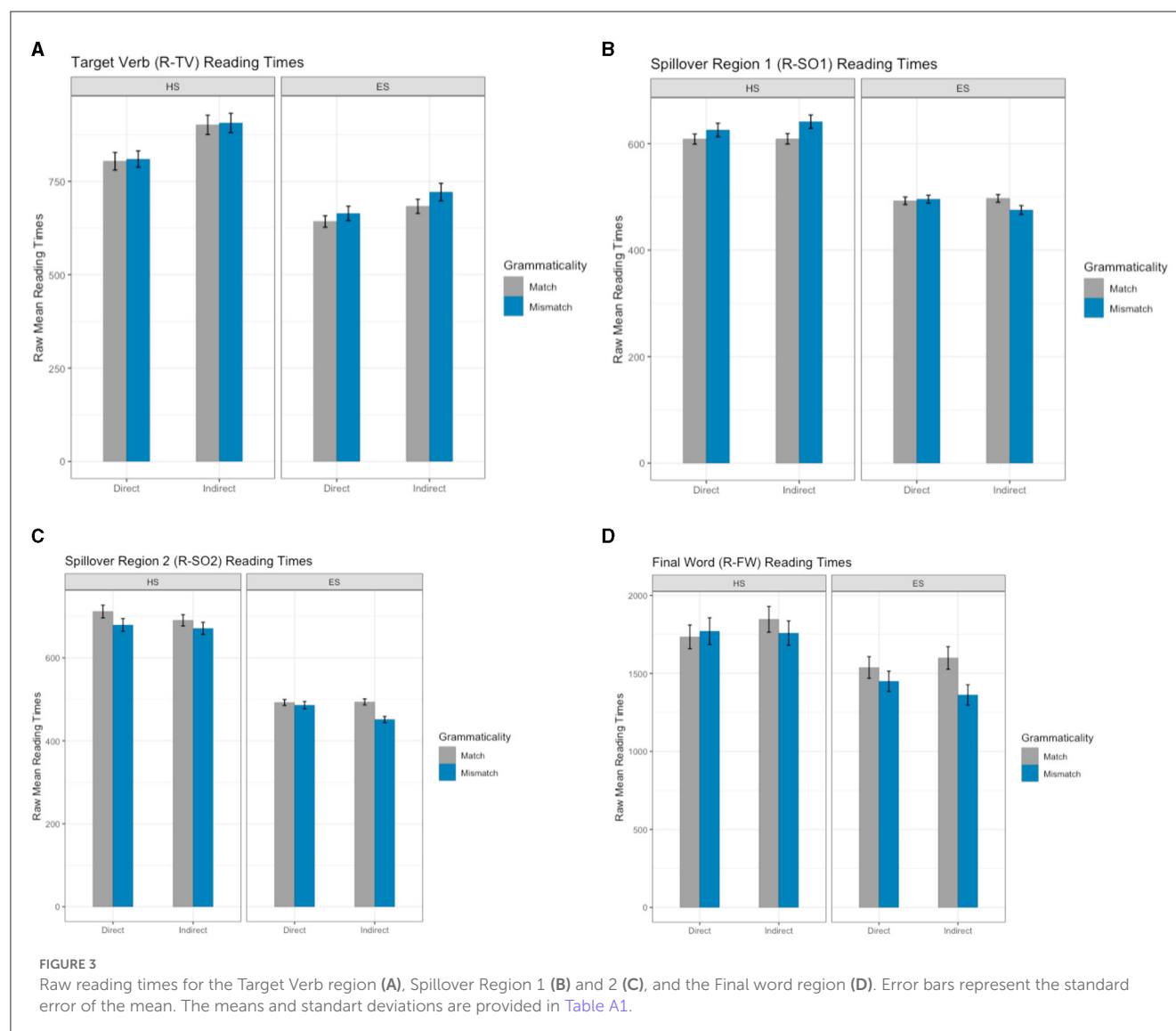
The patterns of response across the analyses are summarised in Table 5.

Reading times during self-paced reading (i.e., online results) showed that the Turkish heritage language speakers were slower in their reading times overall compared to the other bilingual group, i.e., Turkish emigrant speakers. At the target verb, targets marked with direct evidential markers (-DI) were processed

TABLE 4 Mixed-effects estimates for the reading times at the regions of interest.

	Region target verb				Region spillover 1				Region spillover 2				Region final word			
	β	SE	t	p	β	SE	t	p	β	SE	t	p	β	SE	t	p
(Intercept)	6.47	0.04	171.59	<0.001*	6.19	0.03	223.01	<0.001*	6.21	0.03	182.91	<0.001*	6.95	0.06	109.68	<0.001*
Region length	0.05	<0.001	9.80	<0.001*	0.01	<0.001	3.07	0.002*	0.03	<0.001	13.60	<0.001*	0.02	<0.001	4.33	<0.001*
Age	> -0.01	<0.001	-0.70	0.483	0.01	<0.001	1.69	0.097	0.01	<0.001	1.64	0.108	0.01	<0.001	1.74	0.088
Verbal working memory	-0.03	0.02	-1.40	0.167	-0.04	0.02	-2.42	0.019*	-0.05	0.02	-2.27	0.027*	-0.07	0.04	-1.84	0.072
Speaker group	-0.09	0.04	-2.09	0.042*	-0.11	0.03	-3.53	<0.001*	-0.16	0.04	-4.06	<0.001*	-0.13	0.07	-1.73	0.089
Evidential	0.01	0.01	1.58	0.116	> -0.01	<0.001	-1.15	0.252	-0.01	0.01	-2.36	0.020*	> -0.01	0.01	-0.32	0.748
Grammaticality	0.01	0.01	1.06	0.291	-0.01	<0.001	-2.50	0.016*	-0.04	0.01	-7.00	<0.001*	-0.05	0.01	-4.40	<0.001*
Evidential—grammaticality	<0.001	0.01	0.23	0.817	-0.01	<0.001	-1.30	0.194	-0.01	0.01	-1.13	0.260	-0.01	0.01	-1.19	0.237
Evidential—group	-0.01	0.01	-2.40	0.016*	-0.01	<0.001	-2.29	0.022*	> -0.01	<0.001	-0.93	0.354	-0.01	<0.001	-1.48	0.138
Grammaticality—group	<0.001	0.01	0.24	0.811	-0.01	<0.001	-2.34	0.019*	> -0.01	<0.001	-0.62	0.535	-0.02	<0.001	-1.68	0.092
Evidential—grammaticality—group	<0.001	0.01	0.23	0.821	-0.01	<0.001	-2.46	0.014*	-0.01	<0.001	-2.55	0.010*	> -0.01	<0.001	-0.98	0.327
Observations	<5504				<11013				11010				<5472			
Marginal R ² /conditional R ²	0.084/0.334				<0.097/0.267				0.154/0.368				<0.065/0.323			

lmer (log(RT) ~ Evidential * Grammaticality * Group + c. (RegionLength) + c. (VerbalWorkingMemory) + c. (Age) + (1 |Participant) + (1 |Item), data = data, REML = FALSE, control = lmerControl (optimizer = "bobyqa")). Note: Values of significant effects ($p < 0.05$) are printed in bold and asterisked.



faster than those marked with the indirect marker (-mİş) by the heritage language speakers whereas emigrant speakers showed no processing differences.

Emigrant speakers showed significantly faster reading times for mismatch than match sentences with indirect evidential markers at both Spillover Region 1 and Spillover Region 2. However, for the heritage language speakers, this pattern appeared later in the sentence, with a significant difference only at Spillover Region 2. Moreover, in Spillover Region 2, the same effect was apparent for sentences where the verb was inflected with the direct evidential: Both groups were faster at processing mismatched direct conditions compared to the direct match conditions.

In sum, in the second spillover region, both groups showed similar processing: they processed mismatching sentences faster. This faster processing of mismatching sentences was sustained in the final word region without the condition effect for both groups.

The speed of the end-of-sentence acceptability judgements (i.e., offline results) revealed no effect of evidentiality condition

or sentence grammaticality for the heritage language speakers. But the emigrant speakers, showed a similar pattern to their online responses in the spillover regions, responding marginally faster in mismatching sentences when the verb was inflected with the indirect evidential marker following a firsthand information source compared to sentences when the indirect evidential marker matched with a nonfirsthand information source. In the acceptable, match, sentences, their responses were significantly faster when the sentences included the direct evidential marker than the indirect evidential marker.

In terms of accuracy of sentence acceptability judgements, both heritage language speakers and emigrant speakers were more accurate in their judgements of the match vs. mismatch sentences in the direct evidential condition. In the indirect evidential condition, only heritage language speakers showed a significant difference, with match sentences being more accurately responded to than mismatch sentences. Both groups were more accurate in the direct than the indirect condition for the match sentences.

TABLE 5 Summary of analyses examining patterns across reading time (online) and end of sentence judgement (offline) measures by condition.

	Online reading times						Offline end-of-sentence acceptability judgements					
	Target verb region			Spillover region 1			Final word region			Response Time		
	HLS	ES		HLS	ES		HLS	ES		HLS	ES	ES
Match vs. mismatch												
Indirect												
Direct												
Direct vs. indirect												
Match												
Mismatch												

A < B, A is significantly faster, or less error prone (more accurate), than B; A ≤ B, A is marginally significantly faster than B (p = 0.06); HLS, Heritage Language Speakers; ES, Emigrant Speakers; Dir, Direct; Ind, Indirect; Mat, Match; Mis, Mismatch.

Discussion

This study was motivated by two issues in the study of the language processing of heritage language speakers: (1) increasing interest in the use of, and additional awareness of the importance of, online methods; (2) the importance of including a reference group that is comparable to the heritage language speakers. We, therefore, aimed to provide new insights into the processing of evidentiality in Turkish heritage language speakers using a task, self-paced reading, that provides online and offline measures, and by comparing their performance to that of a reference group of late bilingual emigrant speakers of Turkish also living in Australia. During the self-paced reading task, participants were presented with sentences that started with an indication of either a firsthand or a nonfirsthand information source followed by matching or mismatching evidentiality markers. We will first discuss the offline, behavioural data, from end of sentence judgements before moving to the online data and a comparison of the two.

In sentence acceptability judgements, both groups were more accurate in their judgements of the matching vs. mismatching sentences in the direct evidential condition. However, in the indirect evidential condition, the ES group were equally accurate for matching and mismatching sentences, whereas the heritage language speakers showed poor response accuracy for mismatching sentences relative to matching sentences for both evidential contexts. The heritage language speaker group showed poor response accuracy for mismatching sentences relative to matching sentences for both evidential contexts. Simply put, the heritage language speaker group showed a tendency to judge mismatching sentences to be acceptable in about 50% of all the trials: they lacked sensitivity in judging the acceptability of both the evidential forms in reference to given information source contexts. Such a finding is completely consistent with the pattern found in a listening task for the same heritage language speaker group (see Tokaç-Scheffer et al., to appear) and for a separate group of heritage language speakers residing in the Netherlands reported in Arslan et al. (2017).

As mentioned above, we did not observe a significant difference between direct and indirect evidentials in either groups' accuracy to mismatching sentences, suggesting that they considered mismatches of both information sources to evidential forms equally unacceptable to an extent. Following Arslan et al. (2017) data from Turkish monolingual speakers, we may have expected to find an asymmetry here with higher response accuracy in detecting violations when the firsthand information source mismatches to the indirect evidential than when the nonfirsthand information source mismatches to the direct evidential. Recall that such an evidential context, mismatching use of indirect evidential marker following firsthand information source, is what Aikhenvald (2004, p. 217) refers to as counter-intuitive. Note that in the previous literature this behavioural asymmetry was found for monolingual speakers (Arslan et al., 2017; Karaca, 2018; Schmid and Karayayla, 2019), and it is not surprising that this counter-intuitiveness was not reflected in our bilingual groups' responses who may have reduced sensitivity to evidentiality, as even the emigrant speaker group were living in conditions where Turkish was not the language of the society (i.e., what could be considered heritage language

conditions). It is important to note, however, as we discuss below, that we found such an asymmetry towards faster reading times in firsthand information sources mismatched to indirect evidential forms at the post-critical regions during word-by-word processing for the emigrant speakers. This suggests that living under heritage language conditions and being exposed to a majority language which lacks grammatical evidentiality limits and/or influences the representation of evidentiality in our groups of bilinguals but in online processing we still find a clear response to counter-intuitive evidential contexts by the late bilingual group who grew up in Türkiye.

We turn now to online processing in more detail. Our first research question asked whether the online processing of the Turkish heritage language speakers and the Turkish emigrant speakers differed during the time course of reading these sentences with evidentiality marking. As expected, the heritage language speakers were slower in their reading times across the critical segments compared to the emigrant speakers. This is not a surprising outcome when heritage language research is considered (e.g., [Sekerina and Trueswell, 2011](#); [Montrul, 2016](#)). For example, in [Hulsen et al. \(2002\)](#), second-generation heritage language speakers of Dutch born in New Zealand, an experimental group similar to ours, were slower in word retrieval compared to first-generation adult migrant speakers of Dutch. However, as we discuss below, when we examined the pattern of responses of our participants in more detail, we found similarities as well as differences in the performance of the two groups.

Our second research question asked whether the speakers' online processing of evidentiality differed depending on whether the evidentiality marker on the verb was direct or indirect, and whether the two groups differed in this regard. We examined the reading times in four segments of interest including the segment with the target evidential-marked verb, a first spillover region comprised of the first two words following the target verb, a second spillover region with the next two words following the first spillover region, and the final word of the sentence. At the target verb region, the heritage language speaker group was faster in their reading times for the direct than the indirect evidential condition (irrespective of the grammaticality of the sentence). This condition difference seems to be consistent with previous heritage language research conducted with children (e.g., [Aarssen, 2001](#); [Karakoç, 2007](#)) and adults ([Schmid and Karayayla, 2019](#); [Arslan, 2020](#); [Arslan et al., 2020](#)), all of which reported heritage language speakers' better processing of direct evidential forms. Note that, however, direct evidential is the default form in the Turkish evidentiality paradigm as opposed to the indirect evidential which is conceptually rather complex and semantically more "marked" as it refers to an assortment of contexts including reportative and inferential. Outcomes from language acquisition studies often mirror this asymmetry in monolingual Turkish-speaking children's acquisition trajectories in that the direct evidential marker emerges earlier in children's language, both in comprehension and production, and is also fully mastered prior to the indirect evidential marker ([Aksu-Koç, 1988](#); [Öztürk and Papafragou, 2008](#); [Aksu-Koç et al., 2009](#); [Ünal and Papafragou, 2016](#)). Borrowing insights from these acquisition studies, we suggest that heritage language speakers growing up in homes

where Turkish is spoken as a heritage language are likely to better acquire the direct evidential marker than the indirect evidential marker. A possibility here is that, as the majority language (English) which lacks grammatical evidentiality gains dominance after early childhood, the Turkish evidentiality paradigm may have been impacted by negative language transfer effects, and as a consequence, Turkish heritage speakers are tending to take direct evidentiality as a default past tense form (see also [Arslan et al., 2015](#)). The slower reading times of our heritage language speakers for the indirect than the direct evidential conditions, in fact, reflect the erosion of the indirect evidential in Turkish heritage grammar. The lack of such condition differences between the direct and indirect evidential forms in the emigrant speakers' reading times suggests that the individual bilingualism profiles of our bilingual participants (including dominant language setting during childhood, language of formal education) influence the way the evidentiality system erodes in Turkish heritage grammar.

The pattern of results in the spillover regions demonstrated that the emigrant speakers differed from the heritage language speakers in terms of timing, which suggested delayed processing for heritage language speakers compared to emigrant speakers. In the first spillover region, the emigrant speaker group was faster at reading mismatching sentences in the indirect condition, than matching sentences. In the mismatching sentences, a phrase indicating a firsthand information source was followed by an indirect evidential marker. This pairing is not plausible in any circumstance and is therefore easy for native Turkish speakers to reject. The heritage language speakers on the other hand showed a similar pattern but not until the second spillover region (where this pattern was still evident for the emigrant speakers). Hence, while the heritage language speakers were sensitive to the mismatch of firsthand information, they were slower to do so. A similar latency effect in evidentiality processing was shown by [Karaca \(2018\)](#): In comparison to monolinguals, bilingual first-generation migrants were slow to show sensitivity to the mismatch of firsthand information, relative to monolinguals. The significantly slower listening that monolinguals showed for the incongruent indirect evidential compared to congruent in the second segment did not appear until the third segment for first-generation immigrants. As pointed out by [Jegerski \(2014\)](#), any effects on latency during self-paced reading should be interpreted cautiously and that it is not simply that heritage language speakers are merely slower overall. Importantly, we showed that, despite their overall slower latencies, heritage language speakers showed qualitative similarities to the reference group with both groups showing similar sensitivity to the same evidential condition (direct vs. indirect) and grammaticality (match vs. mismatch) combination: when mismatch indirect sentences are compared to match sentences.

With regard to our results from the analysis of the second spillover region, the two groups were similarly faster at reading the mismatch sentences compared to match sentences in the direct condition. In contrast, in both the first and second spillover regions, only emigrant speakers exhibited a difference between the direct and indirect evidential conditions, and this was only in the mismatching sentences. That is, when a firsthand information source was violated by the use of the indirect evidential marker (mismatching indirect condition), they were faster than when

a nonfirsthand information source was violated by the direct evidential marker (mismatching direct condition). This was similar to the pattern shown by monolinguals in previous studies (Arslan et al., 2017; Tokaç-Scheffer et al., to appear). According to Aikhenvald (2004, p. 217) upon being given firsthand information, it is “counterintuitive” to challenge this experience with the use of indirect evidential and native speakers are faster at rejecting such mismatches.

It is worth reiterate that while the heritage language speakers showed a difference between direct and indirect sentences at the target verb (for both match and mismatch sentences), they did not show it in the spillover regions. The lack of an effect of condition in later processing shows some similarity to previous (offline and online) studies: Neither of the studies using (offline) auditory verification (go/no go) tasks (Arslan et al., 2017; Tokaç-Scheffer et al., to appear) revealed any effect of evidentiality condition for heritage language speakers. Similarly, during a self-paced listening task, heritage language speakers in Karaca (2018) did not show any significant processing differences between the conditions in any of the sentence segments. This is at odds with Arslan et al. (2015) findings from the eye-movement monitoring experiment which demonstrated that both early and late bilingual Turkish speakers were slower and less accurate to respond to direct evidential than indirect evidential conditions. Arslan et al. (2015) used a visual representation of evidence while their participants listened to sentences with evidential forms, and it seems these condition differences are reflecting a somewhat different aspect of evidentiality processing. The precise nature of these condition differences requires further studies critically investigating grammaticality/acceptability judgement tasks as opposed to naturalistic tasks. We will turn to this issue below.

In the spillover regions, it is also important to underline that there was a significant difference between match and mismatch sentences in the indirect condition. This effect was observed in both post critical regions for emigrant speakers, only in the second spillover region for heritage language speakers, possibly due to delayed processing speed. Both bilingual groups rejected the indirect mismatch sentences faster both in-between and within conditions. However, the difference between match and mismatch sentences in the direct condition was only observed in the second spillover region for both groups. We argue that participants were faster at reading times when presented with a violation of firsthand direct information with a mismatching indirect evidential marker.

The final word reading times were longer than those for any other word/segment of the sentence and this was true for both groups. It is important to note that in this experiment, the last word of the sentence was indicated by the presence of segment lines (see the Method section for details). Consequently, participants knew that with the next key press they would be asked to make a decision about the acceptability of the sentence they had just read. This resulted in participants generally having longer reading times for this segment compared to the rest of the sentence (see Figure 1). This is quite a common phenomenon in self-paced reading and eye-tracking studies known as the “wrap-up effect” (Just and Carpenter, 1980). At this stage, the parser evaluates and assesses all the presented information and resolves any “inconsistencies” (Just and Carpenter, 1980, p. 345). Longer reading times at the end of

sentence have also been hypothesised to be due to readers preparing for the execution of the next task (Stowe et al., 2018). It is not possible to tease apart the role of these two effects in our experiment nor was it the purpose of this study. At this time point (the final word), qualitatively, the two groups showed similar processing with no effect of evidentiality condition and faster reading times for sentences with unacceptable evidentiality marker use following a mismatching information source. The re-evaluation taking place at the final word was much faster for the violated (mismatching) sentences for both groups. The similarities between the two groups that were observed in this final segment provide support for the validity of using a self-paced reading task. As highlighted out by previous researchers, self-paced reading offers qualitative insight on the nature of processing differences between heritage language speakers and the reference group they are being compared to. This extends beyond quantitative comparison of slower processing, and, furthermore, enables seeing whether the differences between monolinguals and non-monolinguals are diminishing or becoming less apparent (Jegerski et al., 2016; Jegerski, 2018a,b).

Our third research question sought to address the benefits of including online methods in heritage language studies and asked whether there were differences between online and offline processing of evidentiality for heritage language speakers and emigrant speakers. As noted above, the examination of the offline results, end-of-sentence judgement questions, revealed that overall, heritage language speakers were slower and less accurate than the emigrant speakers. End-of-sentence acceptability judgements are thought to measure metalinguistic knowledge gained most likely through formal teaching (Bayram et al., 2021). At this point, the individuals parsing the sentences are aware that they need to make a decision, so they re-evaluate their processing, complete the missing information, and solve the “linguistic problems” to make their judgements (Keating and Jegerski, 2015, p. 3). The emigrant speakers, who had received formal education in the language under investigation, showed better accuracy. However, we cannot ignore the fact that the heritage language speakers of the current study, who were residing in Australia had also participated in learning activities through community schools and, the accuracy of these heritage language speakers was higher than that of the heritage language speakers in other studies, who had not received schooling in their heritage language (Arslan et al., 2017; discussed in Tokaç-Scheffer et al., to appear).

The detailed examination of the offline patterns in comparison to the online patterns showed some differences for both groups. The differences between direct and indirect evidentiality conditions that were captured in the target verb region for heritage language speakers were reflected in the accuracy results but only for the match sentences. Their online processing of direct evidentiality markers was faster than that of the indirect markers, and at the (offline) end-of-sentence judgements they were more accurate at judging grammatical direct evidentiality sentences (matching between a firsthand information source and a direct evidential marker) compared to the indirect grammatical sentences. Why would the processing of direct evidential markers be more accurate and quicker as compared to their indirect counterparts? Above, we mentioned the impact of potential processing asymmetries given the primacy of the direct evidential in Turkish monolingual

children. This is based on the idea that indirect evidential is semantically more “marked” as it refers to a number of indirect information sources (i.e., inference, reportative), and the ability to monitor indirect information sources develop with a delay in children’s acquisition. In adults, by contrast, such an asymmetry is not necessarily reflected in behavioural responses. For example, Arslan (2020), using a similar design to ours, showed that a group of Turkish monolingual speakers judged both evidential forms with around 90% task accuracy. Therefore, explaining the presence of strong asymmetry in our adult bilingual speakers as enhanced performance in direct evidential forms over indirect forms based solely on the markedness of the indirect form seems rather unreasonable.⁵ In the absence of data on the developmental trajectories of evidential forms in children acquiring Turkish as a heritage language, we also cannot speculate on whether this asymmetry emerged as result of any possible maturational constraints. The only possible explanation that we can offer at present is the lack of a grammatical indirect evidential form in English, which is the dominant societal language for all our participants. It is conceivable that cross-linguistic transfer effects were at play here and these Turkish speakers in Australia developed a greater tendency towards accepting the direct evidential as the more plausible form in a past time context. Such instances of restructuring of Turkish grammar in contact with English have been attested before. For instance, Gürel (2002) showed that Turkish speakers in North America have attuned to English-like overt subject pronouns as opposed to null pronouns.

A second difference in online and offline processing was that our participants exhibited faster reading times for mismatching sentences for both indirect and direct conditions during moment-by-moment reading, while later, at the end of the trial, their judgements were more accurate for the matching sentences. A similar finding was reported for second language learners of English by Juffs and Harrington (1996), who found that participants were more accurate in their judgements of those sentences on which they spent more reading time.

A final point that is important to note is that there might be a task effect in measuring heritage language outcomes. Grammaticality/acceptability judgement tasks in heritage speakers have been suggested to lead to biased outcomes as opposed to naturalistic tasks with time-sensitive online measures as these speakers tend to have low metalinguistic awareness of their heritage language (see Polinsky, 2018). Therefore, it is conceivable that the heritage language speakers under examination in this study were inaccurate in their judgements of sentence acceptability due either to insensitivity in their grammar and/or weakened language awareness of their heritage language. This argument fits in well with why some studies report no condition differences in heritage speakers’ evidentiality processing (Arslan et al., 2017; Tokaç-Scheffer et al., to appear) while some others report critical condition differences (Arslan et al., 2015).

Evidentiality studies with heritage language speakers are scarce, which leaves limited room for comparisons. On the other hand, there are studies comparing online and offline results with other grammatical phenomena in other languages (e.g., Keating et al.,

2016; Jegerski, 2018a,b). Keating et al. (2016), who studied the differential object marking in Spanish, obtained similar results that heritage language speakers showed processing differences between conditions during online self-paced reading but not in their reaction times to the comprehension questions. The lack of difference between the conditions on the sentence judgement results prove that what is observed via offline measures may not be as informative as online measures for non-monolingual processing.

In addition to providing detailed information on heritage language speakers’ evidentiality processing patterns, while showing similarities and differences between online vs. offline tasks, this study also underlines the importance of choosing an appropriate reference group to compare heritage language speakers. Although the (late bilingual) emigrant speakers were faster and more accurate in their processing compared to the heritage language speakers, they had similar moment-by-moment processing of evidentiality to heritage language speakers. The similarities in sentence processing between bilingual groups (heritage language speakers and late bilinguals) during self-paced reading that have been shown in previous studies (Jegerski, 2018b) were confirmed by our results and the differences between online vs. offline processing were clear.

The results from both online and offline measures in the current study suggest that heritage language speakers do process the grammatical details of evidentiality. Although it has been hard to measure the extent of this knowledge comprehensively with offline measures, the moment-by-moment investigation revealed that heritage language speakers can activate and integrate this knowledge during online reading. Results from this study can motivate future research comparing heritage language speakers’ processing to other bilingual groups with a variety of age of onsets and language backgrounds. This can help us understand and come to grounded conclusions on the important factors affecting evidentiality processing in bilinguals.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Macquarie University Human Research Ethics Committee (approval number: 3531). The patients/participants provided their written informed consent to participate in this study.

Author contributions

ST-S primary author and the editor of the original manuscript collected the data. ST-S and SA prepared the stimuli and the task. All authors participated in project conceptualisation and implementation, data analysis, and preparation of this manuscript.

⁵ This point was also suggested by an anonymous reviewer.

Funding

This research was supported by a Macquarie University International Research Excellence Scholarship (iMQRS) held by ST-S.

Acknowledgments

The authors are grateful to Turkish Welfare Association and Cultural Centre in Sydney and their president Beşir Karasu, also the teachers and parents at Atatürk School who helped greatly with participant requirement; to Roelien Bastiaanse and Srđan Popov for their priceless feedback and discussion during the construction of this study. We also would like to thank Serje Robidoux and Solène Hameau for their statistical advice and support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcomm.2023.1070510/full#supplementary-material>

References

- Aarsen, J. (2001). "Development of temporal relations in narratives by Turkish-Dutch bilingual children," in *Narrative Development in a Multilingual Context*, eds L. T. Verhoeven and S. Stromqvist (Amsterdam: Benjamins), 209–231.
- Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford: Oxford University Press.
- Aksu-Koç, A. (1988). *The Acquisition of Aspect and Modality: The Case of Past Reference in Turkish*. Cambridge: Cambridge University Press.
- Aksu-Koç, A., Ögel-Balaban, H., and Alp, I. E. (2009). "Evidentials and source knowledge in Turkish," in *Evidentiality: A Window Into Language and Cognitive Development, New Directions for Child and Adolescent Development*, eds S. A. Fitneva and T. Matsui (San Francisco: Jossey-Bass), 13–28.
- Anderson, J. A., Mak, L., Keyvani Chahi, A., and Bialystok, E. (2018). The language and social background questionnaire: assessing degree of bilingualism in a diverse population. *Behav. Res. Methods* 50, 250–263. doi: 10.3758/s13428-017-0867-9
- Arslan, S. (2020). When the owner of information is unsure: epistemic uncertainty influences evidentiality processing in Turkish. *Lingua* 247, 102989. doi: 10.1016/j.lingua.2020.102989
- Arslan, S., Bastiaanse, R., and Bayram, F. (2020). "First language exposure predicts attrition patterns in Turkish heritage speakers' use of grammatical evidentiality," in *Studies in Turkish as a Heritage Language*, ed F. Bayram (Amsterdam: John Benjamins), 105–126.
- Arslan, S., Bastiaanse, R., and Felser, C. (2015). Looking at the evidence in visual world: eye movements reveal how bilingual and monolingual Turkish speakers process grammatical evidentiality. *Front. Psychol.* 6, 1387. doi: 10.3389/fpsyg.2015.01387
- Arslan, S., De Kok, D., and Bastiaanse, R. (2017). Processing grammatical evidentiality and time reference in Turkish heritage and monolingual speakers. *Bilingual. Lang. Cogn.* 20, 457–472. doi: 10.1017/S136672891500084X
- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Memory Lang.* 68, 255–278. doi: 10.1016/j.jml.2012.11.001
- Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Statistic. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01
- Bayram, F., Kubota, M., Luque, A., Pascual y Cabo, D., and Rothman, J. (2021). You can't fix what is not broken: contextualizing the imbalance of perceptions about heritage language bilingualism. *Front. Educ.* 6, 628311. doi: 10.3389/educ.2021.628311
- Bayram, F., Rothman, J., Iverson, M., Kupisch, T., Miller, D., Puig-Mayenco, E., et al. (2017). Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *Int. J. Biling. Educ. Bilingual.* 22, 919–939. doi: 10.1080/13670050.2017.1324403
- Carminati, M. N. (2002). *The Processing of Italian Subject Pronouns*, PhD thesis, University of Massachusetts Amherst.
- Chen, E., Gibson, E., and Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *J. Memory Lang.* 52, 144–169. doi: 10.1016/j.jml.2004.10.001
- Clackson, K., Felser, C., and Clahsen, H. (2011). Children's processing of reflexives and pronouns in English: evidence from eye-movements during listening. *J. Mem. Lang.* 65, 128–44. doi: 10.1016/j.jml.2011.04.007
- De Haan, F. (1999). Evidentiality and epistemic modality: setting boundaries. *Southwest J. Linguistics* 18, 83–101.
- De Haan, F. (2005). "Encoding speaker perspective: evidentials," in *Linguistic Diversity and Language Theories*, eds Z. Frajzyngier, A. Hodges, and D. S. Rood (Amsterdam: John Benjamins B. V.), 379–417.
- Di Pisa, G., Rothman, J., and Marinis, T. (2022). *Gender and Number Agreement in Italian as a Heritage Language: A Self-paced Reading Study* [Poster presentation]. Heritage Languages Around the World, Lisbon, Portugal. Available online at: <http://cehum.ilch.uminho.pt/languages#registration> (accessed October 2022).
- Drummond, A. (2013). *Ibex farm*. Available online at: <http://spellout.net/ibexfarm> (accessed September 30, 2021).
- Felser, C., and Cummings, I. (2012). Processing reflexives in a second language: The timing of structural and discourse-level constraints. *Applied Psycholinguistics* 33, 571–603. doi: 10.1017/S0142716411000488
- Felser, C., Sato, M., and Bertenshaw, N. (2009). The on-line application of Binding Principle A in English as a second language. *Bilingual. Lang. Cogn.* 12, 485–502. doi: 10.1017/S1366728909990228
- Filiaci, F. (2011). *Anaphoric Preferences of Null and Overt Subjects in Italian and Spanish: a Cross-linguistic Comparison*, PhD thesis, The University of Edinburgh.
- Fuchs, Z., Polinsky, M., and Scontras, G. (2015). The differential representation of number and gender in Spanish. *Linguistic Rev.* 32, 703–737. doi: 10.1515/trlr-2015-0008
- Gürel, A. (2002). *Linguistic characteristics of second language acquisition and first language attrition: Overt vs. null pronouns*. [doctoral dissertation], McGill University.
- Hofmeister, P. (2011). Representational complexity and memory retrieval in language comprehension. *Lang. Cogn. Processes* 26, 376–405. doi: 10.1080/01690965.2010.492642

- Hulsen, M., Bot, K. D., and Weltens, B. (2002). Between two worlds social networks, language shift, and language processing in three generations of Dutch migrants in New Zealand. *Int. J. Sociolinguistic*. 153, 27–52. doi: 10.1515/ijsl.2002.004
- Jegerski, J. (2014). “Self-paced reading,” in *Research Methods in Second Language Psycholinguistics*, eds J. Jegerski and B. VanPatten (New York, NY: Routledge), 20–49.
- Jegerski, J. (2015). The processing of case in near-native Spanish. *Sec. Lang. Res.* 31, 281–307. doi: 10.1177/0267658314563880
- Jegerski, J. (2016). Number attraction effects in near-native Spanish sentence comprehension. *Stud. Sec. Lang. Acquisit.* 38, 5–33. doi: 10.1017/S027226311400059X
- Jegerski, J. (2018a). Sentence processing in Spanish as a heritage language: a self-paced reading study of relative clause attachment. *Lang. Learn.* 68, 598–634. doi: 10.1111/lang.12289
- Jegerski, J. (2018b). The processing of the object marker *a* by heritage Spanish speakers. *Int. J. Bilingual.* 22, 585–602. doi: 10.1177/1367006916681083
- Jegerski, J., Keating, G. D., and VanPatten, B. (2016). On-line relative clause attachment strategy in heritage speakers of Spanish. *Int. J. Bilingual.* 20, 254–268. doi: 10.1177/1367006914552288
- Juffs, A., and Harrington, M. (1996). Garden path sentences and error data in second language processing research. *Lang. Learn.* 46, 286–324. doi: 10.1111/j.1467-1770.1996.tb01237.x
- Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychologic. Rev.* 87, 329–354. doi: 10.1037/0033-295X.87.4.329
- Just, M. A., Carpenter, P. A., and Wooley, J. D. (1982). Paradigms and processes in reading comprehension. *J. Exp. Psychol.* 111, 228–238.
- Karaca, F. (2018). *Comprehension of Evidentiality in Spoken Turkish: Comparing Monolingual and Bilingual Speakers* (Unpublished Master's thesis), University of Alberta.
- Karakoç, B. (2007). “Connectivity by means of finite elements in monolingual and bilingual Turkish discourse,” in *Connectivity in Grammar and Discourse*, eds J. Rehbein, L. Pietsch, and C. Hohenstein (Amsterdam: John Benjamins), 199–227.
- Karayayla, T., and Schmid, M. S. (2019). First language attrition as a function of age at onset of bilingualism: first language attainment of Turkish–English bilinguals in the United Kingdom. *Lang. Learn.* 69, 106–142. doi: 10.1111/lang.12316
- Keating, G. D., and Jegerski, J. (2015). Experimental designs in sentence processing research. *Stud. Sec. Lang. Acquisit.* 37, 1–32. doi: 10.1017/S0272263114000187
- Keating, G. D., Jegerski, J., and VanPatten, B. (2016). Online processing of subject pronouns in monolingual and heritage bilingual speakers of Mexican Spanish. *Bilingual. Lang. Cogn.* 19, 36–49. doi: 10.1017/S1366728914000418
- Keating, G. D., VanPatten, B., and Jegerski, J. (2011). Who was walking on the beach? anaphora resolution in Spanish heritage speakers and adult second language learners. *Stud. Sec. Lang. Acquisit.* 33, 193–221. doi: 10.1017/S0272263110000732
- Kim, M., Crossley, S. A., and Skalicky, S. (2018). Effects of lexical features, textual properties, and individual differences on word processing times during second language reading comprehension. *Read. Writ.* 31, 1155–1180. doi: 10.1007/s11145-018-9833-x
- Lago, S., Stutter, A., G., and Felsler, C. (2018). The role of native and non-native grammars in the comprehension of possessive pronouns. *Sec. Lang. Res.* 35, 319–349. doi: 10.131219/osf.io/v72gu
- Lee, E., and Fraundorf, S. (2022). Do L1-L2 differences in discourse processing reflect processing demands or difficulty of form-function mapping?: evidence from self-paced listening of contrastive prosody. *Stud. Sec. Lang. Acquisit.* 44, 942–966. doi: 10.1017/S0272263121000619
- Lenth, R. (2019). *Emmeans: Estimated marginal means, aka least-squares means. R package version 1.3.5.1*. Available online at: <https://cran.r-project.org/web/packages/emmeans/index.html>
- Litcofsky, K. A., and Van Hell, J. G. (2017). Switching direction affects switching costs: behavioral, ERP and time-frequency analyses of intra-sentential codeswitching. *Neuropsychologia* 97, 112–139. doi: 10.1016/j.neuropsychologia.2017.02.002
- Luce, R. D. (1991). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford: Oxford University Press.
- Mikhaylova, A. (2018). Morphological bottleneck: the case of Russian heritage speakers. *J. Lang. Contact* 11, 268–303. doi: 10.1163/19552629-01102005
- Montrul, S. (2002). Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingual. Lang. Cogn.* 5, 39–68. doi: 10.1017/S1366728902000135
- Montrul, S. (2004). Subject and object expression in Spanish heritage speakers: a case of morphosyntactic convergence. *Bilingual. Lang. Cogn.* 7, 125–142. doi: 10.1017/S1366728904001464
- Montrul, S. (2009). Knowledge of tense-aspect and mood in Spanish heritage speakers. *Int. J. Bilingual.* 13, 239–269. doi: 10.1177/1367006909339816
- Montrul, S. (2016). *The Acquisition of Heritage Languages*. Cambridge: Cambridge University Press.
- Montrul, S., and Bowles, M. (2009). Back to basics: Differential object marking under incomplete acquisition in Spanish heritage speakers. *Bilingual. Lang. Cogn.* 12, 363–383. doi: 10.1017/S1366728909990071
- Montrul, S., Foote, R., and Perpiñán, S. (2008). Gender agreement in adult second language learners and Spanish heritage speakers: the effects of age and context of acquisition. *Lang. Learn.* 58, 503–553. doi: 10.1111/j.1467-9922.2008.00449.x
- Montrul, S., and Sánchez-Walker, N. (2013). Differential object marking in child and adult Spanish heritage speakers. *Lang. Acquisit.* 20, 109–132. doi: 10.1080/10489223.2013.766741
- Nicklin, C., and Plonsky, L. (2020). Outliers in L2 research in applied linguistics: a synthesis and data re-analysis. *Ann. Rev. Appl. Linguistic.* 40, 26–55. doi: 10.1017/S0267190520000057
- Öztürk, Ö., and Papafragou, A. (2008). The acquisition of evidentiality in Turkish. *Univ. Pennsylvan. Working Papers Linguistic.* 14, 297–309. Available online at: <https://repository.upenn.edu/handle/20.500.14332/44673>
- Polinsky, M. (2008). Relative clauses in heritage Russian: fossilization or divergent grammar? *Form. Approach. Slavic Linguistic.* 16, 333–358. Available online at: <http://nrs.harvard.edu/urn-3:HUL.InstRepos:3403062>
- Polinsky, M. (2018). *Heritage Languages and Their Speakers* (Vol. 159). Cambridge: Cambridge University Press.
- Polinsky, M., and Kagan, O. (2007). Heritage languages: In the ‘wild’ and in the classroom. *Lang. Linguistics Compass* 1, 368–395. doi: 10.1111/j.1749-818X.2007.00022.x
- Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06*, University of Cambridge, Cambridge.
- R Core-Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Available online at: <http://www.r-project.org/>
- Rothman, J. (2007). Heritage speaker competence differences, language change, and input type: inflected infinitives in Heritage Brazilian Portuguese. *Int. J. Bilingual.* 11, 359–389. doi: 10.1177/13670069070110040201
- Schmid, M. S., and Karayayla, T. (2019). The roles of age, attitude, and use in first language development and attrition of Turkish–English bilinguals. *Lang. Learn.* 70, 54–84. doi: 10.1111/lang.12361
- Sekerina, I. A., and Trueswell, J. C. (2011). Processing of contrastiveness by heritage Russian bilinguals. *Bilingual. Lang. Cogn.* 14, 280–300. doi: 10.1017/S1366728910000337
- Stowe, L. A., Kaan, E., Sabourin, L., and Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition* 176, 232–247. doi: 10.1016/j.cognition.2018.03.011
- Tokaç-Scheffer, D. (2023). *An investigation of heritage language speakers of Turkish: Evidentiality processing and the effects of language experience* (Doctoral dissertation), University of Groningen, Groningen, Netherlands. doi: 10.33612/diss.593422246
- Tokaç-Scheffer, S. D., Nickels, L., and Arslan, S. (to appear). “One suitcase, two grammars: What can we conclude about Australian Turkish heritage speakers’ divergent processing of evidentiality?” in *Cognitive mechanisms driving contact-induced language change* (LangVang).
- Ünal, E., and Papafragou, A. (2016). Production–comprehension asymmetries and the acquisition of evidential morphology. *J. Memory Lang.* 89, 179–199. doi: 10.1016/j.jml.2015.12.001
- Vasishth, S., and Drenhaus, H. (2011). Locality in German. *Dialog. Discourse* 2, 59–82. doi: 10.5087/dad.2011.104
- Xu, X., Chen, Q., Panther, K. U., and Wu, Y. (2018). Influence of concessive and causal conjunctions on pragmatic processing: online measures from eye movements and self-paced reading. *Discourse Process.* 55, 387–409. doi: 10.1080/0163853X.2016.1272088

Appendix

Table A1 Means (M) and standard deviations (SD) for each condition of end-of-sentence judgement accuracy (SJ-Acc), response times (SJ-RT), and reading times at each region.

		SJ-Acc		SJ-RT		R-TV		R-SO1		R-SO2		R-FW	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
HLS	Direct match	0.75	0.43	1,847.9	1,898.1	803.7	577.4	608.6	330.9	711.9	530.6	1,733.8	1,859.8
	Direct mismatch	0.51	0.50	1,804.9	1,855.3	809.6	537	625.6	439.8	679.4	528.9	1,770.7	2,082.8
	Indirect match	0.64	0.48	1,871.8	2,003.8	901.4	632.5	609	337.3	690.5	468.7	1,846.5	2,018.1
	Indirect mismatch	0.51	0.50	1,945.5	2,082.6	906.5	634	641.3	437.3	671.1	507.2	1,758	1,903.9
ES	Direct match	0.85	0.36	1,500	1,441.2	642.4	428.5	492.7	281	492.5	284.1	1,538	1,926.2
	Direct mismatch	0.58	0.49	1,612.7	1,634.8	664.2	538.6	495.7	291.2	486.1	346.7	1,449.3	1,802.1
	Indirect match	0.69	0.46	1,662.8	1,585.3	683.1	526.4	497	285.7	493.8	288.4	1,598.7	2,012.5
	Indirect mismatch	0.65	0.48	1,537.8	1,557.3	721.3	657.6	475.4	331.3	451.5	291.8	1,361.7	1,818



OPEN ACCESS

EDITED BY
Behcet Özncar,
Near East University, Cyprus

REVIEWED BY
John Lipski,
The Pennsylvania State University (PSU),
United States
Ad Backus,
Tilburg University, Netherlands

*CORRESPONDENCE
Fatih Bayram
✉ fatih.bayram@uit.no

RECEIVED 01 February 2023
ACCEPTED 30 October 2023
PUBLISHED 01 December 2023

CITATION
Antonova-Unlu E and Bayram F (2023) The role
of external factors on the reactivation of the
heritage language of Turkish-German
returnees.
Front. Psychol. 14:1156779.
doi: 10.3389/fpsyg.2023.1156779

COPYRIGHT
© 2023 Antonova-Unlu and Bayram. This is an
open-access article distributed under the terms
of the [Creative Commons Attribution License
\(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction
in other forums is permitted, provided the
original author(s) and the copyright owner(s)
are credited and that the original publication in
this journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted which
does not comply with these terms.

The role of external factors on the reactivation of the heritage language of Turkish-German returnees

Elena Antonova-Unlu¹ and Fatih Bayram^{2*}

¹Department of Foreign Language Education, Hacettepe University, Ankara, Türkiye, ²Department of Language and Culture, UiT The Arctic University of Norway, Tromsø, Norway

Introduction: This study investigates the heritage language performance of Turkish-German returnees upon their reintegration into Turkey and explores the impact of external factors on their proficiency in the (re-)activated heritage language (HL).

Methods: Data collection involved the participation of 28 Turkish heritage speakers and a control group of 28 monolingual speakers. The language proficiency of both groups was assessed through a cloze test and an error correction task with a focus on converbial constructions, evidentiality and direct object case marking in Turkish. A sociolinguistic background questionnaire was used to obtain information about their language experiences. The study focused on understanding the individual and group differences in returnee's heritage language performance. Additionally, random forest analysis was employed to investigate the relative influence of external factors on individual variability within the returnee group.

Results and Discussion: The analysis of results revealed notable group differences between the returnees and the control group, emphasizing the unique linguistic challenges faced by those who returned to Turkey. Within the returnee group, there was considerable individual variability in heritage language performance. The subsequent exploration of individual variation highlighted the significant role of external factors. Notably, the length of residence in Germany, the age at which participants returned to Turkey, and the frequency of Turkish language use in their migration context emerged as significant predictors of the returnee participants' proficiency in their (re-)activated HL. Surprisingly, formal contact with the dominant German language did not exert a substantial impact on the returnees' language proficiency, suggesting the nuanced influence of various external factors on heritage language development.

KEYWORDS

heritage language, returnees, external factors, language reactivation, bilingualism, Turkish-German heritage speakers, language performance, heritage language maintenance

Introduction

Returnees represent a unique subset of heritage speakers, typically born in an immigrant setting or spending a significant portion of their early childhood and/or adulthood in that setting before returning to their home or heritage country (Yoshitomi, 1999; Flores, 2020; Flores and Snape, 2021). Much like heritage speaker bilinguals,

returnees grow up as bi-/multilingual individuals. They acquire one or more languages as their heritage language, primarily used within their home environment, while they are exposed to the societal language in the broader community, particularly through their schooling. This pattern often results in varying outcomes in their heritage language, typically differing from the established baselines, and the societal language tends to become dominant (for an extensive review, see Montrul, 2016, 2022; Polinsky, 2018). What sets returnees apart is their unique journey of reactivating or relearning their heritage language, leading to a shift in dominance from their former dominant language to their heritage language. This shift occurs as a result of immersion in the home language context upon returning to their home country.

Existing literature on heritage speaker bilinguals consistently underscores the defining characteristic of their linguistic competence in their heritage language: individual variation. Heritage speakers are a diverse group, with their linguistic competence spanning a wide spectrum of differences shaped by their specific contexts and experiences with the languages around them (see, for example, Gathercole and Thomas, 2005; Rodina and Westergaard, 2015; Unsworth, 2015; Montrul, 2016; Correia and Flores, 2017; Antonova-Unlu and Wei, 2018; Kupisch and Rothman, 2018; Armon-Lotem and Meir, 2019; Polinsky and Scontras, 2019, 2020; Bayram et al., 2021; Montrul, 2022). When returnees, who, by nature, are former heritage speakers, settle in their new home country environment, the exposure to the home country language, which is now the dominant societal language, has consequences for their linguistic competencies in both their heritage language and their second language—the societal dominant language in their former country of residence. Similar to the context of typical heritage language bilinguals, the context in which returnees continue to live in their new home country plays a crucial role in determining their linguistic outcomes.

Therefore, the interesting question that arises is whether all returnees end up with the same linguistic outcome in their heritage language. What happens to the HL when HSs return to their country of origin? Will the potential variations in the HL exist after many years spent in an environment where their HL is dominant? If there is still variation, what experiential factors will predict the attainment in the (re-)activated HL? The present study aims to examine these questions, focusing on the (re-)activation of the (re-)activated heritage Turkish of Turkish-German returnees.

(Re-)activation of returnees' HL

Returning to the home country entails one significant linguistic consequence, namely the transition of the once-heritage language back to the dominant societal language. This shift naturally offers various advantages, e.g., among others, increased exposure to a broader spectrum of formal and informal language use with an increased number of opportunities for interaction with various interlocutors. Limited research suggests that the reactivation of the heritage language yields positive outcomes. For instance, within the first year of returning for lexicon, and, on average, after approximately 7 years for morphosyntax, the heritage language of returnees can become nearly indistinguishable from that of monolingual speakers

in their home country (Daller and Yildiz, 1995; Treffers-Daller et al., 2016).

However, it is also important to recognize that not all aspects of the heritage language may fully reactivate at the same level (Treffers-Daller et al., 2007; Kaya-Soykan et al., 2023). For instance, in the context of heritage Turkish, Treffers-Daller et al. (2007) examined the use of syntactic embeddings by Turkish-German bilinguals residing in Germany, Turkish-German returnees who had been living in Turkey for 8 years, and Turkish monolinguals. This study revealed that Turkish-German bilinguals used fewer and less complex embeddings than the returnee group. While some returnees exhibited similar performance outcomes to the control group, both the heritage speaker group and the returnees, on average, performed relatively worse than the monolingual control group. Similarly, Kaya-Soykan et al. (2023) looked at the use of evidentiality markers in the heritage Turkish of Turkish-German returnees who returned to Turkey as adults and had resided in the country for more than a decade. The findings show differences in the heritage Turkish of the returnees compared to the Turkish control group. Even after many years of residence in their home country, the returnees preserved features typically found in heritage speakers.

To our knowledge, only two studies have explored the factors influencing competence in the reactivated heritage language. Flores and Rato (2016) examined the accent of returnees in their heritage language and found that the age at which they immigrated to the host country, Germany, influenced the variability in their heritage language pronunciation, while the length of residence in the home country post-return did not. In a more recent study, Kubota et al. (2021) investigated the narrative skills of Japanese-English returnees in their heritage Japanese immediately upon return and after 1 year of residence in the home country. Their findings indicated that the age at which the returnees returned and their relative proficiency in the heritage language predicted developments in their heritage language skills within the first year after returning.

Current study

The present study aims to examine the linguistic attainment of Turkish-German returnees in their (re-)activated heritage Turkish after several years ($M = 18.96$, $SD = 12.05$) of residence in the country of origin and the role of external factors in the attainments of the (re-)activated HL.

The study aims to answer the following research questions:

- 1 What are, if any, the differences between the (re-)activated HL Turkish and the baseline Turkish?
- 2 What external factors modulate the attainment in the (re-)activated HL of the returnee participants?

To assess the attainment in the (re-)activated HL of the returnees, we examine their overall use of the Turkish language using a c-test as well as an error correction task, including morphosyntactic structures (converbs, evidentiality, and direct object marking) that have been reported as vulnerable in Turkish (see for a review, Arslan et al., 2015; Antonova-Unlu, 2015; Antonova-Unlu and Wei, 2020; Bayram, 2020)

as well as other HLs (Montrul and Polinsky, 2011; Polinsky and Scontras, 2019).

Converbial constructions

A converb is a non-finite verb form that marks adverbial subordination (Haspelmath, 1995, p. 3) operating as a clause-linking device (Coupe, 2006). From a syntactic point of view, converbs are divided into strict and non-strict (Nedjalkov, 1998). Strict converbs are used with an adverbial function only. Non-strict converbs represent forms derived from participles, verbal nouns, or infinitives, and that is why they are often called participles or gerunds used as an adverb. Turkish converbs are strict and non-finite verb forms that function to express time, manner, purpose and result, cause, condition, degree, place, and concession (Goksel and Kerslake, 2005). Example 1 is illustrative:

Example 1:

Çocuk	top-un-u	al-ıp	ev-e	git-ti.
Child	ball-POSS-ACC	take-CONV	house-DAT	go-PAST(3P.SG)
Having taken the ball, the child went home.				

Previous research on the bilingual acquisition of converbial constructions demonstrated that HSs of Turkish diverged from the monolingual baseline in that they tended to use converbs significantly less than the monolingual control group, create sentences where both finite verbs and converbs were used with the subject, which caused ambiguity, as well as place converbs in a detached position loosening the relationship between the converb and the finite verb (Rehbein and Herkenrath, 2015; Turan et al., 2020).

Evidentiality

Evidentiality is a grammatical category that indicates the source of information (Chafe-Nichols, 1986; Aikhenvald, 2004). In Turkish, tense-aspect-modality markers of *-DI* or *-mİş* are used on predicates as evidentiality markers (Aksu-Koç and Slobin, 1986; Aksu-Koç, 1988). *-DI* is used to indicate that the speaker observed/experienced the event (Example 2), while *-mİş* is used to mark indirect experience for cases when events were not observed by the speaker (Example 3).

Example 2:

Çocuk	pasta-yı	ye-di.
Child	cake-ACC	eat-PAST(3P.SG)
The child has eaten the cake.		

Example 3:

Çocuk	pasta-yı	ye-miş.
Child	cake-ACC	eat-EVD(3P.SG)
The child has eaten the cake.		

Depending on the context, *-mİş* may be used to indicate that the information is obtained from another person (hearsay/reportative) or inferred by relying on resultative evidence.

A number of studies demonstrated that HSs tended to replace the indirect evidentiality forms with direct ones, ignoring the source of information, and shift between the two even though there were no reasons for that (e.g., Arslan and Bastiaanse, 2014). Furthermore, HSs were reported to be slower and less sensitive to violations than the monolingual baseline. Evidentiality marking also diverged from the baseline in the (re-)activated heritage Turkish of returnees after residing many years in Turkey (Kaya-Soykan et al., 2023).

Case-marking on direct objects

Case-marking on direct objects is a morphology-syntax-pragmatics interface with two options: (1) the accusative-case ending-*I*, which, depending upon the preceding vowel sound in the stem and the syllable-final phoneme (i.e., whether it is a vowel or a consonant), may have eight different forms (*İ, I, U and Ü, and (y)İ, (y)I, and (y)U and (y)Ü*), and (2) the zero-case ending, in which the form of the direct object is identical with the nominative form of nouns. Four contexts determining the case-marking of direct objects have been defined (Enç, 1991; Kornfilt, 1997; Goksel and Kerslake, 2005; Johanson, 2006). A direct object is accusative-marked if it is definite and specific, that is, being a subset of or standing in some recoverable relation to a familiar object (Enç, 1991, p. 24). A direct object is also accusative-marked if it is indefinite/non-specific and appears before the predicate but not in the closest position to it. Finally, a direct object is zero-case marked if it is indefinite/non-specific and appears in the closest position before the predicate in the sentence.

Thus, the speaker marks a direct object depending on its syntactic position and the oppositions [\pm specific] and [\pm definite] determined by the discourse and speaker-listener knowledge.

Early and late Turkish L2 users also encounter difficulties in case-marking on direct objects even at advanced levels of proficiency and independently from their L1 backgrounds (Gürel, 2000; Altunkol and Balci, 2013; Antonova-Unlu and Wei, 2020). Case-marking on direct objects also diverged from the baseline in the (re-)activated heritage Turkish of returnees after residing many years in Turkey.

Methodology

Tools

A background questionnaire and two tasks (a c-test and an error correction task) were utilized to measure the attainments of the returnees in their (re-)activated heritage Turkish. The c-test and the error correction task were developed by an expert in testing and validated by two instructors in Turkish, who were also native speakers of the language.

Background questionnaire

The information about the participants was obtained from their responses to the questionnaire. Self-reports have been often used in bilingual research for assessing the background and linguistic profiles

of bi-/multilinguals (Marian et al., 2007; de Bruin et al., 2017; Anderson et al., 2018; Marian and Hayakawa, 2021). The questionnaire consisted of 30 questions regarding the background of the returnees: their gender, place and date of birth, levels of education, family status, the age of moving to and returning from Germany, the quantity of Turkish language use, the number of social contacts while in Germany, and perceived levels of proficiency in both languages at the moment of returning to Turkey and at the present time. A 5-point Likert scale was used to rate the use of the Turkish language as 1—“never,” 2—“seldom,” 3—“sometimes,” 4—“often,” and 5—“always.”

C-test

The c-test has been proven to be a useful and reliable tool for measuring holistic proficiency in foreign and native languages (Klein-Braley, 1985; Dörnyei and Katona, 1992; Chapelle, 1994; Koller and Zahn, 1996; among others) as it requires language users to incorporate knowledge from all linguistic levels. Moreover, c-tests have also been used in previous studies investigating heritage Turkish (re-)activation (Daller and Yıldız, 1995; Treffers-Daller et al., 2016).

The c-test used in this study consisted of two authentic texts chosen from the reading materials of the advanced level (C2) of the Turkish teaching coursebook *Hitit* (Uzun, 2018). Text 1 consisted of 263 words and 17 sentences, and text 2 consisted of 248 words and 15 sentences. The expert deleted 20 items in each of the texts. In all but two cases, the deleted item was every 10th word of the text. The two cases that have been decided as unsuitable for deletion included a proper name and a coordinating conjunction *ve* (*and*) for simplicity. The participants were requested to fill in the gaps with a suitable word. The missing words implied the use of roots and/or roots and inflectional and derivational morphemes. There were 20 roots and 8 derivational and 26 inflectional morphemes required in the first text, and 20 roots and 9 derivational and 27 inflectional morphemes required in the second text.

The c-test was piloted on 10 native speakers of Turkish. The test-retest reliability coefficient was calculated as 0.91 over a period of 3 weeks. The participants were requested to fill in the gaps with a suitable word. The c-test was scored using the acceptable method (Alderson, 1979) by which gaps were expected to be filled not with the exact word from the original text but with any appropriate word. For example, in the sentence below taken from the c-test, both *kelimelerden* (*from words*) and *tanımlardan* (*from definitions*) would be correct. Although the word *tanımlardan* was used in the original text, both variants were accepted as correct when assessing the performance of the participants in the c-test.

Example 4:

Toplum temel birimi olan ailenin yaşadığı ev için Türkçedeki 1. _____ biri “huzur ve sükûnet içerisinde yaşanan yer” anlamında kullanılmakta olan “mesken” dir.

One of the Turkish words for the house where the family, the basic unit of society, lives is “residence,” which is used to mean “the place where you live in peace and tranquility.”

Error correction task

The error correction task (ECT) has been proven useful for assessing grammar knowledge, especially of specific domains/structures (Azar, 2007). The ECT was used in this study to

examine the perception of grammatical and ungrammatical uses of Turkish by the returnees and their ability to produce the correct forms. The task consisted of 30 ungrammatical and 30 grammatical items. The items covered three morphosyntax domains (two of which also required the activation of pragmatics) that have been reported as vulnerable in the available research on heritage Turkish acquisition and (re-)activation: evidentiality, direct object case-marking, and converbial constructions (Arslan and Bastiaanse, 2014; Akkus et al., 2017; Turan et al., 2020; Kaya-Soykan et al., 2023).

The participants were requested to judge the task items regarding their grammaticality, as grammatically correct or incorrect, and correct them if considered incorrect. The task was piloted on 10 native speakers of Turkish. The test-retest reliability coefficient was 0.92 over a period of 3 weeks.

Participants

The study participants were 28 Turkish-German bilinguals (*Women* = 18 and *Men* = 10) whose ages varied from 19 to 59 years ($M = 32.79$, $SD = 11.24$). As for the educational level of the participants, 9 were university students, 11 were university graduates, 7 had a PhD degree, and 1 had an MA degree. Among all the participants, 9 were studying German language and literature or German translation and interpreting in Turkey, 9 were instructors of German at universities, 3 participants were employees of a firm, 2 were working as German language specialists at ministries, and 5 were unemployed at the moment of the data collection.

A total of 22 participants were born in Germany and 6 were born in Turkey. Among those 6 participants who were born in Turkey, the age during the move to Germany ranged from 3 to 7 years. Thus, the average age of the returnee participants during the move to Germany was approximately 1 year ($M = 1.04$, $SD = 2.16$).

Both parents of all the participants were Turkish and had lived in the Central Anatolian region before moving to Germany. The communication among the family members was in Turkish. The onset of participants' formal contact with the German language varied from the age of 3 years, when they started a German kindergarten, to the age of 6 years, when they started a primary school in Germany ($M = 4.46$, $SD = 1.78$). As for secondary school education, 12 participants reported that they completed their secondary school education in Germany, 9 studied the final 2 years of secondary school in Turkey, and 7 participants finished their secondary school in Turkey. As for high school education, 24 participants received it in Turkey, 3 in Germany, and 1 started a high school in Germany but finished in Turkey. All the participants pursued their university education at various departments in Turkey.

All the participants stated that they had used Turkish while they were in Germany to varying degrees from “seldom” to “always.” The participants also indicated the social groups (parents, relatives, neighbors, friends, and teachers) with whom they had been using Turkish while in Germany. The sum of the latter two variables (the frequency of Turkish use, from 1 for “seldom” to 5 for “always,” and the number of interlocutors) was defined as the perceived frequency of Turkish language use in the German-dominant context.

The participants' age when returning to Turkey varied from 7 to 20 years ($M = 13.64$, $SD = 3.18$). The length of residence of the returnee participants in Germany varied from 6 to 20 years ($M = 12.68$, $SD = 3.83$). At the time of the study, the participants had been residing in Turkey for 5–45 years ($M = 18.96$, $SD = 12.05$). All the participants reported that after returning to Turkey, they had been using Turkish daily in all public places and interacting with family members, friends, and colleagues. All the participants reported that their Turkish had improved significantly after return, and they all considered themselves as monolingual-like in Turkish in the four skills: speaking, listening, writing, and reading.

Control group

The control group consisted of 28 Turkish speakers (*Women* = 22 and *Men* = 6) who had lived all their lives in Turkey and whose ages varied from 18 to 60 years ($M = 30.17$, $SD = 10.56$). As for the educational level of the participants, 3 were university students, 9 were university graduates, 9 had an MA degree, and 7 had a PhD degree. All the participants in the control group were from the Central Anatolian region.

Data analysis and results

C-test

Descriptive statistics were used to examine the performance of the returnee participants and the control group on the c-test (see [Table 1](#)).

The data analysis showed that the mean score of the returnee participants on the c-test was 36.21 (91%). The results of the returnee group ($M = 36.21$, $SD = 3.83$) were significantly different ($W = 134$, $p < 0.000$) from the results of the control group ($M = 39.68$, $SD = 0.61$) when compared with the help of the Wilcoxon signed-rank test ([RStudio Team, 2020](#)).

Error correction task

Descriptive statistics were used to examine the performance of the returnee participants and the control group on the ECT (see [Table 2](#)).

The data analysis showed that the mean score of the returnee participants on the error correction task was 26.14 (87%). The results of the returnee group ($M = 26.14$, $SD = 2.95$) were significantly different ($W = 181.5$, $p = 0.000$) from the results of the control group ($M = 28.8$, $SD = 1.21$) when compared with the help of the Wilcoxon signed-rank test ([RStudio Team, 2020](#)).

Furthermore, the performance of the returnee participants was examined for each of the three domains included in the error correction task separately. [Table 3](#) presents the descriptive statistics for each of the domains.

The Kruskal–Wallis rank sum test was run to see whether there was a significant difference in the performance of the returnee participants in the domains of evidentiality, direct object marking, and converbs. The Kruskal–Wallis rank sum test showed that the returnee participants performed significantly differently on the three domains ($H(2) = 32.922$, $p < 0.000$). Pairwise comparisons using the Wilcoxon

TABLE 1 Performance of the returnee and control groups on the c-test.

	Returnee group	Control group
Minimum	25.00	38.00
Maximum	40.00	40.00
Mean	36.21	39.68
SD	3.83	0.61
Median	36.5	40.00

TABLE 2 Performance of the returnee and control groups on the error correction task.

	Returnee group	Control group
Minimum	21.00	26.00
Maximum	30.00	30.00
Mean	26.14	28.8
SD	2.95	1.21
Median	26.00	29.00

rank sum test with continuity correction were used to compare the returnee participants' performance in all the three domains. The difference between the domain of converbs and the two other domains of evidentiality and direct object markings was significant ($p < 0.000$), while the performance of the returnee participants on the domain of evidentiality did not differ ($p = 0.9$) from their performance on the domain of direct object marking as shown in [Figure 1](#).

When the performance of the returnee group was compared with the results of the control group for each of the domains using the Wilcoxon rank sum test, significant differences were revealed between the returnee group and the control group in the domains of evidentiality ($W = 226$, $p = 0.004$) and direct object marking ($W = 224.5$, $p = 0.004$), while no difference was found between the groups in the converb domain ($W = 406$, $p = 0.571$).

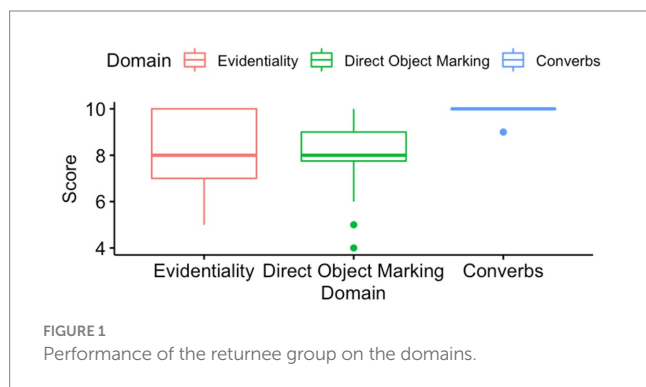
The data analysis also showed that there were five returnee participants whose scores on the tasks were compatible with the mean of the control group. Two of these participants moved to Germany at the ages of 6 and 7 years and returned to Turkey when they were 12 and 13 years old; the other two participants were born in Germany and returned to Turkey when they were 6 and 7 years old, and the last participant moved to Germany at the age of 1 year and returned to Turkey at the age 7 years. Thus, the average length of these participants' residence in Germany was approximately 6 years ($M = 6.6$, $SD = 0.548$).

In addition to these five participants, there were six other returnee participants who scored within the minimum–maximum range of the control group (*min* 65 out of 70 for both tasks). The age at the return of these six participants varied from 11 to 20 years ($M = 14.17$; $SD = 3.31$); however, all six participants reported higher than the group average values for the independent variable of frequency of Turkish language use in the migration context, which varied from 8 to 10 years ($M = 9.00$, $SD = 0.89$).

Along with the overall divergence of the returnee group from the baseline, 39% of the returnees performed compatibly with the control group. This finding, together with a pretty high standard deviation for both tasks, suggests the impact of external factors on the attainments in the reactivated HL.

TABLE 3 Performance of the returnee group on the domains.

	Evidentiality	Direct object marking	Converbs
Minimum	5.00	4.00	9.00
Maximum	10.00	10.00	10.00
Mean	8.071	8.071	9.964
SD	1.804	1.585	0.1890
Median	8.00	8.00	10.00



Effect of sociolinguistic/external factors on the returnees' attainments in the reactivated heritage Turkish

As previously mentioned, one of the primary objectives of this study is to explore the connection between the linguistic outcomes of returnees and their linguistic experiential factors. To achieve this, we employed the random forest method (Breiman, 2001) and implemented it using the *Ranger* package (Wright and Ziegler, 2017). This analysis aimed to assess how sociolinguistic factors influence the cloze test and error correction scores of the returnees.

Random forests are built upon decision trees, which employ a series of binary rules to predict a response variable. Decision trees, used with numerical and categorical response variables, are statistical models that employ recursive partitioning as their primary algorithm. Put more simply, the algorithm initially tests the association of independent variables with the response variable. If it identifies multiple independent variables associated with the response variable, the model assesses the strength of each association. The variable with the strongest association is selected for the initial binary split. For instance, if the independent variable is binary with values "M" and "F," one subset will comprise all observations with the "M" value, while the other subset will include those with the "F" value. Each subset forms a branch in the tree. This process is iteratively repeated until all independent variables have been evaluated. A random forest is constructed by aggregating a large number of decision trees. To create diverse trees, random forests employ two key procedures: bootstrap aggregating and random predictor subset selection.

Bootstrapping involves generating subsamples of the dataset with replacement, allowing each observation to be chosen more than once in a subsample. Consequently, the subsample contains two-thirds of the observations, while the remaining one-third constitutes the

out-of-bag sample. Each tree in a random forest is trained on a distinct bootstrapped sample.

Random predictor selection refers to the procedure in which the algorithm chooses a random subset of predictor variables to train each tree in the forest, denoted as "mtry." For categorical predictors, this value is typically the square root of the total number of predictor variables, whereas for continuous predictors, it is the number of predictors divided by 3 (Hastie et al., 2001; Strobl et al., 2009).

The choice of random forest over more traditional analyses, such as linear regression, was based on two main reasons: the high number of predictor variables derived from the questionnaire, which is more than the number of participants, meaning that there are more predictors than observations, a problem usually known as $p > n$. Linear regression models are not recommended in this scenario (Bühlmann and van de Geer, 2011; Chakraborty et al., 2012). The second reason is the fact that several of the questions were highly correlated. The presence of correlated variables would have made the results uninterpretable and inaccurate. Since we were interested in determining the effect of each of the variables targeted in the questionnaire, we did not want to do a principal component analysis, because this type of analysis, while taking care of the correlation among the variables, obscures the effect of the individual predictors.

Random forests can handle scenarios with more predictors than observations and manage correlated predictors. They are versatile, accommodating both continuous and categorical predictors, and are robust to variable scaling. Additionally, random forests provide variable importance rankings, helping identify the most significant predictors. The *Ranger* package's random forest implementation adds the ability to calculate value of ps, enhancing our ability to assess the statistical significance of each variable's contribution to explaining the outcome.

The *Ranger* package offers two value of p calculation methods. We opted for the Altmann method (Altmann et al., 2010), which involves performing 1,000 permutations, recommended for greater precision by the *Ranger* package creators.

We ran two random forest analyses, one for each of the two scores: cloze test and error correction. Each random forest consisted of 5,000 trees and employed the default *mtry* value, which is the square root of the number of predictors. In total, each random forest included 23 variables after removing the surplus variables. A complete list of the variables is available online.

Results of random forest models

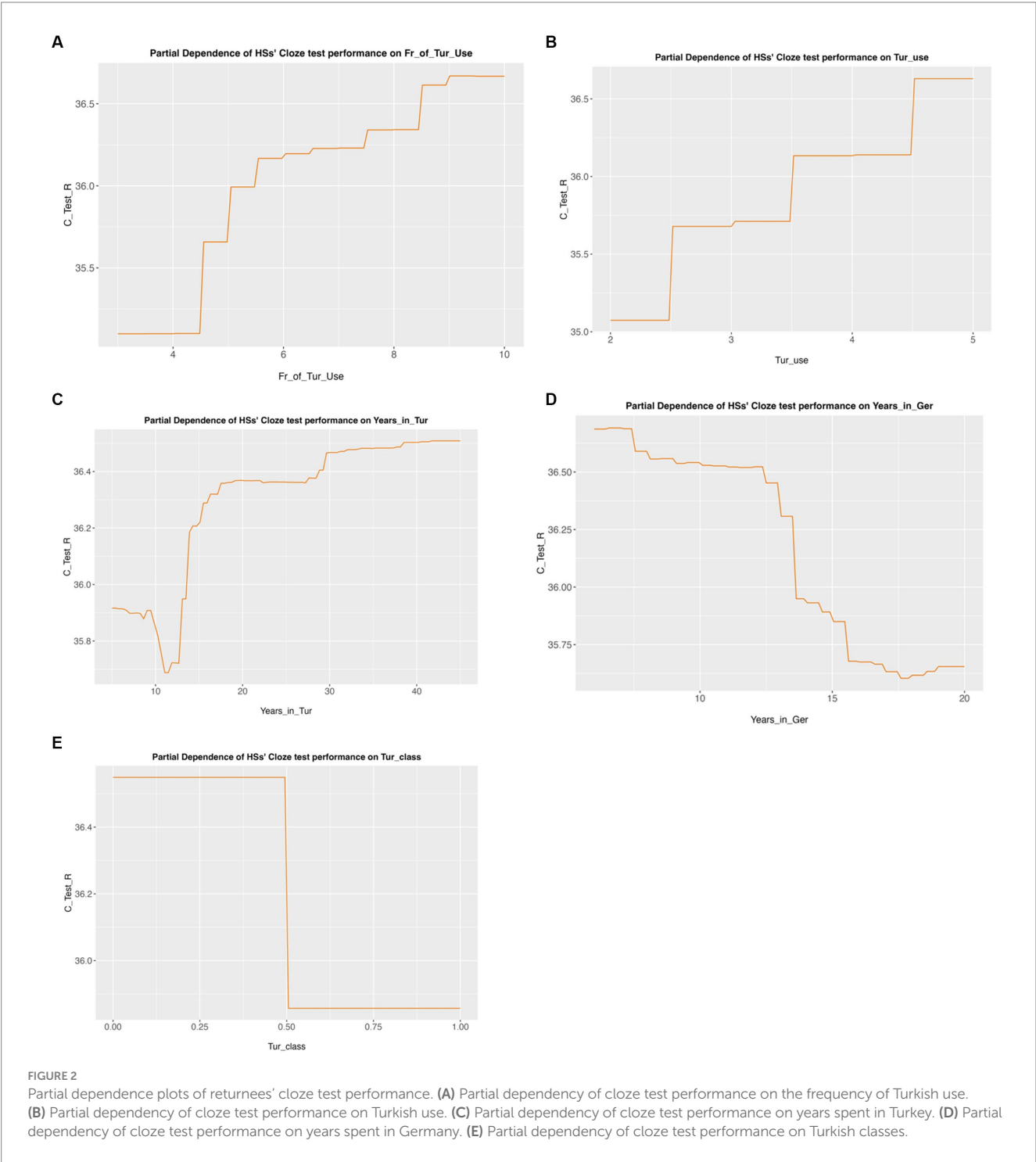
Cloze test

The next model in Table 4 shows the results of the cloze test. The model determined that *frequency of Turkish use* ($p < 0.05$) is the most important predictor of cloze test performance, followed by *Turkish use* ($p < 0.05$), *years spent in Turkey* ($p < 0.05$), *years spent in Germany* ($p < 0.05$), and *Turkish use* ($p < 0.05$).

In Figure 2, we show partial dependence plots of returnees' cloze test performance on the five variables selected as significant by the model. Figure 2A shows that as returnees' frequency of use of Turkish increases, so does their task. Similarly, in Figure 2B, it is

TABLE 4 Significant predictor variables for the cloze test.

Variable	Importance	Importance #	<i>p</i> -value	<i>p</i> significance
Fr_of_Tur_Use	1.441432	1	0.01998	<i>p</i> < 0.05
Tur_use	1.339889	2	0.023976	<i>p</i> < 0.05
Years_in_Tur	1.313394	3	0.03996	<i>p</i> < 0.05
Years_in_Ger	1.279082	4	0.028971	<i>p</i> < 0.05
Tur_class	1.008857	5	0.020979	<i>p</i> < 0.05



shown that a higher use of Turkish results in a higher task score. Figure 2C shows that if the returnees have spent more time in Turkey, then their task performance increases too. Figure 2D shows that spending more time in Germany will have a negative effect on returnees' performance. And finally, in Figure 2E, we see that those returnees who claimed that they did not have any Turkish classes outperform those who did.

Error correction

In Table 5, we show the random forest for the error correction task. The most important predictor, in this case, is the time a returnee spent in Germany ($p < 0.05$), followed by whether they had any Turkish classes ($p < 0.05$). The final most important variable is the amount of Turkish use ($p < 0.05$).

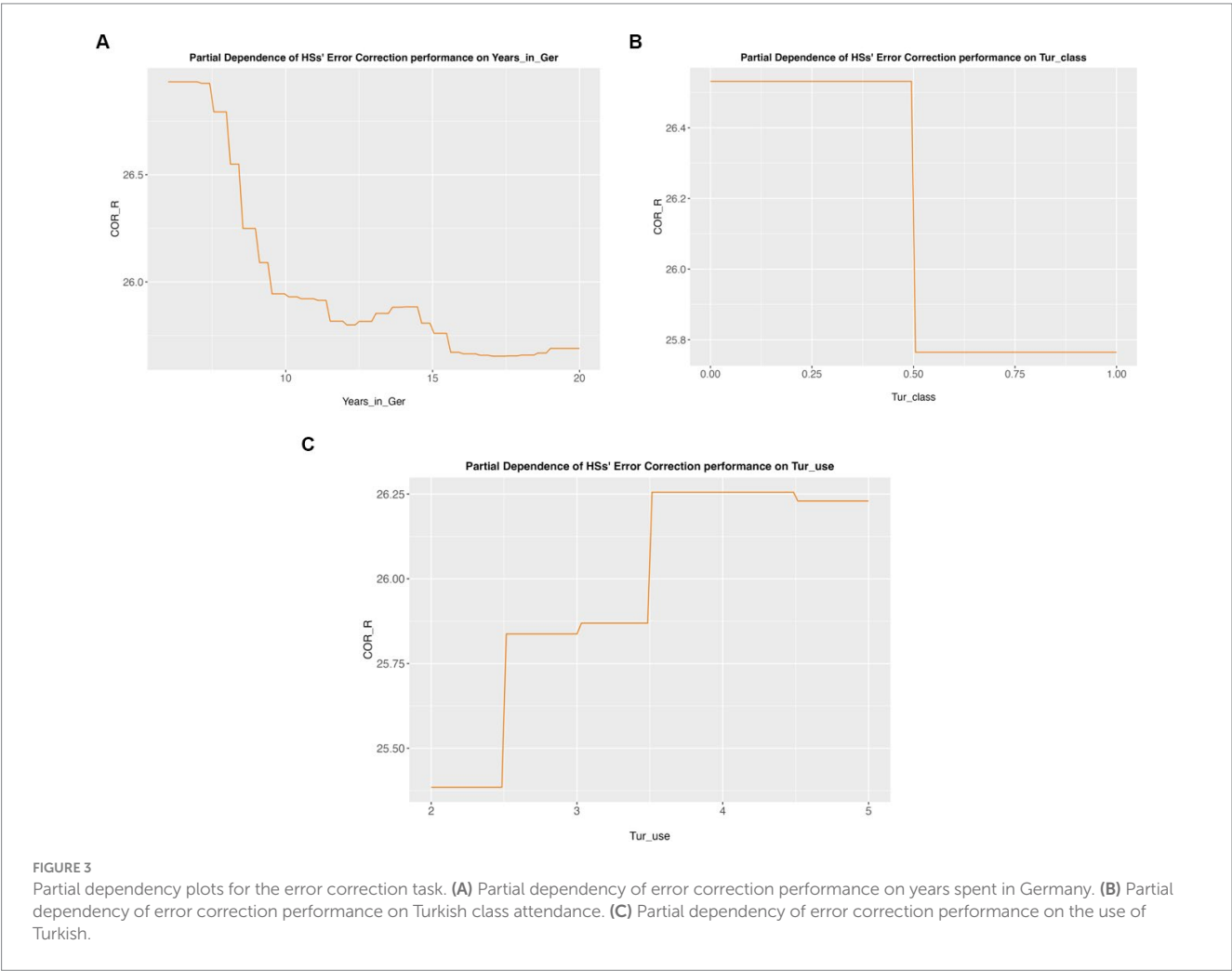
As above, Figure 3 shows the partial dependence plots for the error correction task. We observe in Figure 3A that there seems to be a negative relationship between the time spent in Germany and task performance. That is, the more time a returnee spends in Germany, the lower their error correction task performance. Similarly, in Figure 3B, we see that attending Turkish classes also has a negative effect on the returnees' performance in this task. Figure 3C shows that the more Turkish returnees use Turkish, the more likely it is that they perform better in the error correction task.

Discussion

Most of the evidence regarding individual variation in the heritage speaker bilingualism research comes from research on *typical* heritage speakers: those who are born and/or grow up in a dominant host

TABLE 5 Significant predictor variables for the error correction task.

Variable	Importance	Importance #	<i>p</i> -value	<i>p</i> significance
Years_in_Ger	1.60235613	1	0.002997003	$p < 0.05$
Tur_class	0.72749702	2	0.017982018	$p < 0.05$
Tur_use	0.4220541	3	0.047952048	$p < 0.05$



language environment with their home language(s) differing from the societal one (Rothman, 2009; Kupisch and Rothman, 2018; Polinsky and Scontras, 2019). Examining the reactivation of returnees' heritage language upon re-exposure to the home environment offers valuable contributions to our understanding of heritage language development and maintenance, and reactivation, as well as the influence of various factors on the attainment of language competence among returnee heritage speakers.

Selective vulnerability of grammatical domains

The analysis of the cloze test and error correction task revealed significant differences in performance between the returnees and the control group. These findings highlight the selective vulnerability of different grammatical domains when heritage speakers return to their country of origin. The cloze test and error correction task demonstrated that certain linguistic domains, especially when more than one domain of grammar is involved, such as interfaces of syntax and pragmatics (evidentiality and direct object marking), can be more susceptible to change, as evidenced by the significant differences between the returnee group and the control group in these domains. In contrast, other domains of the grammar morphosyntax domain, represented by converbs in this particular study, appeared to exhibit greater resilience, with no significant difference observed between the two groups. These results are in line with previous research (e.g., Treffers-Daller et al., 2007; Kaya-Soykan et al., 2023) that showed that not all linguistic structures may be (re-)activated and converge toward the baseline for granted once a HS is immersed in the environment where the HL is dominant again.

From a cross-linguistic influence perspective, the category of evidentiality is not available in German, the source of information is marked lexically (Diewald and Smirnova, 2010; Haßler, 2015), and the category of definiteness and specificity, which is involved in direct object marking in Turkish, is similarly available in German grammar (Dodd et al., 2003). Nevertheless, no benefit in the performance of the returnee participants in the domain of direct object marking has been revealed in comparison with the domain of evidentiality. Such language behavior of the returnees in their (re-)activated HL might be considered as a piece of evidence supporting the view that restricted resources of bilinguals in integrating information from different modules but not (only) cross-linguistic influence is the underlying reason for the vulnerability in different domains of grammar (Hopp, 2009; Antonova-Unlu and Wei, 2020).

Experiential factors and individual variability

Our findings offer valuable insights into the role of specific experiential factors in explaining individual variability observed with the returnee group. Our study aligns with recent trends in bilingualism (e.g., DeLuca et al., 2019; Rothman et al., 2023) as well as heritage speaker bilingualism research (e.g., Rodina et al., 2020; Bayram et al., 2021; Tomić et al., 2023), recognizing the dynamic and complex

nature of bilingual experiences and understanding heritage speakers within their own right (Polinsky and Scontras, 2019).

Role of formal language education

In this line, the findings shed light on the potential influence of formal language education in the heritage language. Returnee heritage speakers who reported not attending Turkish classes outperformed those who did in the cloze test and the error correction task. While formal language education has the potential to impact heritage language performance positively (e.g., Kupisch and Rothman, 2018; Bayram et al., 2019, 2021; Gharibi et al., 2023), the results of this study are nuanced. The negative effect observed in both tasks suggests that formal education may not always align with enhanced performance, reinforcing the necessity to develop comprehensive heritage language maintenance strategies that go beyond formal education. While formal language education can be beneficial in certain contexts, it may not always guarantee enhanced performance, suggesting that the effectiveness of these programs may vary depending on individual circumstances.

Frequency of language use and sociolinguistic factors

These findings have significant implications for heritage language maintenance and (re-)activation. The research underscores the importance of consistent language use and maintaining sociolinguistic networks. The frequency of Turkish language use emerged as a critical factor influencing the (re-)activation and proficiency in the heritage language after returning to the home language environment. Returnee participants who reported using Turkish more frequently and having a broader social network for communication in Turkish exhibited advantages in heritage Turkish (re-)activation after returning to their home country. This finding highlights the significance of creating opportunities for heritage speakers to engage in regular language use, even when they are outside the heritage language-dominant environment.

Residence in non-heritage language-dominant environments

The findings also reveal the potential challenges faced by heritage speakers who spend prolonged periods in non-heritage language-dominant environments. The length of time spent in a non-heritage language environment, as revealed by the analysis, negatively correlates with task performance (see Figure 2A). A longer residence in the context where another language is dominant implies fewer opportunities to get sufficient input into and use of the HL. This suggests that heritage speakers may experience difficulties in maintaining their heritage language competence when exposed to prolonged periods in a non-heritage language-dominant environment. These findings are consistent with numerous previous studies (Hoff and Naigles, 2002; Gutiérrez-Clelle and Kreiter, 2003; Gathercole and Thomas, 2005; Blom, 2010; Rodina and Westergaard, 2015; Unsworth, 2015; Correia and Flores, 2017 among others) demonstrating that the amount of input is a significant predictor of

language development, especially in contexts where the language is not supported by the community. It is possible to suggest that the returnee participants who reported that they had used Turkish more frequently and had a wider social network to communicate in Turkish acquired their heritage Turkish in Germany better and, by extension, had an advantage in the HL (re-)activation after their return to the country of origin.

Conclusion

In summary, the findings from this study offer important implications for heritage language maintenance and (re-)activation. The results underscore the dynamic and multifaceted nature of heritage language development, emphasizing the influence of both internal linguistic structures and external factors such as language use frequency and formal education. These findings offer valuable insights for developing strategies to support heritage speakers in preserving, activating, and enhancing their heritage language proficiency.

Ultimately, this study contributes to the broader discourse on heritage language development and acknowledges the unique linguistic journeys that heritage speakers undertake. The research encourages a more comprehensive and individualized approach to heritage language (re-)activation, recognizing that heritage speakers are not simply recipients of their linguistic environment but active agents in shaping their bilingualism.

References

- Aikhenvald, A. Y. (2004). *Evidentiality*. Oxford: Oxford University Press.
- Akkus, M., Sagin Simsek, C., and Backus, A. (2017). Türkçe-Hollandaca: İki dilli Bireylerde Kullanım-Odaklı Dilbilim Bağlamında Zarf-Fiil Yapılarının Kullanımı. Proceedings 31. Ulusal Dilbilim Kurultayı (pp. 10–15).
- Aksu-Koç, A. (1988). *The acquisition of aspect and modality: the case of past reference in Turkish*. Cambridge: Cambridge University Press.
- Aksu-Koç, A., and Slobin, D. I. (1986). "A psychological account of the development and use of evidentials in Turkish" in *Advances in discourse processes* 20. eds. W. Chafe and J. Nichols (Norwood, NJ: Ablex), 159–167.
- Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Q.* 4, 219–227. doi: 10.2307/3586211
- Altmann, A., Tolosi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics* 26, 1340–1347. doi: 10.1093/bioinformatics/btq134
- Altunkol, E., and Balci, B. (2013). The usage of Turkish grammatical morphemes by learners of Turkish as a second language. *Atiner's Conference Papers Series*, No: Lng2013-0711.
- Anderson, J. A., Mak, L., Keyvani Chahi, A., and Bialystok, E. (2018). The language and social background questionnaire: assessing degree of bilingualism in a diverse population. *Behav. Res. Methods* 50, 250–263. doi: 10.3758/s13428-017-0867-9
- Antonova-Unlu, E. (2015). Testing the Interface hypothesis: the evidence from fossilized errors in the use of Turkish case markers. *Dilbilim Araştırmaları Dergisi* 26, 1–23.
- Antonova-Unlu, E., and Wei, L. (2018). Examining the effect of reduced input on language development: the case of gender acquisition in Russian as a non-dominant and dispreferred language by a bilingual Turkish-Russian child. *Int. J. Biling.* 22, 215–233. doi: 10.1177/1367006916666390
- Antonova-Unlu, E., and Wei, L. (2020). Examining possible sources of L2 divergence at the pragmatics interface: Turkish accusative in the end-state grammar of L1 Russian and L1 English users of L2 Turkish. *Lingua* 244:102868. doi: 10.1016/j.lingua.2020.102868
- Armon-Lotem, S., and Meir, N. (2019). The nature of exposure and input in early bilingualism. In HouwerA. De and L. Ortega (Eds.), *The Cambridge handbook of bilingualism* (pp. 193–212). Cambridge: Cambridge University Press
- Arslan, S., and Bastiaanse, R. (2014). Tense/aspect and evidentiality in narrative speech of Turkish-Dutch bilingual speakers. In Proceedings of the 17th International Conference on Turkish Linguistics. France: University of Rouen.
- Arslan, S., Bastiaanse, R., and Felser, C. (2015). Looking at the evidence in visual world: eye-movements reveal how bilingual and monolingual Turkish speakers process grammatical evidentiality. *Front. Psychol.* 6:1387.
- Azar, B. (2007). Grammar-based teaching: a practitioner's perspective. *Tesl-ej* 11:n2.
- Bayram, F., (ed). (2020). *Studies in Turkish as a heritage language*. Amsterdam: John Benjamins
- Bayram, F., Pisa, G., Rothman, J., and Slabakova, R. (2021). "Current trends and emerging methodologies in charting heritage language grammars" in *The Cambridge handbook of heritage languages and linguistics*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press), 545–578.
- Bayram, F., Rothman, J., Iverson, M., Kupisch, T., Miller, D., Puig-Mayenco, E., et al. (2019). Differences in use without deficiencies in competence: passives in the Turkish and German of Turkish heritage speakers in Germany. *Int. J. Biling. Educ. Biling.* 22, 919–939. doi: 10.1080/13670050.2017.1324403
- Blom, E. (2010). Effects of input on the early grammatical development of bilingual children. *Int. J. Biling.* 14, 422–446. doi: 10.1177/1367006910370917
- Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324
- Bühlmann, P., and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Berlin: Springer Science & Business Media.
- Chafe-Nichols, W. (1986). *Evidentiality: the linguistic coding of epistemology*. Norwood, NJ: Ablex.
- Chakraborty, S., Ghosh, M., and Mallick, B. K. (2012). Bayesian nonlinear regression for large p small n problems. *J. Multivar. Anal.* 108, 28–40. doi: 10.1016/j.jmva.2012.01.015
- Chapelle, C. A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second. Lang. Res.* 10, 157–187. doi: 10.1177/026765839401000203
- Correia, L., and Flores, C. (2017). The role of input factors in the lexical development of European Portuguese as a heritage language in Portuguese-German bilingual speakers. *Languages* 2:30. doi: 10.3390/languages2040030
- Coupe, A. R. (2006). "Converbs" in *Encyclopaedia of language and linguistics*, vol. 3. ed. K. Brown (Amsterdam: Elsevier Science), 145–152.
- Daller, H., and Yıldız, C. (1995). "Die türkischen Sprachfähigkeiten von Rückkehrern aus Deutschland (the Language proficiency of Turkish returnees from Germany)" in *Zwischen den Sprachen. Sprachgebrauch, Sprachmischung und Sprachfähigkeiten*

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

EA-U: data collection, handling and analysis, conceptualisation, and writing. FB: data analysis, conceptualisation, and writing. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

- türkischer Rückkehrer aus Deutschland. [Between the languages. Language use, language mixing and language proficiency of Turkish returnees from Germany]. eds. J. Treffers-Daller and H. Daller, vol. 2 (The Language Center: Bogaziçi University, Turkey), 83–94.
- De Bruin, A., Carreiras, M., and Duñabeitia, J. A. (2017). The BEST dataset of language proficiency. *Front. Psychol.* 8, 522–536. doi: 10.3389/fpsyg.2017.00522
- DeLuca, V., Rothman, J., Bialystok, E., and Platsikas, C. (2019). Redefining bilingualism as a spectrum of experiences that differentially affects brain structure and function. *Proc. Natl. Acad. Sci.* 116, 7565–7574. doi: 10.1073/pnas.1811513116
- Diewald, G., and Smirnova, E. (2010). *Evidentiality in German: linguistic realization and regularities in grammaticalization*. Berlin: Walter de Gruyter.
- Dodd, B. J., Whittle, R., Zoer, H., Klapper, J., and Eckhard-Black, C. (2003). *Modern German grammar: a practical guide*. London: Psychology Press.
- Dörnyei, Z., and Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Lang. Test.* 9, 187–206. doi: 10.1177/026553229200900206
- Enç, M. (1991). The semantics of specificity. *Linguist. Inq.* 22, 1–25.
- Flores, C. (2020). Attrition and reactivation of a childhood language: the case of returnee heritage speakers. *Lang. Learn.* 70, 85–121. doi: 10.1111/lang.12350
- Flores, C., and Rato, A. (2016). Global accent in the Portuguese speech of heritage returnees. *Heritage Lang. J.* 13, 161–183. doi: 10.46538/hlj.13.2.5
- Flores, C., and Snape, N. (2021). “Language attrition and heritage language reversal in returnees” in *The Cambridge handbook of heritage languages and linguistics (on-line first)*. eds. S. Montrul and M. Polinsky (Cambridge: Cambridge University Press)
- Gathercole, V. C. M., and Thomas, E. M. (2005). Minority language survival: input factors influencing the acquisition of Welsh. In *Proceedings of the 4th International Symposium on Bilingualism*, 74–84. Somerville, MA: Cascadia Press.
- Gharibi, K., Bayram, F., and Guajardo, G. (2023). Lexical and morphosyntactic variation in Persian heritage language outcomes. *Linguist. Approaches Biling.* doi: 10.1075/lab.21052.gha
- Goksel, A., and Kerslake, C. (2005). *Turkish: a comprehensive grammar*. London: Routledge.
- Gürel, A. (2000). “Missing case inflection: implications for second language acquisition” in *Proceedings of the 24th annual Boston University conference on language development* 45. eds. C. Howell, S. A. Fish and K. L. Tea (Somerville, MA: Cascadia Press), 379–390.
- Gutiérrez-Clelle, V. F., and Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Appl. Psycholinguist.* 24, 267–288.
- Haspelmath, M. (1995). “The converb as a cross-linguistically valid category” in *Converbs in cross-linguistic perspective: structure and meaning of adverbial verb forms-adverbial participals, gerunds*. eds. M. Haspelmath and E. König (Berlin: Mouton de Gruyter), 3–55.
- Haßler, G. (2015). Evidentiality and the expression of speaker’s stance in romance languages and German. *Discourse Stud.* 17, 182–209. doi: 10.1177/1461445614564522
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2001). *The elements of statistical learning: data mining, inference, and prediction*. Springer-Verlag, New York.
- Hoff, E., and Naigles, L. (2002). How children use input to acquire a lexicon. *Child Dev.* 73, 418–433.
- Hopp, H. (2009). The syntax-discourse interface in near-native L2 acquisition: off-line and on-line performance. *Biling. Lang. Cogn.* 12, 463–474. doi: 10.1017/S1366728909990253
- Johanson, L. (2006). Two approaches to specificity. In L. Kulikov, A. Malchukov and SwartP. de, eds., *Case, valency and transitivity* (pp. 247–256). Amsterdam: John Benjamins Publishing Company
- Kaya-Soykan, D., Antonova-Unlu, E., and Sagin-Simsek, C. (2023). The production and perception of Turkish evidentiality markers by Turkish-German returnees. *Appl. Linguist. Rev.* 14, 251–270. doi: 10.1515/applirev-2020-0042
- Klein-Braley, C. (1985). A cloze-up on the C-test: a study in the construct validation of authentic tests. *Lang. Test.* 2, 76–104. doi: 10.1177/026553228500200108
- Koller, G., and Zahn, R. (1996). Computer based construction and evaluation of C-tests. In R. Grotjahn (Ed.), *Der C-Test: Theoretische Grundlagen und praktische Anwendungen*, 3, 401–418. Bochum, Germany: Brockmeyer.
- Kornfilt, J. (1997). *Turkish*. London: Routledge.
- Kubota, M., Chondrogianni, V., Clark, A. S., and Rothman, J. (2021). Linguistic consequences of toing and froing: factors that modulate narrative development in bilingual returnee children. *Int. J. Biling. Educ. Biling.* 1, 1–19, 2363–2381. doi: 10.1080/13670050.2021.1910621
- Kupisch, T., and Rothman, J. (2018). Terminology matters! Why difference is not incompleteness and how early child bilinguals are heritage speakers. *Int. J. Biling.* 22, 564–582. doi: 10.1177/1367006916654355
- Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The language experience and proficiency questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* doi: 10.1044/1092-4388(2007/067)
- Marian, V., and Hayakawa, S. (2021). Measuring bilingualism: the quest for a “bilingualism quotient”. *Appl. Psycholinguist.* 42, 527–548. doi: 10.1017/S0142716420000533
- Montrul, S. (2016). *The acquisition of heritage languages*. Cambridge, UK: Cambridge University Press.
- Montrul, S. (2022). *Native speakers, interrupted: differential object marking and language change in heritage languages*. Cambridge: Cambridge University Press.
- Montrul, S., and Polinsky, M. (2011). Why not heritage speakers? *Linguist. Approaches Biling.* 1, 58–62. doi: 10.1075/lab.1.1.07mon
- Nedjalkov, I. (1998). “Converbs in the languages of Europe,” in *Adverbial constructions in the languages of Europe*. ed. J. Auwera (Berlin, New York: De Gruyter Mouton), 421–456.
- Polinsky, M. (2018). *Heritage languages and their speakers*. Cambridge: Cambridge University Press.
- Polinsky, M., and Scontras, G. (2019). Understanding heritage languages. *Biling. Lang. Cogn.* 23, 4–20. doi: 10.1017/S1366728919000245
- Polinsky, M., and Scontras, G. (2020). A roadmap for heritage language research. *Biling. Lang. Cogn.* 23, 50–55. doi: 10.1017/S1366728919000555
- Rehbein, J., and Herkenrath, A. (2015). Converbs in monolinguals’ and bilinguals’ Turkish. *Ankara Papers in Turkish and Turkic Linguistics* 493–513.
- Rodina, Y., Kupisch, T., Meir, N., Mitrofanova, N., Urek, O., and Westergaard, M. (2020). Internal and external factors in heritage language acquisition: evidence from heritage Russian in Israel, Germany, Norway, Latvia and the United Kingdom. *Front. Educ.* 5, 20–35. doi: 10.3389/educ.2020.00020
- Rodina, Y., and Westergaard, M. (2015). Grammatical gender in Norwegian: language acquisition and language change. *J. Ger. Linguist.* 27, 145–187. doi: 10.1017/S1470542714000245
- Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: romance languages as heritage languages. *Int. J. Biling.* 13, 155–163. doi: 10.1177/1367006909339814
- Rothman, J., Bayram, F., DeLuca, V., Di Pisa, G., Dunabeitia, J. A., Gharibi, K., et al. (2023). Monolingual comparative normativity in bilingualism research is out of “control”: arguments and alternatives. *Appl. Psycholinguist.* 44, 316–329. doi: 10.1017/S0142716422000315
- RStudio Team (2020). *RStudio: integrated development for R*. Boston, MA: RStudio, PBC. Available at: <http://www.rstudio.com/>.
- Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol. Methods* 14:323. doi: 10.1037/a0016973
- Tomić, A., Rodina, Y., Bayram, F., and De Cat, C. (2023). Documenting heritage language experience using questionnaires. *Front. Psychol.* 14:1131374. doi: 10.3389/fpsyg.2023.1131374
- Treffers-Daller, J., Daller, M., Furman, R., and Rothman, J. (2016). Ultimate attainment in the use of collocations among heritage speakers of Turkish in Germany and Turkish-German returnees. *Biling. Lang. Cogn.* 19, 504–519. doi: 10.1017/S1366728915000139
- Treffers-Daller, J., Özsoy, A. S., and Van Hout, R. (2007). (In) complete acquisition of Turkish among Turkish-German bilinguals in Germany and Turkey: An analysis of complex embeddings in narratives. *Int. J. Biling. Educ. Biling.* 10, 248–276.
- Turan, D., Antonova-Unlu, E., Sagin-Simsek, C., and Akkus, M. (2020). Looking for contact-induced language change: converbs in heritage Turkish. *Int. J. Biling.* 24, 1035–1047. doi: 10.1177/1367006920926263
- Unsworth, S. (2015). “Amount of exposure as a proxy for dominance in bilingual language acquisition” in *Language dominance in bilinguals: issues of measurement and operationalization*. eds. C. Silva-Corvalán and J. Treffers-Daller (Cambridge: Cambridge University Press), 156–173.
- Uzun, E. (2018). *Hitit*. Ankara University Press.
- Wright, M. N., and Ziegler, A. (2017). Ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.* 77, 1–17. doi: 10.18637/jss.v077.i01
- Yoshitomi, A. (1999). “On the loss of English as a second language by Japanese returnee children” in *Second language attrition in Japanese contexts*. ed. L. Hansen (Oxford, UK: Oxford University Press), 80–113.

Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

[See more →](#)

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact

