frontiers | Frontiers in Language Sciences

TYPE Original Research
PUBLISHED 03 January 2025
DOI 10.3389/flang.2024.1400587

OPEN ACCESS

EDITED BY
Susanne Brouwer,
Radboud University, Netherlands

REVIEWED BY
Clara Fridman,
Bar-Ilan University, Israel
Chantal Van Dijk,
Technical University of
Braunschweig, Germany

*CORRESPONDENCE
Alicia Luque
✉ aluque@nebrija.es

RECEIVED 13 March 2024
ACCEPTED 14 November 2024
PUBLISHED 03 January 2025

CITATION
Luque A, Koronkiewicz B, Issa B, Faretta-Stutenberg M and Bowden HW (2025) Ecological validity and inclusivity in heritage bilingualism research: Examining objective and subjective Spanish proficiency assessments and language experience factors. *Front. Lang. Sci.* 3:1400587. doi: 10.3389/flang.2024.1400587

Ecological validity and inclusivity in heritage bilingualism research: Examining objective and subjective Spanish proficiency assessments and language experience factors

Alicia Luque[1,2,3]*, Bryan Koronkiewicz[4], Bernard Issa[5], Mandy Faretta-Stutenberg[6] and Harriet Wood Bowden[5]

[1]Department of Applied Language Studies, Nebrija University, Madrid, Spain, [2]Nebrija Research Center in Cognition, Nebrija University, Madrid, Spain, [3]Department of Language and Culture (ISK), UiT The Arctic University of Norway, Tromsø, Norway, [4]Department of Modern Languages and Classics, The University of Alabama, Tuscaloosa, AL, United States, [5]Department of World Languages and Cultures, The University of Tennessee, Knoxville, Knoxville, TN, United States, [6]Department of World Languages and Cultures, Northern Illinois University, DeKalb, IL, United States

The multidimensional nature of bilingualism demands ecologically valid and inclusive research methods that can capture its dynamism and diversity. This is particularly relevant when assessing language proficiency in minoritized and racialized communities, including heritage speakers (HSs). Motivated by a paradigm shift in bilingualism research, the present study joined current efforts to establish best practices for assessing language proficiency among bilingual individuals accurately and consistently, promoting ecological validity and inclusivity. Specifically, we examined the reliability and validity of objective and subjective proficiency assessments ubiquitously used in second language (L2) and bilingualism research to assess Spanish proficiency, within a sample of HSs of Spanish in the United States (US). We also sought to understand the relationships between these proficiency assessments and a subset of heritage language (HL) experience factors. To our knowledge, this is the first study to examine the reliability and validity of these proficiency assessments and their relationship with HL experience factors with HSs of Spanish in the US in a multidimensional way. Forty-three HSs of Spanish completed the Bilingual Language Profile questionnaire, including self-reports of proficiency and information about HL experience and two objective proficiency assessments: a lexical decision task, namely the LexTale-Esp, and a vocabulary and grammar task, often referred to as the "Modified DELE". Our findings revealed high internal consistency for both objective proficiency assessments and medium correlations between them, supporting their reliability and validity. However, our results also revealed inconsistent relationships between subjective proficiency assessments and HL language experience factors. These findings underscore the dynamic interplay between these HSs' objective and subjective proficiency, and HL experiences and use across different contexts. Additionally, they highlight the limitations of relying on any single proficiency assessment, aligning with previous research that emphasizes the need for multidimensional proficiency assessments and language experience factors to capture the dynamic and diverse nature of bilingualism. By critically evaluating the reliability and validity of existing objective

and subjective proficiency assessments alongside HL experience factors, our study aims to shed light on the best practices of assessing language proficiency among bilingual individuals, specifically HSs of Spanish in the US, in an ecologically valid and inclusive manner.

KEYWORDS

heritage bilingualism, ecological validity, language proficiency assessment, bilingual experience factors, inclusivity in bilingualism

1 Introduction

Bilingualism, characterized by regular engagement with two (or more) languages in daily life– regardless of level of proficiency in each language–, is a subject of profound academic interest due to its influence on multiple dimensions of the human experience, including identity, language development, communication, sociocultural engagement, and neuro/psychological functioning (e.g., Birdsong, 2014; Dewaele et al., 2003; Grosjean, 2010). Within this fascinating landscape, heritage language (HL) bilingualism occupies a distinct and significant place, emerging as a complex and dynamic phenomenon, reflecting the experiences of individuals who grow up speaking a native language, their HL, which differs from the dominant language in their wider societal context.

Heritage speakers (HSs) constitute a unique group of bilinguals. Although these speakers acquire their HL early and naturalistically, they often navigate a sociolinguistic landscape characterized by challenges. These include reduced linguistic input (especially written input in academic contexts), and/or have fewer opportunities to meaningfully engage with, use, or be formally trained in their HL (e.g., Flores, 2015; Rothman and Treffers-Daller, 2014; Valdés, 2005). Additionally, HLs are frequently marginalized within broader societal contexts, facing systematic neglect in educational, governmental, and cultural domains. This marginalization is deeply intertwined with raciolinguistic ideologies that both reflect and reinforce societal hierarchies based on race and language (e.g., Rosa and Flores, 2017; Zou and Cheryan, 2017). For example, within the United States (US), HSs of Spanish often encounter policies and practices that prioritize English proficiency and use, sometimes to the detriment of their HL development. This can manifest in educational settings where English-only instruction predominates, limiting opportunities for HL development and contributing to potential loss over time (e.g., Beaudrie and Fairclough, 2012; Christoffersen, 2019; Flores and García, 2017; García and Solorza, 2021; Kelly, 2018; Lee and Wright, 2014; Leeman, 2015; Leeman and Martínez, 2007; Sánchez-Muñoz, 2016). Furthermore, societal attitudes toward Spanish and bilingualism are mixed, with some segments of society viewing Spanish and/or bilingualism as an asset, while others may perceive it negatively (e.g., Achugar and Pessoa, 2009; Barrett et al., 2023; Fuller and Leeman, 2020; Surrain and Luk, 2023). Additionally, the availability of educational opportunities for HSs of Spanish may vary depending on factors such as geographic location, socioeconomic status, and access to resources, leading to disparities in outcomes (e.g., Bohman et al., 2010; Paradis, 2023; Rothman, 2009). Navigating this complex sociolinguistic landscape presents unique challenges for HSs of Spanish, as they strive to maintain their HL and bicultural identity while also adapting to the linguistic and cultural hegemonic norms of their environment (e.g., Holguín Mendoza et al., 2023; Pascual y Cabo and Prada, 2018).

We acknowledge that, as researchers, we have an obligation to contribute knowledge that can help address these challenges to promote linguistic diversity, cultural preservation, and equitable educational opportunities for HSs of Spanish in the US (e.g., Flores, 2020; Flores and Rosa, 2015, 2023; Flores and Schissel, 2014; García et al., 2021).

For researchers and practitioners working with HSs, assessing language proficiency takes on a multifaceted character and involves considering cultural identity, communication, and sociolinguistic engagement (e.g., Pascual y Cabo and Prada, 2015; Valdés, 2005). However, many tasks utilized to evaluate language proficiency rely on standardized assessments based on monolingual benchmarks, prioritizing prescriptive linguistic norms and language usage (e.g., Bachman and Palmer, 1996; Bayram et al., 2021b; Cummins, 2013). While these proficiency assessments can offer useful data for the purposes of HL bilingualism research, they also present limitations in capturing the rich diversity and dynamic nature of bilingual experiences and the sociocultural and linguistic abilities of bilingual individuals in a holistic way. Moreover, the exclusive use of such tasks can inadvertently perpetuate negative stereotypes and disregard the sociocultural dimensions inherent in bilingualism, which are a core part of HSs' lived experiences (Flores and Rosa, 2015; Ortega, 2020).

Recognizing these limitations, there has been a notable paradigm shift in bilingualism research to establish best practices for assessing language proficiency among bilingual individuals while promoting ecological validity and inclusivity (e.g., De Bruin, 2019; López et al., 2023). Within this context, ecological validity refers to the extent to which research findings about bilingualism apply to real-world bilingual settings, such that results can be generalized to everyday bilingual experiences beyond the controlled conditions of a research laboratory. Inclusivity ensures that bilingual individuals' diverse experiences and backgrounds are accurately represented and respected in research. An integral part of this shift acknowledges that bilingual individuals are not simply two monolinguals in one person; instead, bilingualism is viewed as multifaceted and dynamic, including unique phenomena such as code-switching and translanguaging, where speakers fluidly alternate between languages across conversations and/or contexts. These practices, inherent to the bilingual experience, reflect the adaptive nature of bilingualism across diverse contexts of language use, contexts which are essential to understanding the full breadth of bilingual realities. To advance this understanding, researchers have begun to propose and incorporate new methodologies that

better capture the diverse and dynamic language proficiencies and experiences of bilingual individuals. By prioritizing ecological validity and inclusivity, these new approaches aim to reflect the complex nature and dynamics of bilingualism more accurately, making research findings more relevant and applicable to real-world settings (e.g., Ali, 2023; Bayram et al., 2019, 2021a; Cacoullos and Travis, 2018; Grosjean, 1989, 2010; Gullifer et al., 2021; Higby et al., 2023; Leivada et al., 2023; Prada, 2021, 2022; Rothman et al., 2023; Toribio and Duran, 2018).

Despite this shift, there remains a lack of consensus in the field regarding which proficiency assessments best capture the multifaceted nature of bilingualism in an accurate and consistent way, especially for HSs. Furthermore, the wide variety of proficiency measurements used across studies limits the generalizability of results and complicates cross-study comparability, thereby hindering the advancement of knowledge in the field (Olson, 2023a). Thus, the first step to creating best practices for assessing proficiency in HSs is to better examine and understand these various proficiency assessments, both their reliability and validity. Key aspects, such as internal consistency within reliability, and construct and ecological validity within overall validity, play crucial roles in this process. Note that reliability does not tell us the specific nature of what is being measured—only that the measurement is consistent across items. In contrast, construct validity focuses on determining whether the test accurately measures the intended concept or construct, while ecological validity examines how well the test results apply to real-world contexts (Brown, 2013; Crocker and Algina, 1986; Kline, 2013; Tavakol and Dennick, 2011). These aspects can be evaluated by assessing the internal consistency of the test items to ensure reliability, examining how well the test relates to other assessments designed to measure the same construct, which supports construct validity, and by assessing how well the results reflect real-life situations, which is crucial for ecological validity. Through these comprehensive examinations, researchers can better identify reliable and valid proficiency assessments.

Enhancing the robustness of bilingualism research depends significantly on using the most reliable and valid methodologies, including proficiency assessments. This approach facilitates knowledge development in the field, especially if researchers can converge on a smaller set of the most robust measures, which are then used consistently across studies. As Olson (2023a) states, "given the important role that proficiency plays in the field, and notably in the comparability of results across multiple studies, proficiency assessment remains a key methodological consideration" (p. 7). Thus, by these aspects, researchers can ensure their work contributes to a more reliable and valid understanding of bilingual proficiency, advancing the field in a more ecologically valid and inclusive way.

Our study aimed to contribute to this effort by examining the reliability and validity of a set of objective and subjective proficiency assessments focusing on a specific group of bilinguals: HSs of Spanish in the US. Specifically, we investigated the reliability and validity of two widely used objective and subjective proficiency assessments in the field of L2 and bilingualism research. These were a lexical decision task, in particular, the Lextale-Esp (Izura et al., 2014), and a vocabulary-grammar task often called the "Modified DELE" (Montrul and Ionin, 2012). The subjective assessments were derived from proficiency self-reports in the Bilingual Language Profile questionnaire (BLP; Birdsong et al., 2012). Our goal was to examine these tasks within a sub-group of college-educated HSs of Spanish, who were not the target population for which the assessments were developed.

Additionally, as part of our assessment of ecological validity, we sought to understand the relationships between these proficiency assessments and particular HL experience factors, including years of exposure to Spanish, years of Spanish schooling, and the social diversity of bilingual language use, as assessed by language entropy (following Gullifer and Titone, 2018). This multidimensional approach aimed to shed light on the effectiveness of these assessments in capturing and characterizing the dynamic and diverse nature of HSs' proficiency and experiences and of HL bilingualism in an ecologically valid and inclusive way. To our knowledge, this is the first study to examine the reliability and validity—specifically internal consistency, construct validity, and ecological validity—of these proficiency assessments and their relationship with HL experience factors with HSs of Spanish in the US.

2 Background

To enhance our ability to characterize the Spanish proficiency of HSs in the US in an ecologically valid and inclusive manner, it is crucial to disentangle the different concepts and terms that are used in the field to describe and evaluate the receptive and productive linguistic proficiency of HSs. In this section, we begin by unpacking the concept of proficiency, a term that is ubiquitous in the literature, but often used without a clear operationalization. Following, we review studies that have explored interactions between bilingual language proficiency and exposure, comparing objective and subjective assessments of proficiency across different modalities and bilingual populations. Finally, we review previous research that has provided insights into the reliability and validity of the objective and subjective proficiency assessments examined in our study.

A recent review of proficiency assessment methods in bilingualism research (Olson, 2023a) traces the evolution of definitions of the term proficiency. Such definitions begin with notions of general competence in a language (e.g., Thomas, 1994) and expand to communicative competence in different sociocultural contexts (e.g., Canale and Swain, 1980; Hymes, 1972). Other definitions conceptualize this construct as (at least) two-dimensional, composed of a linguistic knowledge dimension (e.g., morphosyntactic, lexical) and a language skills dimension (reading, writing, speaking, listening) (e.g., Carroll and Freedle, 1972). More recent conceptualizations merge linguistic knowledge, language abilities or skills, and communicative competence into multidimensional models (e.g., Hulstijn, 2015; Hyltenstam, 2016). These models ultimately converge in the notion that proficiency is a combination of skills and knowledge that allow speakers to comprehend and produce language successfully (Olson, 2023a).

Several different types of assessment methods have been used to characterize proficiency in bilinguals, including (a) standardized language-specific tests such as the TOEFL, (b) self-ratings, (c) area-specific tests (e.g., vocabulary tests, picture-naming tasks, etc.),

(d) multiple component tests (e.g., an elicited imitation task), (e) holistic assessments such as the Oral Proficiency Interview, or (f) characterization based on curricular level (Olson, 2023a). Each of these approaches has its theoretical or practical justification but also has specific methodological limitations [as discussed by Menke and Malovrh (2021); Olson (2023a)]. Assessments of proficiency among bilinguals can be (and are) used for different purposes in research, including to examine as a variable of interest, to characterize (e.g., "intermediate level"), to group, and/or to exclude participants, as well as to make cross-study comparisons (Olson, 2023a). Thus, they are critical to examine from a methodological viewpoint. As revealed by Surrain and Luk (2019), there is a tendency to oversimplify the construct of proficiency based on a single metric. This oversimplification can lead to the categorization or assignment of potentially misleading labels to bilingual speakers, overlooking the multidimensional nature of bilingual language proficiency and the diverse factors that contribute to it.

A number of studies have explored relationships between objective and subjective proficiency assessments to evaluate and characterize different aspects of language proficiency among diverse bilingual populations with varying results. For instance, Gollan et al. (2012) investigated language dominance among Spanish-English bilinguals, including young and older adults. Their study examined the Multilingual Naming Test (MINT) and the Boston Naming Test (BNT), both picture-naming tasks, as objective measures, and proficiency interviews and subjective self-reports of language proficiency as subjective measures. The results revealed that while self-ratings of proficiency and proficiency interviews generally aligned well with the results of the MINT in determining language dominance, the BNT often classified participants as more English-dominant than other assessments. This discrepancy is particularly significant because it highlights a key issue in bilingual language assessment: the potential for tasks originally designed for monolingual speakers, like the BNT, to misrepresent the abilities of bilingual individuals. Specifically, the BNT appeared to underestimate proficiency in Spanish, suggesting that such tools may not be fully reliable for assessing language proficiency in bilingual populations. Furthermore, the study found that a substantial portion of participants—up to 60%—performed better on tasks involving their self-reported non-dominant language. This finding suggests that bilinguals may possess a higher level of proficiency in their non-dominant language than they perceive or that certain tasks may be more sensitive to different aspects of language proficiency in dominant vs. non-dominant languages.

Similarly, Sheng et al. (2014) explored the relationship between subjective and objective assessments of language dominance among Mandarin-English bilinguals, employing similar assessments as those used by Gollan et al. (2012), such as the MINT and self-reported proficiency. Their findings echoed the earlier study in that discrepancies existed between self-reports and objective assessments. Specifically, self-ratings of language dominance did not always align with the results of the MINT, suggesting that self-perceptions of language abilities can be influenced by factors other than actual proficiency, such as cultural attitudes or confidence levels in using a particular language. A key finding from the study by Sheng et al. (2014) was that the degree of convergence or divergence between subjective and objective assessments could vary depending on the language pair and the context in which the languages are used. For instance, Mandarin-English bilinguals who used both languages regularly in different domains (e.g., Mandarin at home, English at work) were more likely to have self-ratings that diverged from their MINT results. The results of these studies highlight the importance of using multiple, carefully chosen tasks to assess bilingual proficiency comprehensively. The observed differences between subjective assessments, like self-ratings, and certain objective assessments underscore the need to use assessment methods that accurately reflect and characterize bilingual individuals' dynamic and diverse linguistic abilities. This complexity underscores that a single approach may not suffice, and a more multidimensional strategy, which integrates both subjective experiences and objective linguistic abilities, is crucial for a more accurate representation of bilingual proficiency.

Other studies have focused on the role of language experience factors in influencing the reliability and validity of proficiency assessments. Tomoschuk et al. (2019) examined the relationship between self-ratings and picture-naming tasks across Spanish-English and Chinese-English bilinguals with varying acquisition backgrounds. Their findings showed discrepancies between self-ratings and picture-naming results across different language groups, with some individuals rating their proficiency higher or lower than what was reflected in their performance on the objective task. These discrepancies suggest that individual biases or differing interpretations may influence subjective assessments, while objective assessments, such as picture-naming tasks, were more reliable indicators of proficiency. Additionally, their results underscored the importance of considering language experience factors, including the amount and context of language exposure, as these were found to impact the reliability of both subjective and objective assessments significantly. Relatedly, Gullifer and Titone (2020) explored bilingual language proficiency among French-English bilinguals (with varying experience backgrounds) using a combination of objective and subjective assessments. Objective assessments included picture-naming ability and verbal fluency tests, while subjective assessments encompassed self-reports. Their study investigated how factors such as timing and amount of HL exposure influence proficiency outcomes across different communicative contexts. The findings revealed nuanced patterns in language exposure and proficiency, indicating that subjective assessments can sometimes provide more accurate insights than expected, particularly for assessing L2 proficiency. Specifically, the study highlighted how language exposure across various communicative contexts exhibited distinct but interrelated patterns that contributed to a more comprehensive self-assessment of L2 proficiency compared to L1 proficiency.

Additionally, Gehebe et al. (2023) used both objective (ACTFL and DIALANG standardized proficiency tests) and subjective proficiency assessments (self-rated proficiency and Can-Do statements) among young adult bilinguals with varying levels of exposure to English as their L2 (and one of over a dozen different non-English languages as their L1). Their findings revealed that proficiency assessment outcomes varied based on exposure levels to the L2 and domains of language proficiency, revealing the impact of language exposure on proficiency assessments. In their study,

participants with higher English exposure demonstrated more consistent proficiency outcomes across subjective and objective assessments, supporting the validity of standardized assessments in capturing proficiency differences among this group. Yet, subjective assessments provided insights into self-perceptions and confidence in language use, complementing the quantitative data of standardized tests. Hržica et al. (2024) added another layer of complexity by examining the relationship between self-assessment of language proficiency and objective assessments of lexical diversity and syntactic complexity among bilingual HSs of Italian in Croatia. Their study specifically focused on a diglossic community, where individuals regularly navigate between a standard language (Italian) and a regional dialect (Istrovenetian) in different contexts. The findings revealed an intricate interplay between objective and subjective language proficiency assessments, indicating that although subjective assessments can provide valuable insights, they do not always fully align with objective language proficiency assessments. Specifically, the results of their study revealed that self-assessment scores were generally higher for the standard language compared to the regional dialect, reflecting the different social statuses and usage contexts of the two language varieties. However, objective assessments, such as lexical diversity and syntactic complexity, often painted a different picture, sometimes showing higher proficiency for the regional dialect, particularly in spoken contexts.

Taken together, these findings highlight the importance of considering language experience factors when assessing bilingual proficiency. A balanced approach that integrates both objective and subjective assessments is essential for capturing the full scope of bilingual language proficiency, particularly in individuals with diverse linguistic backgrounds. The complexity of bilingualism underscores the need to evaluate proficiency within multiple socio-experiential contexts. This multidimensional approach, supported by previous studies, allows for a more accurate and dynamic understanding of bilingual proficiency. While subjective assessments provide valuable insights, they may not fully capture the intricate relationship between language experience and proficiency, making it crucial to complement them with objective assessments.

Finally, it is worth noting that some studies have found self-ratings to be highly correlated with other well-documented, production-oriented, objective assessments of proficiency in bilinguals, supporting their validity. For example, robust correlations have been found with both the Elicited Imitation Task (EIT) and Simulated Oral Proficiency Interview (SOPI) for L2 learners (Bowden, 2016), and with the EIT for L2 learners and HSs (Faretta-Stutenberg et al., 2023). These findings suggest that self-ratings can serve as more reliable indicators of proficiency when aligned with certain oral and production-oriented tasks. However, it is essential to recognize that the effectiveness of self-ratings may vary depending on the specific tasks and contexts, highlighting the multifaceted nature of language proficiency and the need to consider task-specific characteristics in assessments.

These studies underscore the complexity of assessing bilingual language proficiency due to the interplay between objective and subjective assessments and varying language experience factors. Specifically, they highlight how language proficiency assessments can yield inconsistent or variable results, often influenced by different factors such as language exposure, socio-cultural contexts, and individual perceptions. This variability points to the need for a multidimensional language proficiency assessment approach that captures the full spectrum of bilingual language abilities and experiences.

Our study aimed to contribute to this line of work, highlighting the need for a multidimensional approach to proficiency assessment by focusing specifically on the Spanish proficiency of HSs in the US. As noted above, we explored the reliability and validity of both objective and subjective assessments to assess HL proficiency and examined how these assessments correlate with various HL experience factors. To achieve this, we begin with a detailed analysis of the objective assessments employed in our study, followed by a discussion of relevant prior research examining their development and validation.

2.1 Objective assessments of language proficiency

In the context of our study, we define objective proficiency assessments as a type of language assessment designed to quantify, track, or categorize an individual's language abilities in a systematic manner (Olson, 2023a). Such objective assessments are utilized to evaluate bilingual individuals' proficiency across their different languages, including standardized tests developed by language assessment organizations or researchers, or specifically designed by researchers for the purpose of their studies, and they may focus on one or more domains, such as oral, written, receptive, productive, lexical, and/or grammatical proficiency. Following, we describe the objective assessments used in our study and discuss relevant prior research examining these assessments.

2.1.1 Lexical decision task: Lextale-Esp

The Lexical Test for Advanced Learners of English (LexTALE) was initially developed by Lemhöfer and Broersma (2012) to be a practical and quick (about 5 min) objective tool to assess L2 English vocabulary knowledge. It is intended as a potential proxy to assess overall language proficiency by estimating an individual's vocabulary size. The task uses word frequency as the basic criterion for establishing varying difficulty levels across the proficiency continuum. That is, certain high-frequency words were selected so that they are known by even L2 learners on the lower end of the proficiency spectrum, while other low-frequency words were selected as they would be known only by L2 learners at the higher end. Specifically, participants are presented with a list of words (e.g., *scornful*) and English-like non-words (e.g., *mensible*) and are asked to identify whether each is an existing English word or not. The LexTALE evaluates participants' performance through signal detection theory approaches by considering participants' accurate identification of words and non-words, erroneous identification of a non-word as a word (i.e., false alarms), and failure to recognize a word (i.e., miss rate).[1] There is ample support for the task's

1 Further descriptive and technical information about the task can be found on the LexTALE website: www.lextale.com.

reliability and validity as an estimate of vocabulary size, knowledge and processing speed. This evidence comes from correlations with individual differences in language processing abilities across various task, including studies on reaction time dynamics on masked priming tasks (Andrews and Hersch, 2010), written word identification strategies (Chateau and Jared, 2000), word-recognition speed lexical-decision task accuracy (Diependaele et al., 2013), and performance on lexical decision tasks (Yap et al., 2008), among others. It has also been shown to have small to medium (all correlation strengths following Plonsky and Oswald, 2014) sized correlations with English proficiency assessments including the TOEIC and Quick Placement Test (Lemhöfer and Broersma, 2012). However, recent findings by Puig-Mayenco et al. (2023) suggest a more nuanced consideration of the LexTALE's applicability. In their study, they critically evaluated the LexTALE's validity as an assessment of global L2 proficiency across learners of English with varying proficiency levels, originating from different L1 backgrounds (Spanish and Chinese) by conducting a partial replication of the work by Lemhöfer and Broersma (2012). The results of their study revealed that the LexTALE, while offering valuable insights into vocabulary size, knowledge, and processing speed, shows only low-to-moderate correlations with a standardized assessment of English global proficiency, such as the Quick Placement Test. These findings underscore the fact that the LexTALE's applicability is not straightforward, as its correlations with other proficiency assessments seem to be inconsistent.

Mirroring the English version, a Spanish version of the task (Lextale-Esp; Izura et al., 2014) was developed to address the growing need for efficient and objective tools to assess Spanish language proficiency among bilingual populations, including HSs (e.g., Hao et al., 2024; Luque et al., 2023). Based on the design and purpose of the original LexTALE, the Lextale-Esp also evaluates vocabulary knowledge by estimating an individual's vocabulary size to gauge overall L2 proficiency. The Lextale-Esp uses a range of words that appear to be influenced by Peninsular Spanish selected from the Subtlex-Esp database (Cuetos et al., 2011), which is based on word frequencies from movies and TV shows subtitles screened between 1990 and 2009, with the same intended goal of having words with very high-frequency rates likely known by even beginning L2 learners to very low-frequency words likely only known by highly proficient *native* speakers.

Regarding its validity, the Lextale-Esp has been shown to be a valuable tool for assessing vocabulary knowledge as a proxy to assess overall language proficiency across different Spanish-speaking bilingual populations. Specifically, a study conducted by Ferré and Brysbaert (2017) supported the discriminative power of the Lextale-Esp in assessing Spanish vocabulary size and processing speed within highly proficient Catalan-Spanish bilinguals with varying degrees of language dominance. The findings showed that the two participant groups performed differently on the Lextale-Esp, with the Spanish-dominant group displaying significantly higher scores than the Catalan-dominant group Thus, these findings provide evidence supporting the Lextale-Esp's validity in capturing variability in vocabulary knowledge among highly proficient bilinguals. Further validation efforts for the Lextale-Esp come from Bermúdez-Margaretto and Brysbaert (2022), exploring translation efficiency in language assessments. Participants in the study were L1 Spanish-dominant adults who identified as bilingual speakers in 26 different languages. The goal of their study was twofold: first, to develop new assessment methods that more accurately reflect vocabulary knowledge by emphasizing meaning recognition rather than form, following the work of Vermeiren et al. (2022) and second, to explore the broader question of convergent validity, which involved assessing the extent to which their newly developed vocabulary test and the already established Lextale-Esp measured the same construct of vocabulary knowledge. Findings revealed medium-sized correlations between the Lextale-Esp and their vocabulary test, suggesting that both assessments tap into the same construct to a significant degree. These results suggest that the LexTale-Esp is specifically suited to assessing overall vocabulary knowledge. Overall, these findings support the validity and reliability of the Lextale-Esp as an objective assessment of vocabulary knowledge among bilingual individuals.

Despite the growing body of research supporting the Lextale-Esp's use across different linguistic contexts, there remains a critical need to specifically investigate its reliability and validity within the domain of HL bilingualism, especially regarding its potential to tap into individual language abilities and overall proficiency more broadly in an ecologically valid and inclusive way. Additionally, it is important to recognize that the LexTale-Esp appears to be heavily influenced by Peninsular Spanish norms, especially the low-frequency items. This poses challenges in terms of ecological validity and inclusivity, particularly for HSs of Spanish in the US given their potential lack of familiarity with this particular variety of Spanish. Such unfamiliarity could negatively impact their score on the task, potentially leading to a mischaracterization of their Spanish proficiency.

2.1.2 Spanish vocabulary and grammar task, also known as the "modified DELE"

A written Spanish vocabulary and grammar task widely used as a proficiency assessment in L2 and HL research is often referred to as the "Modified DELE". The task was in fact compiled from two sources in the 1990s by Montrul and Bruhn de Garavito (Hoot, 2020). Its first published use was in Duffield and White (1999), as a measure for grouping adult L2 Spanish learners by proficiency level. They described the task as being comprised of:

> sections from standardized Spanish as a second language proficiency tests, namely the reading/vocabulary section of the MLA Cooperative Foreign Language Test (Educational Testing Service, Princeton, NJ) and a cloze test from the *Diploma de Español como Lengua Extranjera* (DELE) (Embajada de España, Washington, DC) (p. 139)[2].

The test consists of 50 multiple-choice fill-in-the-blank items in two sections. The first section–the reading/vocabulary section–contains 30 separate sentences with a blank in each one, with all

2 Various studies, including Montrul (2005) and Montrul and Slabakova (2003), have cited this task as "parts of" or "adapted from" the DELE, with no mention of the MLA test. Montrul has confirmed (Hoot, 2020), that the first/vocabulary portion came from the MLA test, whereas the second/cloze portion came from a sample DELE test available in the 1990s, as cited in Duffield and White (1999). The compiled test is freely available at https://nhlrc.ucla.edu/. Here, the task is listed as DELE Proficiency Test, Author: Dr. Silvina Montrul, Date: June 12, 2012.

items and choices targeting vocabulary knowledge. The second section–the cloze section–consists of a multi-paragraph reading passage with 20 blanks, with 10 items targeting vocabulary and 10 items targeting grammar knowledge (4 related to tense/aspect/mood, 4 related to prepositions, and 2 related to relatives and conjunctions). (See Section 3.2.1.1 for examples from each section and Note 2 for a link to the full test). Scoring usually consists of the total number of correct responses out of 50. Duffield and White (1999) proposed score ranges to categorize L2 Spanish proficiency levels as follows: 37–50 for "advanced", 25–36 for "intermediate", and 0–25 for "low". Montrul and Slabakova (2003) subsequently used the task with both L1 and L2 Spanish speakers and defined a score range of 45–50 for "near-native" proficiency (as 45 was the minimum score in the L1 group). It should also be noted that some researchers have at least one additional version of the task in circulation (Hoot, 2020). In that version, both sections differ from the original test but follow the same format.

To our knowledge, the first study to employ the task with HSs was Montrul (2005). Investigating the impact of early linguistic exposure on language development, she examined adult L2 and HSs of Spanish. Interestingly, Montrul (2005) noted that the "test might not be entirely suitable to predict the linguistic performance of heritage speakers or early bilinguals" (p. 237), given that this measure invites participants to make explicit grammatical and vocabulary judgments (which may not align with the implicit linguistic competencies inherent to such speakers; Carreira and Potowski, 2011). Nonetheless, this task (in particular, the one available on the UCLA National Heritage Language Resource Center (NHLRC) website) has been widely adopted as a proficiency assessment and as a means of cross-study comparison in L2 and HL bilingualism research (e.g., Faretta-Stutenberg and Morgan-Short, 2018; Sánchez Walker and Montrul, 2020; Solon et al., 2022; Torres, 2018; among many others). The task has often been passed down from researcher to researcher and is available publicly, as mentioned, facilitating the task's adoption in research.

Although few studies have attempted to validate this proficiency assessment, such studies have so far provided support for the test's internal reliability and external validity in L2 and HS samples. In particular, Montrul et al. (2008) and Montrul and Ionin (2012) explored these aspects among L2 learners and HSs. They found the internal reliability of the task to be moderate (Brown, 2013), as evidenced by a Cronbach's Alpha (i.e., α) coefficient of 0.827. This suggests that the test items were reliable and uniform for the two samples. Regarding validity, these studies examined correlations between scores on this task and performance on other measures of linguistic knowledge, including judgment accuracy for gender agreement and verb tense. The positive correlations observed ($r = 0.807$ for the HS group and $r = 0.653$ for the L2 group) provide some support for the construct validity of the task for these samples. Additionally, a recent study by Solon et al. (2022) with L2 speakers and HSs revealed significant correlations between the this task and other validated language proficiency assessments, such as the EIT, which is often utilized to assess Spanish oral proficiency (e.g., Faretta-Stutenberg et al., 2023; Kostromitina and Plonsky, 2022; Solon et al., 2022; see Bowden, 2016 for a validation study of the Spanish EIT).

Given this test's common use in the L2/HL fields, together with the limited evidence regarding its relationship with other proficiency assessments and experience factors, especially for HSs of Spanish, additional research examining this task is warranted. As such, this test was examined in the current study. However, given the fact that the test is (1) only partly from the DELE (and not a current version at that) and (2) that the task requires sentence and paragraph-level reading comprehension, with questions that target vocabulary (40 questions), along with some grammar knowledge (10 questions), we here refer to the task as a Spanish Vocabulary and Grammar Test (VGT).[3]

2.2 Subjective assessments of language proficiency

In the context of our study, we define subjective proficiency assessments as an approach for evaluating an individual's language proficiency that emphasizes subjective personal perceptions rather than objective metrics. These assessments rely on individuals' self-reports of their own individual language abilities, often elicited through surveys, interviews, or expert feedback (Olson, 2023a). Unlike objective assessments, subjective assessments explore personal views on language abilities, incorporating factors such as confidence, comfort level, and self-rated proficiency across language domains (the most predominantly used; Gertken et al., 2014). While subjective assessments might initially seem less precise compared to objective assessments, the use of Likert scales to quantify these subjective evaluations facilitates a systematic analysis of individuals' perceptions, thereby transforming subjective ratings into structured, quantifiable data. However, it is crucial to recognize that, despite our ability to quantify subjective assessments, the resulting data from subjective proficiency assessments can still be influenced by biases, socio-cultural and political factors, and variability in individual self-awareness (e.g., De Bruin, 2019; Hulstijn, 2012). This underscores the need for careful, contextualized interpretation of these assessments, considering the diverse factors that may impact individuals' subjective perceptions of their language proficiency.

2.2.1 Bilingual language profile questionnaire

The Bilingual Language Profile (BLP; Birdsong et al., 2012) was developed as a succinct and accessible self-report tool for assessing bilingual language dominance across bilingual languages

3 As Montrul and Slabakova (2003) state, "[t]he DELEs (Diplomas of Spanish as a Foreign Language) are the official accreditation of the degree of fluency in the Spanish language, issued and recognized by the Ministry of Education, Culture, and Sport of Spain" (p. 389). Thus, the use of the term "DELE" for this assessment has lended official weight to the test, while the DELEs themselves are currently quite different from this task. Further details on the DELE exams can be found at the Instituto Cervantes' website–the official body responsible for its administration, affiliated with Spain's Ministry of Education, Vocational Training and Sports (MEFPD): https://examenes.cervantes.es/es/dele/que-es. The authors would like to acknowledge members of the Hispanic and Lusophone Linguistics Facebook group and attendees at the 2022 UIC Bilingualism Forum for noting this inaccuracy in naming and sparking a deeper investigation into the origin and history of the task.

(see Treffers-Daller, 2019) and a general bilingual profile. It provides a continuous (and composite) dominance score, alongside a general profile of bilinguals' language history, use, attitudes, and proficiency.

Since its development, the BLP has been used across different areas of bilingualism research, including but not limited to research on language processing, language acquisition and psycho/neurolinguistics (e.g., Amengual and Chamorro, 2016; Kubota et al., 2023; Poarch et al., 2019). The availability of the BLP in multiple languages, coupled with its ease of use and open access, has likely contributed to its broad adoption, making it a widely used measure of bilingual language use, experience, and proficiency, including for HSs (Solís-Barroso and Stefanich, 2019). As of February 21, 2024, the BLP had been cited 197 times, according to Google Scholar, highlighting its widespread recognition and impact in the academic community. Several studies have provided positive evidence supporting its construct validity as well as its concurrent validity and test-retest reliability (see Dass et al., 2024; Gertken et al., 2014; Mallonee Gertken, 2013; Olson, 2023b; Solís-Barroso and Stefanich, 2019).

However, the BLP's reliance on self-reported data introduces the potential for subjective bias(es). This can lead participants to either overestimate or underestimate their language proficiency. Such underestimation is a notable concern among HSs, as highlighted by Bayram et al. (2021b). This underscores the importance of interpreting BLP results with caution and, where feasible, integrating objective measures to support (and enhance) the available self-reported data.

2.2.2 Language entropy

In their 2020 study, Gullifer and Titone proposed a novel approach to examining bilingualism through the lens of language entropy. Language entropy is defined as a metric for estimating the diversity of language use in social contexts, particularly focusing on the various contexts in which bilinguals engage with their languages. According to Gullifer and Titone, language entropy can serve as a relevant tool for understanding and quantifying individual differences in how bilinguals navigate their different linguistic environments, by exploring the extent to which bilingual individuals engage in environments that require the use of both languages simultaneously (i.e., dual language contexts) vs. those that are more segregated, relying on a single language mode (i.e., compartmentalized language contexts). This construct of language entropy as defined by Gullifer and Titone is particularly relevant when considered alongside theoretical and empirical findings, such as the adaptive control hypothesis (ACH; Abutalebi and Green, 2016; Green and Abutalebi, 2013). The ACH suggests that how bilinguals use their languages across different social settings—their interactional context—plays a significant role in shaping how they represent, access, and control these languages. According to this hypothesis, bilinguals who frequently navigate dual language contexts (integrated bilinguals) face distinct language and executive control demands compared to those who engage with their languages in more compartmentalized, single-language settings. By quantifying the social diversity of language use through language entropy, researchers can gain quantifiable estimations of how bilinguals manage and navigate multiple linguistic systems and the factors that influence language choice, language switching, and language adaptation in bilingual contexts.

The construct of language entropy is starting to gather significant attention in studying bilingualism thanks to its relationships with the neural, cognitive, and social dynamics of bilingual language proficiency and use. For instance, Sulpizio et al. (2020) investigated the impact of bilingual experience—considering factors such as age of acquisition, proficiency, and language entropy—on the functional connectivity within and between language and executive control networks in the brain. They found that higher language entropy, indicating more diverse and integrated language use, was associated with enhanced connectivity in these networks. Building on this, Li et al. (2021) explored the relationship between bilingual language entropy and executive function. Their findings revealed that greater diversity of language use across social contexts, as assessed by language entropy, seemed to be associated with enhanced brain network specialization and segregation in brain networks associated with executive control. Additionally, Kałamała et al. (2023) introduced a novel psychometric network modeling approach to capture the complexity of bilingual experience, focusing on language entropy and language mixing as key indicators. Their study suggests that bilingualism is an emergent phenomenon shaped by the interplay of language acquisition background, skills, and usage practices. Finally, Wagner et al. (2023) critically examined how contextual factors influence the effects of language entropy on cognitive performance, comparing bilingual contexts in Toronto and Montréal, with results suggesting that language entropy can vary significantly based on environmental/societal factors that influence language use. Collectively, these studies underscore the critical role of language entropy in understanding the complex interplay between the social diversity of language use, neural and psycho/sociocognitive function, and sociolinguistic contexts.

Language entropy is also intimately connected to code-switching, a bilingual practice often associated with HS populations. An individual speaker's tendency to seamlessly alternate between languages goes hand-in-hand with higher language entropy. Although code-switching is frequently maligned as a sign of "disfluency" among "non-proficient" bilinguals, in reality, more dense switching (i.e., intra-sentential code-switching) is associated with higher proficiency in both languages (Bullock and Toribio, 2009). Code-switching is a form of linguistic flexibility that can be seen as a sign of the vitality of the minority language in a community (Gardner-Chloros, 2009) in that it supports identity formation while also being tied to language proficiency. As such, examining language entropy is crucial for understanding the linguistic behavior of HSs of Spanish in the US, as it provides insights into how they integrate Spanish and English in their daily lives, which in turn influences the intergenerational transmission of the HL as well as the diverse ways in which they develop their proficiency in it. In the current study, we calculated language entropy using questions from the BLP (see details in the Methods Section).

2.3 Study goals and research questions

Consequently, the goals of our study were three-fold:

1. Reliability: First, we evaluated the reliability of the objective proficiency assessments (LexTale-ESP, VGT), by examining their internal consistency. As mentioned above, internal consistency reliability refers to the degree to which different items within a specific assessment yield consistent results (Cronbach, 1951). By analyzing the internal consistency of these items, we aimed to provide evidence that these assessments offer a stable and reliable measure of Spanish proficiency.
2. Validity: Second, we evaluated the validity of these assessments by examining their interrelationships with each other and their relationships with the subjective proficiency assessments. This aspect of the study focused specifically on construct validity, which refers to the extent to which these assessments accurately measure the construct of language in Spanish as an HL among HSs of Spanish in the US (Messick, 1995). By analyzing these relationships, we aimed to provide evidence as to whether the assessments reflect the Spanish proficiency constructs they are intended to evaluate.
3. Validity in Context: Finally, we investigated how objective and subjective proficiency assessments were related to different HL experience factors, specifically years of exposure to Spanish, years of Spanish schooling, and language entropy (following Gullifer and Titone, 2020). This aspect of the study addresses both construct validity and ecological validity. Construct validity, in this context, pertains to whether these assessments accurately capture different dimensions of Spanish proficiency, while ecological validity concerns how well performance on these tasks reflects real-world language use and experience (Bronfenbrenner, 1977). By exploring these relationships, we aimed to understand how these proficiency assessments relate to real HL experiences and usage patterns in everyday life.

These goals guided the formulation of the following research questions (RQs):

RQ1 (Reliability): What is the internal consistency of the selected objective proficiency assessments for this sample of HSs of Spanish?

With regard to RQ1, we hypothesized that the selected objective proficiency assessments will demonstrate high internal consistency within this sample of HSs of Spanish. As noted above, internal consistency is crucial for establishing the reliability of these proficiency assessments (Cronbach, 1951). Our hypothesis was based on previous research indicating high internal consistency reliability for these measures across diverse populations (e.g., Izura et al., 2014; Montrul et al., 2008). Thus, to our knowledge, this has not yet been investigated specifically in our population of HSs. Therefore, we aimed to examine whether similar reliability would be observed in assessing Spanish proficiency for this sample of HSs of Spanish.

RQ2 (Validity): How do the selected objective and subjective proficiency assessments relate to one another for this sample of HSs of Spanish?

With regard to RQ2, we hypothesized finding variable relationships among the assessments, reflecting aspects of construct validity. As noted above, construct validity examines whether these assessments accurately measure the intended constructs of Spanish proficiency and how these constructs interrelate (Messick, 1995). Prior research has shown that objective and subjective proficiency assessments can be related, but the strength and nature of these relationships may vary (e.g., Gullifer and Titone, 2020; Tomoschuk et al., 2019). Specifically, we predicted that the objective assessments would be significantly correlated, reflecting their shared focus on vocabulary knowledge and prior evidence of their intercorrelations (e.g., Bermúdez-Margaretto and Brysbaert, 2022). However, as these intercorrelations have not been investigated specifically in HSs, we aimed to examine whether similar patterns would emerge for this sample of HSs of Spanish. Additionally, we drew on evidence supporting the external validity of the VGT as a proficiency assessment (e.g., Montrul and Ionin, 2012). However, we also predicted that the relationships between the objective and subjective proficiency assessments would be more variable than the relationships among the objective assessments themselves, due to (a) subjective assessments' potential to be influenced by individual biases or differing subjective interpretations (Tomoschuk et al., 2019) and (b) HSs' frequently reported underestimation of their HL abilities and overall HL proficiency compared to their objectively measured HL proficiency (e.g., Bayram et al., 2021b). Consequently, while both types of proficiency assessments may demonstrate construct validity, we hypothesized that the objective measures would provide a more consistent and reliable reflection of Spanish proficiency, with stronger and more consistent relationships observed among the objective assessments as compared to those between the objective and subjective assessments.

RQ3 (Validity in Context): Do the selected objective and subjective proficiency assessments correlate similarly with each HL experience factor—namely years of exposure to Spanish, years of Spanish schooling, and social diversity of HL use (i.e., language entropy) helping to determine if these HL experience factors are equally influential for capturing dimensions of Spanish proficiency and reflecting real-world HL use among Spanish HSs?

With regard to RQ3, we hypothesized that the selected objective and subjective proficiency assessments would be differentially related to the investigated HL experience factors, reflecting both construct and ecological validity. Previous research suggests that language experience factors, such as the amount and social diversity of L2 use and exposure (e.g., language entropy), are critical in shaping proficiency outcomes, with subjective assessments potentially being more sensitive to context-dependent language experiences. However, as demonstrated by the findings of Gehebe et al. (2023) and Gullifer et al. (2021), these effects may vary depending on the context (i.e., where, when, and how bilinguals' languages are used) and type of proficiency measure used (i.e., objective vs subjective proficiency assessments focused on specific language abilities). Although these studies do not directly compare objective and subjective measures, they suggest that different facets of language experience might influence each assessment type uniquely. Therefore, we expected that objective and subjective assessments of Spanish proficiency would not pattern uniformly but rather would tap into distinct components of the HL experience factors, with subjective assessments potentially capturing more context-sensitive, real-world language usage.

3 Methods

3.1 Participants

A total of 45 Spanish-English HSs, with ages ranging from 18 to 38 years ($M = 23.46$; $SD = 5.13$), were recruited through multiple avenues, including undergraduate Spanish courses as well as personal contacts, leading to a somewhat diverse set of profiles not only in terms of prior language experience and proficiency but also in their current exposure to and use of Spanish. Participants were classified as heritage speakers of Spanish if their age of onset for Spanish exposure was before age 6, based on research indicating that early exposure to the heritage language, typically before school age, is crucial for its maintenance and development. This cutoff aligns with findings from Benmamoun et al. (2013), who suggest age 6 as a reasonable cutoff for early bilinguals, and Silva-Corvalán (2014), who emphasizes that exposure by age 5 allows for substantial meaningful exposure and interaction with the HL before formal schooling in the majority language. The average age of onset for exposure to Spanish among participants was 0.33 years ($SD = 0.98$). Participants' use of Spanish and English (discussed below) and their language dominance were assessed using the BLP (Birdsong et al., 2012). For language dominance, the group averaged 32.1 out of a possible range of ±218 ($SD = 30.8$; range −44 to 94), where greater positive values indicate more English dominance, and greater negative values indicate more Spanish dominance. Overall, the group leaned toward English dominance, but scores varied widely, with some Spanish-dominant participants among the group. See Table 1 for a summary of participants' language proficiency (as measured by objective and subjective assessments), experience, and use of both Spanish and English. This information related to the HL will be addressed again in more detail in the Results Section.

Participants reported their gender identities and cultural/ethnic backgrounds via free-response questions. Most participants identified as *female* ($N = 35$; 77.8%); eight participants identified as *male* (17.8%), one participant identified as *trans masculine* (2.2%), and one participant chose not to disclose gender identity (2.2%). For cultural/ethnic background, participants could include as many different identifiers as they wished. The most common response included the term *Mexican* ($N = 18$; 40.0%). Other common identifiers included *Hispanic* ($N = 10$; 22.2%) and *Latinx/o/a* ($N = 9$; 20.0%). Less commonly reported identities (1–2 participants each) were: *Bolivian, Colombian, New Mexican, Peruvian, Puertorriqueña/German, Salvadoran, Texan/Tejana.* Additionally, one individual responded *multi-racial*, *one* responded *unique*, and three participants chose not to answer this question.

3.2 Materials

Given the methodological scope of this study, as mentioned, our assessments fall into three broad categories:

1. Objective assessments of Spanish proficiency: These assessments consist of widely used tasks that quantify vocabulary and/or grammar knowledge and have been used as proxies to assess Spanish proficiency.
2. Subjective self-assessments of Spanish proficiency: These assessments offer a subjective and personal perspective on one's Spanish abilities across various language domains.
3. HL experience factors: These self-reported data (i.e., years of Spanish exposure, years of Spanish schooling, and language usage data, used to calculate language entropy) tap into various aspects of the depth and nature of each individual's engagement with Spanish.

3.2.1 Objective assessments of Spanish proficiency

To critically examine how Spanish proficiency is often measured, two objective, quantitative assessments (rather than one) were employed. These two tasks were chosen for a larger project (Koronkiewicz, 2023) due to their widespread usage in the fields of L2 acquisition and bilingualism research as proxies for characterizing Spanish language proficiency.

3.2.1.1 Spanish vocabulary and grammar task

The first objective measure used was the 50-item written, multiple-choice Spanish VGT, consisting of two sections. As described above, the first section comprises 30 sentence-level items, for which selecting the correct answer depends on understanding the sentence and completing it with a semantically appropriate word or phrase, for example (see footnote for translations[4]):

Al oír del accidente de su buen amigo, Paco se puso_____.
a. alegre b. fatigado c. hambriento d. desconsolado

The second section (20 items) is a multi-paragraph, fill-in-the-blank passage. Multiple-choice options for each blank are oriented toward vocabulary/semantics (10 questions) and prescriptively correct grammar (10 questions). Note that this section is representative of Peninsular Spanish language and culture, as shown through its focus on a Catalan artist and the use of some particular verbal morphology. For example (see footnote for translation):

Hoy se inaugura en Palma de Mallorca la Fundación [Pilar] y Joan Miró, en el mismo lugar en donde el artista vivió sus últimos treinta y cinco años. El sueño de Joan Miró se ha_____ (1). Los fondos donados a la ciudad por el pintor y su esposa en 1981 permitieron que el sueño se_____ (2)...

1. *a. cumplido b. completado c. terminado*
2. *a. inició b. iniciara c. iniciaba*

4 Upon hearing about his old friend's accident. Paco became_____
a. happy b. fatigued c. hungry d. inconsolable
The [Pilar] y Joan Miró Foundation opens today in Palma de Mallorca, in the same place that the artist lived his last thirty-five years. Juan Miró's dream has been_____ (1). The funds donated to the city by the painter and his wife in 1981 allowed the dream to be_____ (2).

1. a. fulfilled b. completed c. finished
2. a. started.PRET.IND b. started.IMPF.SUBJ c. started.IMPF.IND

TABLE 1 Overview of participant Spanish and English language proficiency and experience.

	Measure	Possible score	M	SD	Min	Max
Spanish language proficiency	Vocabulary and grammar test	0–50	35.8	8.0	11	48
	Lextale-Esp	−60–60	19.0	15.9	−2	57
	Self-rated speaking	0–6	4.7	1.1	2	6
	Self-rated understanding	0–6	5.2	1.0	3	6
	Self-rated reading	0–6	4.5	1.2	1	6
	Self-rated writing	0–6	4.2	1.5	0	6
English language proficiency	Vocabulary and grammar test	0–40	34.9	5.5	8	39
	Lexical decision task	0–100	86.9	9.5	59	100
	Self-rated speaking	0–6	5.8	0.4	5	6
	Self-rated understanding	0–6	5.9	0.3	5	6
	Self-rated reading	0–6	5.9	0.3	5	6
	Self-rated writing	0–6	5.7	0.4	5	6
Experience with Spanish	Years of exposure		23.1	5.2	17	38
	Years of schooling		6.7	5.2	0	20
Experience with English	Years of exposure		20.8	5.7	14	38
	Years of schooling		14.7	4.0	0	24
Current use of Spanish	% with friends		27.9	19.5	0	80
	% with family		60.2	23.6	0	100
	% at school/work		26.1	24.4	0	89
Current use of English	% with friends		71.8	19.6	20	100
	% with family		39.8	23.6	0	100
	% at school/work		73.5	24.9	11	100

As detailed in the Background Section, for Spanish, the Vocabulary and Grammar Test was the "Modified DELE" or VGT (with the average above coming from participants' total score), and the Lexical Decision Task was the Lextale-Esp (with the average above reflecting penalty scoring; see Methods). For English, the Vocabulary and Grammar Test was adopted from O'Neill et al. (1981), while the Lexical Decision task was the LexTALE (Lemhöfer and Broersma, 2012). All other measures were taken from portions of the BLP. Also, with regard to current language use, while only Spanish and English are reported in the table, one participant listed 10% Italian use in an average week at school/work, and one participant listed 10% Chinese with friends and at school/work, and 10% Chinese/Korean when talking to themself and counting; no other participants listed additional languages.

In the present analyses, we include three different score calculations from the VGT. The first is simply the total number of correct responses from 0 to 50, which is the most commonly reported score in previous research. However, given the qualitative differences between the two sections, we also separated the calculations for each section (i.e., just the sentence-level vocabulary-oriented responses [from 0 to 30], and just the paragraph-level vocabulary- and grammar-oriented responses [from 0 to 20]).

3.2.1.2 Lextale-Esp

The second objective assessment of Spanish proficiency was the Lextale-Esp lexical decision task (Izura et al., 2014). As mentioned above, the task includes 90 items that are either Spanish words (*pellizcar* 'to pinch'; $n = 60$) or Spanish-like non-words (e.g., *terzo*; $n = 30$), and the participant is asked to simply select *Sí* or *No* for each item to indicate if it is a word or not.

For the analysis, we include three different calculations. The first is simply the total of correct answers from 0 to 90. The second is the calculation recommended by the authors of the Lextale-Esp where there is a penalty for "guessing behavior". This is calculated as the total of correct words minus two times the total of incorrect non-words (e.g., if a participant responded *Sí* to *terzo*), with a range of possible scores from −60 to 60. Finally, we also included a d-prime (d') score, which is a standardized measure following signal detection theory that accounts for response bias in a participant's ability to discriminate words from non-words (Macmillan and Creelman, 1996); specifically, scores can range from −4.65 to 4.65, where a score of zero reflects chance-level discrimination ability.

3.2.2 Subjective assessments of Spanish proficiency

The subjective assessments of Spanish proficiency were self-reported language skill ratings, which were collected as part of the BLP. The questionnaire asks participants how well they speak, understand, read, and write Spanish (and, separately, English), with a 7-point Likert scale from *not well* (0) to *very well* (6) for each question. For the analysis, we include three different calculations. The first is a composite score ("Total") that averages the four questions about Spanish (i.e., the four skills). The second and third scores average together productive

(i.e., speaking, writing) and receptive (i.e., understanding, reading) Spanish abilities separately. We wanted to examine productive and receptive Spanish abilities separately given the receptive nature of the objective proficiency assessments used here, which might be expected to pattern together.

3.2.2.1 Heritage language experience factors

Three HL experience factors were also derived from self-reported data gathered from participants' BLP responses. First, we calculated years of Spanish exposure based on participant responses to the following question: *At what age did you start learning the following languages?*, subtracting their reported age at first exposure to Spanish from the participants' current age. For years of Spanish schooling, we utilized the participant responses to the question, *How many years of classes (grammar, history, math, etc.) have you had in the following languages (primary school through university)?*

Additionally, we examined language entropy as an HL experience-related factor, following Gullifer and Titone (2020)'s methodology. As mentioned above, language entropy assesses the dynamics of an individual's language use across different sociolinguistic contexts, indicating the degree to which their languages are used in a compartmentalized or integrated manner. We used data from the BLP (Birdsong et al., 2012) to calculate language entropy, where participants reported the percentage of time in an average week they use each language in five different contexts: at school/work, with friends, with family, when talking to themselves, and when counting. These percentages were converted into a proportion for each context, which we then used with the *languageEntropy* package in *R* (Gullifer and Titone, 2018) to calculate an entropy score for each context.

The *languageEntropy* package calculates entropy based on Shannon entropy (Shannon, 1948), a concept from information theory originally developed to estimate the unpredictability or diversity in a system of possible outcomes. In information theory, entropy provides a measure of how "spread out" or "integrated" different elements are within a system. In this context, Gullifer and Titone (2020) adapted Shannon entropy to estimate the diversity in bi/multilingual language use, where the entropy score for each context reflects the proportion and dynamics of language use across social settings. Specifically, the *languageEntropy* package uses the formula $H = -\sum (p_i \cdot \log(p_i))$, where H represents the entropy score for a given context, p_i denotes the proportion of time that each language i is used within that context, and the summation is taken over all languages used in that context. Thus, the formula works as follows: for each language used within a context (e.g., English, Spanish), we calculate the proportion of time the participant uses that language and multiply it by the logarithm of that proportion. This product $p_i \cdot \log(p_i)$ is calculated for each language, and the results are then summed. The negative sign in front of the summation ensures the entropy value is positive. Thus, the resulting entropy score reflects the diversity of language use within each context.

A score of 0 indicates complete compartmentalization, where only one language is used exclusively within a context (e.g., 100% use of English and 0% use of Spanish in a particular setting). For instance, a participant who reports using only English in some contexts (e.g., school/work and with friends) and only Spanish in other contexts (e.g., family and self-talk) would have an entropy score close to 0, reflecting complete compartmentalization across different social contexts. In contrast, a score of 1 indicates complete integration, where both languages are used equally within a context (e.g., 50% English and 50% Spanish in a given context). A participant who reports using each language 50% of the time across all contexts—such as school/work, with friends, and with family—would achieve an entropy score close to 1, showing full integration, as both languages are used equally within each context.

To generate an overall measure of language integration across contexts, we computed a composite entropy score by averaging the individual entropy scores across all contexts. This composite score provides a single, interpretable metric that represents the participant's overall level of language integration or compartmentalization in daily life. It is important to note that for multilingual experiences involving more than two languages, the entropy calculation dynamically adapts by incorporating each language's proportion of use in the formula. For instance, if a participant uses three languages (e.g., English, Spanish, and French) in a context, each language's proportion of use is included in the calculation. The maximum entropy score increases as more languages are used, reflecting a greater diversity of language use. For a trilingual context, the maximum entropy score becomes approximately 1.585 (the logarithm of 3), rather than 1, allowing the measure to capture multilingual dynamics effectively. For a more detailed description of entropy calculations and their theoretical basis in the context of bi/multilingualism research (see Gullifer and Titone, 2020).[5]

3.3 Procedure

As mentioned, the data under analysis here came from a larger project. This larger project included three study sessions that were completed on different days. All relevant data for the present study come from the first two sessions, completed independently by study participants via Qualtrics surveys (Qualtrics, Provo, UT).

In the first session (~10–20 min), participants completed the Lextale-Esp (Izura et al., 2014), followed by the English LexTALE (Lemhöfer and Broersma, 2012). They then answered 12 questions targeting language exposure and acquisition and language mixing experience and attitudes. Those 12 questions were used to categorize participants as late L2 learners or HSs of Spanish, as both were targeted in recruitment efforts for the larger project. Any participant who indicated that they learned both languages from a young age and had a parent or primary caregiver who primarily used Spanish with them growing up was categorized as a HS and was included in the current dataset.

The second session (~45–60 min) included a series of acceptability judgment tasks for code-switched sentences,

5 As an anonymous reviewer noted, a composite entropy score does not take into account the varying amounts of time participants spend using their languages in the different contexts that are measured (i.e., an individual may spend considerably more time talking with their family than with friends, or vice-versa). Nonetheless, we believe such composite scores remain helpful for broadly characterizing the diversity of language use in a single metric, as proposed by Gullifer and Titone (2020).

Spanish-only sentences, and English-only sentences (none of which are analyzed here; see Koronkiewicz, 2023, for details). Between the judgment tasks, participants completed the (Spanish) VGT and an English proficiency measure (O'Neill et al., 1981) parallel to the VGT. Finally, participants completed basic demographic questions and the BLP (Birdsong et al., 2012), both in English.

Participants were compensated with a $20 Amazon.com eGift Card for their time completing the first two sessions of the study. Informed consent was also obtained from all participants before each study session.

4 Results

4.1 Descriptive results

First, we present a general overview of participants' data from the different assessments of Spanish proficiency, both objective and subjective. These descriptive statistics, detailed in Table 2, summarize the average scores obtained for the three proficiency assessments. Recall that for each assessment, there were three different score types; the various scores either comprised the total score and subsets of the measure (as in the case of the VGT and self-ratings) or they employed distinct score calculations (as in the case of the Lextale-Esp) (see Section 3.1 and Table 1 for a comprehensive reporting of participant characteristics, including demographics and educational background).

Because many of the score types have different ranges of possible values, direct comparison of mean values across assessments is not descriptively straightforward. Thus, Figure 1 illustrates average performance on each proficiency assessment as a percentage of the maximum for each score type, facilitating a more meaningful comparison. As we can see, overall, the scores for the different assessments were relatively similar, as the average scores were between ~60–80% of their respective maximum scores. Descriptively, participants performed the lowest on Score 3 of the VGT (i.e., the paragraph-level section that requires reading comprehension generally as well as specific vocabulary and grammar knowledge), receiving on average 11.9 out of 20 points (59.5%). Meanwhile, participants performed the highest on Score 3 of the self-ratings (i.e., receptive skills), averaging a 4.9 out of 6 (81.7%). The VGT showed the most variability within the proficiency assessments, where there was a difference of 19.8% between Score 2 and Score 3 (i.e., the sentence-level vocabulary-focused questions and the paragraph-level vocabulary and grammar-focused questions, respectively).

Data analyses were conducted using R version 4.2.2 (2022-10-31), within RStudio Version 2022.12.0+353. A suite of R packages was employed for comprehensive data manipulation, analysis, and visualization, including *tidyverse* for manipulation of data sets, *ggplot2* for creating graphics, *psych* for calculating effect sizes and measures of internal consistency, *stats* for correcting *p* values and *languageEntropy* for calculating and analyzing language entropy scores to address our specific data analysis needs. In order to assess the strength and direction of the relationships we examined in RQ2 and RQ3, we used the non-parametric Spearman's correlation as the data were not normally distributed, and some were ordinal rather than continuous. Note that all correlation sizes were interpreted following Plonsky and Oswald (2014), and *p* values associated with correlations were corrected using the Benjamini-Hochberg (BH) correction using the *p.adjust()* function in the *stats* package in R to account for the false discovery rate in multiple statistical tests (Benjamini and Hochberg, 1995). The BH correction method was chosen due to the characteristics of our study, as the correction does not assume independence of tests, and our study was exploratory in nature, providing a balanced approach to controlling for both Type I and Type II errors (i.e., false positives and false negatives, respectively).

4.2 RQ1 (reliability): what is the internal consistency of the selected objective proficiency assessments for this sample of HSs of Spanish?

To address RQ1, pertaining to whether the objective proficiency assessments employed in this study were reliable (i.e., internally consistent) for our sample, we calculated Cronbach's alpha for the Lextale-Esp and each portion of the VGT using the *alpha* function from the *psych* package in R. The 95% confidence intervals (CIs) for Cronbach's alpha were calculated using the Duhachek method (Duhachek and Iacobucci, 2004). To interpret the levels of Cronbach's alpha, we followed the guidelines provided by Brown (2013).

The overall reliability results were as follows: For the Lextale-Esp, Cronbach's alpha indicated high internal consistency at 0.88, with 95% CIs [0.82 −0.93]. The sentence-level portion of the VGT showed a Cronbach's alpha of 0.87, with 95% CIs [0.82 −0.92], indicating high internal consistency. Additionally, the paragraph-level portion of the VGT revealed a Cronbach's alpha of 0.68, with 95% CIs [0.55 −0.82], indicating moderate internal consistency.

In summary, the objective proficiency assessments, including the Lextale-Esp and the sentence-level portion of the VGT, demonstrated moderate to high internal consistency, revealing that the items were highly correlated and reliably measured the same construct. However, the paragraph-level portion of the VGT showed lower internal consistency, revealing that the items were reasonably consistent in measuring the same construct for this sample, but less so than the other portions of the objective assessments under investigation.

4.3 RQ2 (validity): how do the selected objective and subjective proficiency assessments relate to one another for this sample of HSs of Spanish?

A summary of the correlations used to assess the relationships between objective and subjective proficiency assessments is provided in Table 3. We also provide scatter plots to illustrate these relationships in Figures 2–4, visualized as a percentage of the maximum for each score type.

First, we can see that the VGT was positively correlated with the Lextale-Esp regardless of the specific score type, as all nine statistics were significant, medium-to-large correlations,

TABLE 2 Descriptive statistics for objective and subjective language proficiency assessments, broken down by score types.

Measure	Score type	Possible score	*M*	*SD*	Min	Max
VGT	1 - Total	0–50	34.8	7.8	10.0	47.0
	2 - Sentence-level	0–30	23.8	4.9	6.0	29.0
	3 - Paragraph-level	0–20	11.9	3.7	5.0	19.0
LxE	1 - Standard	0–90	62.3	10.5	37.0	88.0
	2 - Penalty	−60–60	19.0	15.9	−2.0	57.0
	3 - d-prime	−4.65 to −4.65	0.98	0.92	−0.10	3.96
Self-Rating	1 - Total	0–6	4.7	1.0	2.8	6.0
	2 - Productive	0–6	4.5	1.1	2.0	6.0
	3 - Receptive	0–6	4.9	0.9	3.0	6.0

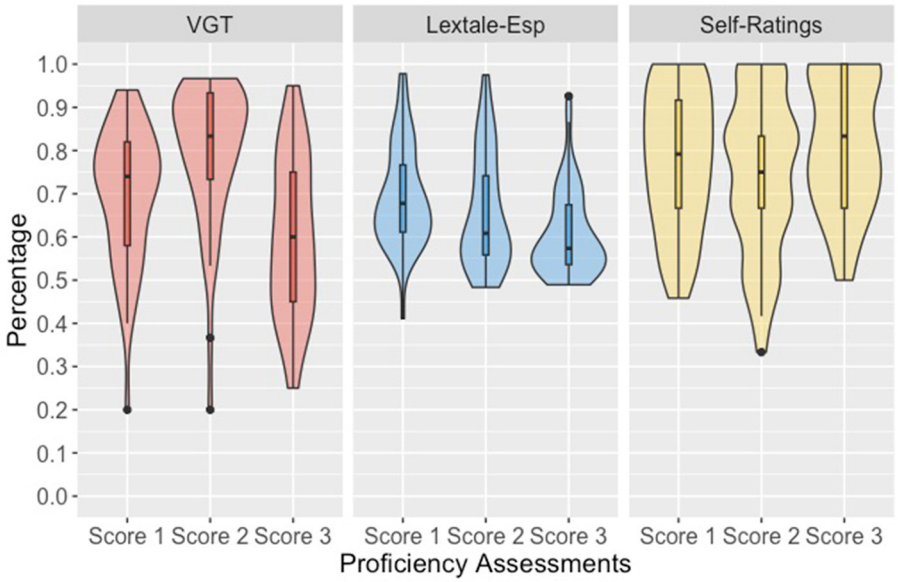

FIGURE 1
Overall mean performance on objective and subjective proficiency assessments as a percentage. This figure presents violin plots depicting the distribution of percentage scores across (y-axis) the different proficiency assessments (x-axis). Specifically, VGT scores are represented in red, LexTALE-Esp in blue, and Self-Ratings in yellow. Recall that each assessment method is divided into three score types: VGT: Score 1 (overall score), Score 2 (sentence-level vocabulary score), Score 3 (paragraph-level vocabulary and grammar score); LexTALE-Esp: Score 1 (standard scoring), Score 2 (penalty-based scoring), Score 3 (*d*-prime scoring); Self-Ratings: Score 1 (overall score), Score 2 (productive skills score), Score 3 (receptive skills score). The width of each violin plot corresponds to the frequency of data points at different percentage levels, and the black lines inside the violins indicate the interquartile range and median of the scores.

ranging from $r_{s(43)} = 0.51$ to.74, $p < 0.01$ (see Table 3). As for the subjective proficiency assessments (i.e., self-ratings), we found no significant correlations with the VGT. Correlations between the Lextale-Esp and subjective assessments were mixed: while Lextale-Esp 1 (i.e., standard scoring) had small, significant correlations with each of the self-rating scores, ranging from $r_{s(43)} = 0.35$ to 0.37, $p < 0.05$. There were no significant correlations with Lextale-Esp 2 (i.e., penalty scoring). or Lextale-Esp 3 (i.e., *d-prime* scoring).

Given the consistent pattern of the VGT and Lextale-Esp correlation results, and in the interest of economy, we decided to condense the number of variables for subsequent analyses. For the objective assessments, VGT 1 will be included because it represents the total score, combining both sentence-level and paragraph-level portions of the VGT. Although we initially considered Lextale-Esp 3 (i.e., *d*-prime scoring) as a potentially more reliable score type for binary-choice tasks (e.g., Rhodes et al., 2019), our current results do not support significant correlations between the LexTale-Esp 3 and the subjective assessments. Instead, LexTale-Esp 1 (i.e., standard scoring) will be used in subsequent analyses because it showed small, significant correlations with each of the self-rating scores, indicating its potential effectiveness in capturing Spanish proficiency as perceived by our study participants. Considering the variation in the pattern of correlations with the subjective assessments, we will retain the division of such skills by reporting both Self-Rating 2 (i.e., productive skills) and

TABLE 3 Spearman correlations between objective and subjective language proficiency assessments.

	VGT 1	VGT 2	VGT 3	LxE 1	LxE 2	LxE 3	Self 1	Self 2	Self 3
VGT 1	–	–	–	**0.59****	**0.58****	**0.60****	0.28	0.27	0.30
VGT 2	–	–	**0.74****	**0.55****	**0.51***	**0.54****	**0.**30	**0.**30	0.30
VGT 3	–	–	–	**0.59****	**0.60****	**0.61****	0.25	0.25	0.26
LxE 1	–	–	–	–	–	–	**0.37***	**0.36***	**0.35***
LxE 2	–	–	–	–	–	–	0.25	0.27	0.21
LxE 3	–	–	–	–	–	–	0.29	0.31	0.25
Self-rating 1	–	–	–	–	–	–	–	–	–
Self-rating 2	–	–	–	–	–	–	–	–	**0.85****
Self-rating 3	–	–	–	–	–	–	–	–	–

VGT 1 = overall score; 2 = sentence-level portion score; 3 = paragraph-level portion score. LxE (LexTale-Esp): 1 = standard score; 2 = penalty-based score; 3 = d-prime score. Self-rating: 1 = composite score; 2 = productive skills score; 3 = receptive skills score. Correlation coefficients and p values are rounded to two decimal places. Bold values indicate statistically significant correlations, with significance levels as follows: $^{*}p < 0.05$ and $^{**}p < 0.01$.

Self-Rating 3 (i.e., receptive skills). This approach ensures that we maintain a comprehensive understanding of the relationships between objective assessments and the different dimensions of subjective self-ratings.

4.4 RQ3 (validity in context): do the selected objective and subjective proficiency assessments correlate similarly with each HL experience factors—, namely years of exposure to Spanish, years of spanish schooling, and social diversity of HL use (i.e., language entropy) helping to determine if these HL experience factors are equally influential for capturing dimensions of spanish proficiency and reflecting real-world HL use among spanish HSs?

Recall that for RQ3, we aimed to examine the validity, specifically construct and ecological validity, of the objective and subjective assessments under investigation to provide evidence as to the extent to which these proficiency assessments reflect Spanish HL proficiency in US-HSs of Spanish and how they relate to relevant HL experience factors—namely, years of exposure to Spanish, years of Spanish schooling, and language entropy—providing insights into practical applicability and relevance in real-world language use scenarios.

We begin by providing descriptive statistics for the selected HL experience factors. First, participants reported an average of 23.1 years of exposure to Spanish and 6.7 years of schooling in Spanish (for more detailed descriptive data, see Participants section, Table 1). Second, language entropy scores (in the different contexts and an overall language entropy composite score) are displayed in violin plots in Figure 5.

Note that to calculate and analyze language entropy scores, we utilized the *languageEntropy* package in R. This specialized package provided the necessary tools to quantitatively assess the degree of language integration and compartmentalization among participants based on their reported language use in various contexts (following Gullifer and Titone, 2018). Also, recall that a score of 0 indicates complete compartmentalization of the two languages, whereas a score of 1 indicates full integration; for participants who report speaking more than two languages[6], the language entropy score can exceed 1, and in such a case a score of 1.585 would indicate complete integration across all languages spoken (see Gullifer et al., 2021 for more details). In general, participants' scores represented integration more so than compartmentalization, as seen in Figure 5.

A summary of the results from correlational analyses for RQ3 is provided in Table 4, where, again, Spearman correlation tests were conducted to assess the strength and direction of the relationships between the language proficiency assessments and the HL experience factors.

Our findings present a multifaceted pattern of small [$r_{s(43)} = 0.29$ to 0.36] to medium [$r_{s(43)} = 0.41$ to 0.44] correlations between the HL experience factors and the objective and subjective/self-rated Spanish proficiency assessments. First, years of exposure to Spanish was significantly correlated with both objective assessments, VGT 1 [$r_{s(43)} = 0.47$, $p < 0.01$] and Lextale-Esp 1 [$r_{s(43)} = 0.42$, $p < 0.05$], but not with self-ratings. Specifically, the more years that participants reported being exposed to the HL, the better their performance on the VGT and the Lextale-Esp. Years of Spanish schooling, on the other hand, were not significantly correlated with any measure.

Turning to language entropy, the analyses revealed more consistent significant correlations with the subjective self-ratings

6 Only two participants indicated speaking a third language in three different contexts, reporting 10% of the time with friends, at school, and when counting. As noted in the manuscript, the entropy calculation adapts to accommodate this additional language proportion, allowing for a slightly higher maximum entropy score (~1.585 for three languages). Both participants show high integration between Spanish and English, with minor use of a third language reflecting a more continuous and diverse multilingual experience.

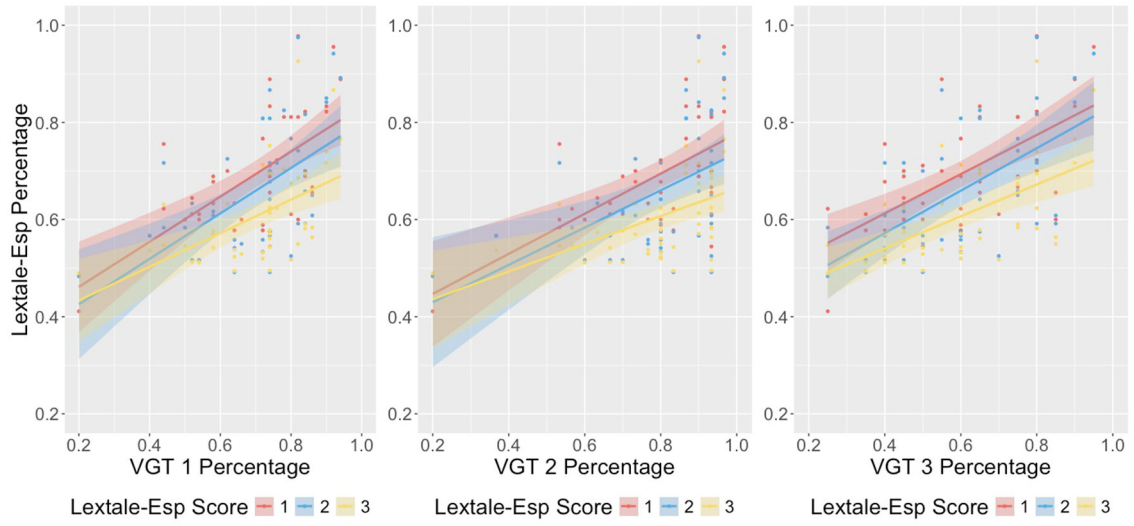

FIGURE 2
Scatter plots illustrating the relationship between VGT scores and LexTale-Esp scores. The scatter plots display the percentage scores on the VGT (x-axis) against the percentage scores on the Lextale-Esp (y-axis) for each score type. Recall that VGT 1 refers to the overall score, VGT 2 includes only the sentence-level score, and VGT 3 includes only the paragraph-level vocabulary and grammar score. The color-coded lines and points represent the different Lextale-Esp scoring methods: standard scoring (red), penalty-based scoring (blue), and *d*-prime scoring (yellow). The shaded areas around the lines represent the 95% confidence intervals for the linear regression fit.

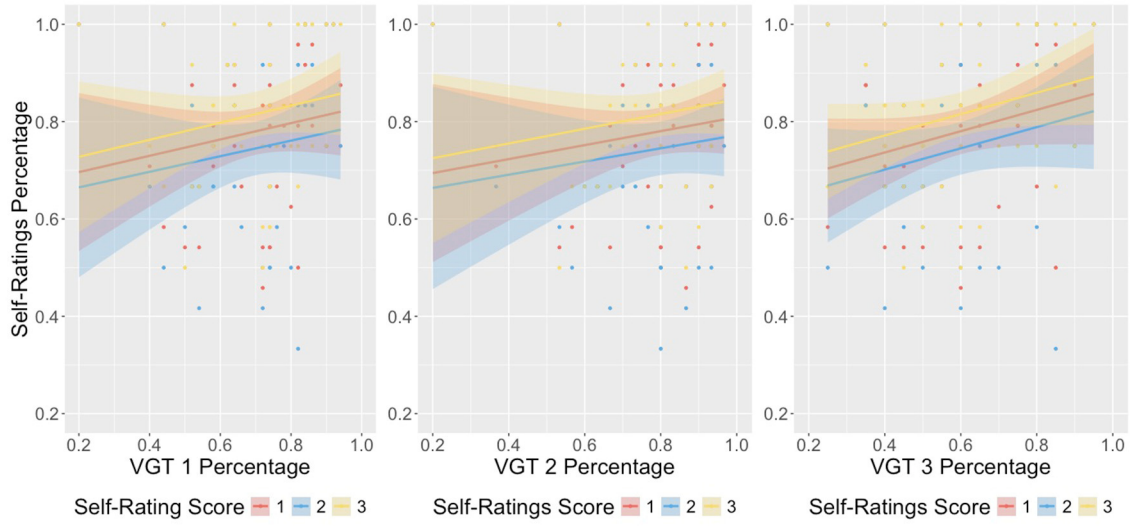

FIGURE 3
Scatter plots illustrating the relationship between VGT scores and Self-Ratings scores. The scatter plots display the percentage scores on the VGT (x-axis) against the percentage scores on the self-ratings (y-axis) for each score type. Recall that VGT 1 refers to the overall score, VGT 2 includes only the sentence-level vocabulary score, and VGT 3 includes only the paragraph-level vocabulary and grammar score. The color-coded lines and points represent the different self-rating scores: composite score (red), productive skills score (blue), and receptive skills score (yellow). The shaded areas around the lines represent the 95% confidence intervals for the linear regression fit.

than objective proficiency assessments. For self-ratings, the language entropy composite score was correlated with Self-Ratings 2 [$r_{s(43)} = 0.35$, $p < 0.05$]. That is, more integrated use of their languages overall was associated with higher self-reported productive Spanish proficiency. Similar positive relationships were also evidenced between language entropy subcategories for work, self, and friends, and self-reported productive Spanish proficiency [$r_{s(43)} = -0.38$, $p < 0.01$]. Note that language entropy in the family context was negatively correlated with productive Spanish proficiency, indicating that more compartmentalization of the two languages in this context was associated with higher self-reported productive skills in the HL. Although the language entropy composite score was not significantly correlated with Self-Rating 3 (i.e., receptive skills), there were significant correlations with 3 out of 5 of the language entropy subcategories, which followed the same pattern as productive self-reported Spanish proficiency. In

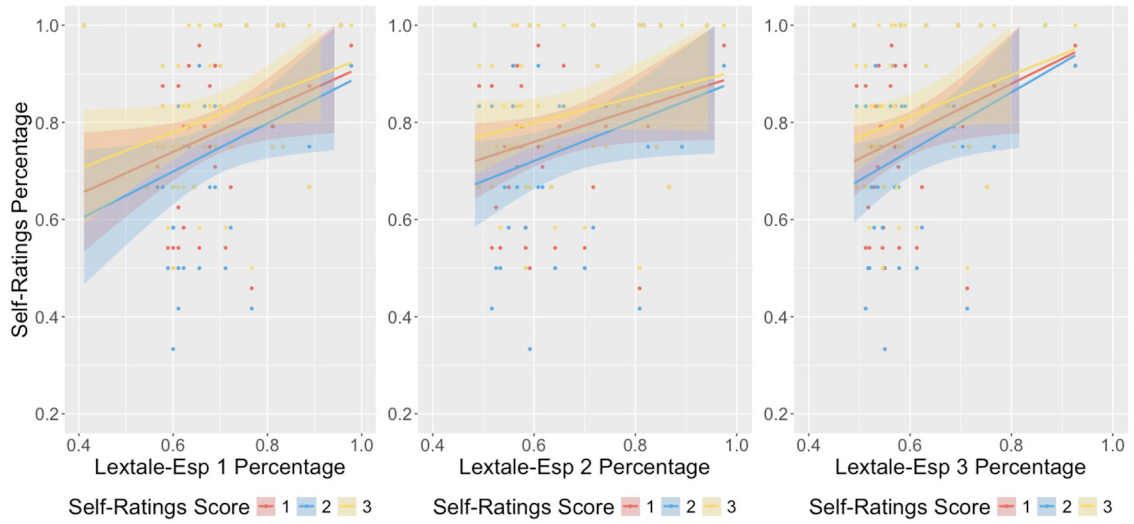

FIGURE 4
Scatter plots illustrating the relationship between LexTale-Esp scores and Self-Ratings scores. The scatter plots display the percentage scores on the Lextale-Esp (x-axis) against the percentage scores on the self-ratings (y-axis) for each score type. Lextale-Esp 1 refers to the standard scoring method, Lextale-Esp 2 refers to the penalty-based scoring method, and Lextale-Esp 3 refers to the *d*-prime scoring method. The color-coded lines and points represent the different self-rating scores: composite score (red), productive skills score (blue), and receptive skills score (yellow). The shaded areas around the lines represent the 95% confidence intervals for the linear regression fit.

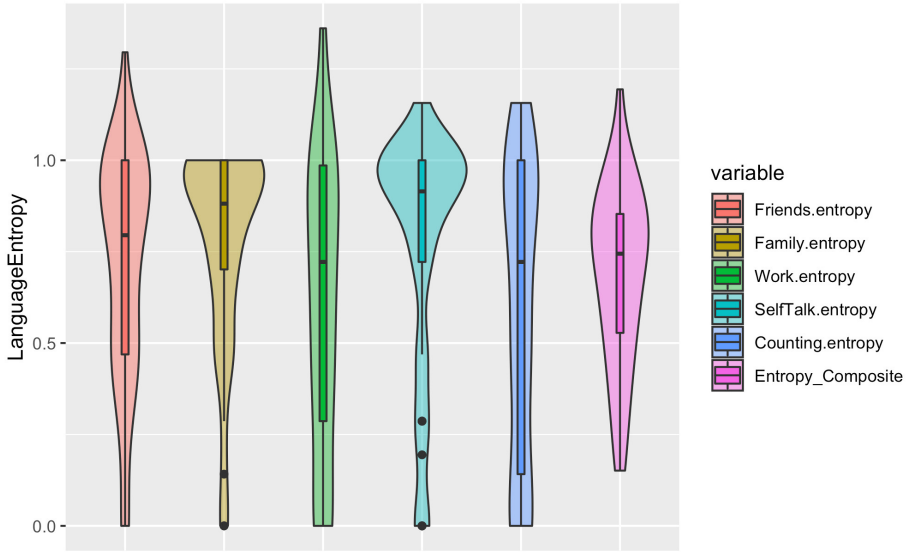

FIGURE 5
Overall language entropy distributions across different contexts of heritage language use. The violin plots illustrate the distribution of Language Entropy scores across different contexts, which are represented on the x-axis. These contexts include Friends, Family, Work, Self-Talk, Counting, and a Composite measure, each color-coded for clarity. The y-axis represents Language Entropy, where a score of 0 indicates complete compartmentalization of languages (languages are used separately in that context), and a score of 1 signifies full integration of two languages (both languages are used interchangeably). For individuals who speak more than two languages, scores can exceed 1, with a score of 1.585 indicating complete integration across all languages spoken (see Gullifer and Titone, 2020). The width of each violin plot reflects the density of data points, with wider sections indicating a higher concentration of values. The black bar within each plot represents the interquartile range, which shows where the middle 50% of the data points fall. This visualization allows for the comparison of language use patterns across various social and cognitive contexts.

particular, correlations with subcategories for friends [$r_{s(43)} = 0.43$, $p < 0.01$] and self [$r_{s(43)} = 0.48$, $p < 0.05$] were positive, whereas the correlation with the family subcategory was negative [$r_{s(43)} = -0.38$, $p < 0.05$].

Regarding relationships between language entropy and objective assessments of Spanish proficiency, the only significant correlation was between the family subcategory and the VGT and Lextale-Esp, which was again a negative correlation [$r_{s(43)} = -0.38$,

TABLE 4 Spearman correlations between objective and subjective language proficiency assessments and HL experience factors.

	Exposure	Schooling	Language entropy					
			Composite	Friends	Family	Work	Self	Count
VGT 1	**0.47****	0.07	−0.05	−0.07	**−0.38****	−0.06	0.00	−0.07
LxE 1	**0.42****	0.05	−0.07	−0.17	**−0.36***	0.04	0.01	0.03
Self-rating 2	−0.12	0.27	0.35	**0.43***	−0.32	0.32	**0.48****	0.24
Self-rating 3	−0.09	0.21	0.28	0.38	**−0.33***	0.20	**0.31***	0.23

VGT 1 = overall score; LxE (LexTale-Esp) 1 = standard scoring. Self-rating 2 = productive skills score; self-rating 3 = receptive skills score. Correlation coefficients and p values are rounded to two decimal places. Bold values indicate statistically significant correlations, with significance levels as follows: $^{*}p < 0.05$ and $^{**}p < 0.01$.

$p < 0.01$]. In other words, greater compartmentalization of their two languages in the family context was associated with higher performance on both types of objective proficiency assessments.

5 Discussion

This study aimed to critically examine the reliability and validity of commonly used objective (i.e., Lextale-Esp, VGT) and subjective (i.e., self-ratings) assessments to accurately and consistently characterize Spanish HL proficiency among a sample of HSs of Spanish in the US. Additionally, we explored the relationships between these assessments and various HL experience factors, including years of Spanish exposure, Spanish schooling, and language entropy. We address each RQ separately below.

RQ1: Reliability: What is the internal consistency of the selected objective proficiency assessments for this sample of HSs of Spanish?

In answering RQ1, our study investigated the internal consistency of the selected objective proficiency assessments. Our results indicated that both the sentence-level portion of the VGT and the Lextale-Esp demonstrated moderate to high internal consistency. Specifically, the Lextale-Esp exhibited a high Cronbach's alpha of 0.88. At the same time, the sentence-level VGT portion showed an alpha of 0.87. These findings suggest that these assessments performed reliably among this group of HSs. However, the paragraph-level portion of the VGT showed a lower Cronbach's alpha of 0.68, indicating moderate internal consistency. This could be due to the higher cognitive demands of integrating grammar and vocabulary knowledge with reading comprehension at the paragraph level, which introduced more complexity and potential for higher variability in performance among participants, or perhaps due to the formal nature of these paragraphs and the Peninsular culture, which participants may not have been familiar with.

These findings mostly support our hypothesis that the selected objective proficiency assessments would demonstrate high internal consistency within this sample of HSs of Spanish. The analyses indicated that within each test, the items appear to be consistently tapping into a single construct (or perhaps different but closely related constructs), although to a somewhat lesser degree for the paragraph-level portion of the VGT. This finding aligns with previous research showing reliable internal consistency for these tasks across diverse populations (e.g., Izura et al., 2014; Montrul et al., 2008). However, these analyses alone can neither determine the exact construct assessed nor confirm whether both assessments measure the same construct.

RQ2: (Validity): How do the selected objective and subjective proficiency assessments relate to one another for this sample of HSs of Spanish?

RQ2 explored the relationships among the different objective and subjective assessments of Spanish proficiency to examine validity, specifically construct validity. Our findings revealed that the objective assessments, the VGT and Lextale-Esp, were consistently positively correlated with each other with medium to large effect sizes [$r_{s(43)}$ =0.47 to 0.74]. This supports our hypothesis that these objective assessments, though qualitatively different, largely tap into similar constructs due to their heavy reliance on the speaker's breadth and depth of Spanish vocabulary knowledge, despite their differences in format (multiple and binary choice, for the VGT and Lextale-ESP, respectively). The tasks' shared reliance on vocabulary knowledge likely accounts for the observed correlations between the two assessments. However, the relationships between these objective assessments and the subjective self-ratings were more variable. Specifically, the VGT and Lextale-Esp showed inconsistent correlations with self-ratings. These findings align with previous research, which has similarly observed that while relationships between objective and subjective assessments exist, they tend to be variable and not as strong. This variability is often attributed to individual biases or differing subjective interpretations, as highlighted in studies by Tomoschuk et al. (2019) and Gullifer and Titone (2020).

Our findings underscore the complexity of assessing language proficiency among HSs of Spanish. While objective assessments such as the VGT and Lextale-Esp demonstrate strong internal consistency and are correlated due to their shared focus on vocabulary, they may not fully capture the diverse and multifaceted nature of language proficiency as experienced by HSs. The variability in the relationships between objective and subjective assessments aligns with previous studies that highlight the influence of individual perceptions and language experience on self-assessments of proficiency (Tomoschuk et al., 2019; Gullifer and Titone, 2020; Gehebe et al., 2023). Moreover, our results, in line with findings from studies such as those by Gehebe et al. (2023) and Hržica et al. (2024), suggest that relying solely on objective assessments could overlook essential aspects of language proficiency that are better captured through self-reports, particularly in contexts where language exposure and socio-cultural factors play significant roles. Therefore, our results indicate

that a balanced approach incorporating objective and subjective assessments is necessary for a comprehensive characterization of bilingual proficiency. Such an approach recognizes the limitations of each type of assessment while leveraging their strengths, providing a more nuanced and valid understanding of language proficiency in bilingual individuals.

When incorporating subjective assessments (self-ratings of proficiency) into our analyses, we found less alignment. In particular, both the VGT and Lextale-Esp only correlated with self-ratings about half the time each. Our findings suggest that the Lextale-Esp is most aligned with the productive-skill self-ratings of HSs of Spanish in the US. Although this may seem surprising given that the Lextale-Esp is a receptive-skills task, recall that the correlations, while significant, are small ($0.27 \leq \rho \leq 0.36$). Overall, the lack of a strong correlation between these objective and self-reported proficiency assessments highlights a disconnect between formal, objective assessments and HSs' self-perceptions of their own HL proficiency as previous research has also shown (Bayram et al., 2021b).

This aligns with our hypothesis that subjective and objective assessments would relate differently than the relationships between only the objective measures. As discussed in the Background section, previous research with both L2 learners and HSs has shown self-reports to be highly correlated with other well-documented, production-oriented, objective assessments of proficiency [e.g., specifically with both the EIT and SOPI for L2 learners in Bowden (2016); with the EIT for L2 learners and HSs in Faretta-Stutenberg et al. (2023)]. Why, then, did we find fewer and weaker correlations between self-ratings and the objective assessments in the present study? Some critical differences between prior work and the present study may lie in the objective assessments themselves. While the VGT and the Lextale-Esp are exclusively receptive tasks and require metalinguistic judgments for task completion, the EIT and SOPI are (at least in part) productive tasks and do not require the participant to reflect upon the language or provide a judgment. As such, our results highlight an important disconnect between how HSs perceive their Spanish proficiency and how objective proficiency assessments as an overarching construct tap into those same individual HL abilities.

RQ3:(Validity in Context): Do the selected objective and subjective proficiency assessments correlate similarly with each HL experience factors —, namely years of exposure to Spanish, years of Spanish schooling, and social diversity of HL use (i.e., language entropy) helping to determine if these HL experience factors are equally influential for capturing dimensions of Spanish proficiency and reflecting real-world HL use among Spanish HSs?

RQ3 aimed to examine the validity, specifically construct and ecological validity, of the objective (VGT, Lextale-Esp) and subjective assessments (self-ratings of productive and receptive HL skills) under investigation to determine the extent to which these proficiency assessments reflect Spanish proficiency in HSs of Spanish in the US, and how they relate to relevant HL experience factors, namely years of exposure to Spanish, years of schooling in Spanish, and language entropy scores (in particular, compartmentalization vs. integration across five distinct contexts of HL use: friends, family, work, self-talk, counting; and as a composite).

Correlations overall were of similar magnitudes and directions. However, we found no robust relationships between the Spanish proficiency assessments and these HL experience factors. For objective proficiency assessments, a total of four significant correlations with HL experience factors were revealed. Specifically, VGT and Lextale-Esp scores correlated with years of exposure and language compartmentalization with family. In contrast, for the subjective assessments, self-reported productive Spanish proficiency correlated with five HL experience factors (language integration with friends, work, and self; and language compartmentalization with family; as well as with the composite entropy score) and self-reported receptive Spanish proficiency correlated with three experience factors (language integration with friends and self, as well as language compartmentalization with family). Interestingly, although correlations between objective and subjective Spanish proficiency assessments and HL experience factors were not very strong, the experience factor *language entropy score for family* stood out, with small but significant negative correlations with all four proficiency assessments. Specifically, more compartmentalized language use in the family setting was associated with higher VGT and Lextale-Esp scores and higher receptive and productive Spanish self-ratings. Note that a more compartmentalized score does not indicate which language is being used more in a given context; thus, to better understand this result, we considered the self-report data, which revealed that Spanish was the more common language reported as being used in the family setting. These data suggest that HSs who reported interacting more with family in a single language (usually Spanish) also reported higher receptive and productive self-ratings and performed better on vocabulary-based objective assessments for Spanish. This finding aligns with existing research, which posits that extensive language exposure and engagement are cornerstones of bilingual proficiency (Kroll and Bialystok, 2013). Also of note, language entropy scores for friend and self-talk contexts showed positive, small-to-medium correlations with self-ratings of both productive and receptive Spanish proficiency. That is, more integrated language use in these contexts was associated with higher self-ratings.

On the surface, this combination of results may appear surprising and/or conflicting, but in fact, it makes sense that both greater compartmentalization in the family context, for example, if many older members of the family speak the HL, and greater integration exclusively in the friends and self-talk contexts, where HSs could potentially translanguage or codeswitch frequently, would develop their HL proficiency consistently. That is to say, it is perhaps imminently reasonable that higher self-perceptions of proficiency are aligned with differing degrees of language integration and compartmentalization in different contexts. The complex interplay between language entropy and proficiency assessments explored in RQ3 brings to light the intricate nature of language integration in everyday life for bilingual individuals. As such, the results support the argument for adaptive bilingualism, where individuals tailor their language use to specific contexts, thus developing a more dynamic and fluid language proficiency (e.g., DeLuca et al., 2019; Pliatsikas et al., 2020; Tiv et al., 2022).

To summarize, our results support the reliability of the VGT and Lextale-Esp, as evidenced by their internal consistency,

indicating that the test items within each assessment are consistently measuring similar constructs. However, the varying degrees of correlation with subjective proficiency assessments suggest a potential limitation in the construct validity of these objective measures for this group of HSs, implying that neither type alone can fully capture the multifaceted nature of HL proficiency. Additionally, the inconsistent alignment of the VGT and Lextale-Esp with HL experience factors further raises questions about their construct validity, as they may not fully reflect the diverse language experiences of HSs. In contrast, subjective proficiency assessments showed a stronger alignment with participants' HL experience factors, suggesting that they may better capture aspects of HL proficiency closely related to HSs' experiences and perceptions. From a validity perspective, this means that subjective assessments might provide a more comprehensive view of HL proficiency by incorporating elements of an individual's language use and context that objective measures may overlook. This enhanced alignment with personal and contextual factors suggests that subjective assessments could offer more valid insights into the practical and experiential dimensions of HL bilingualism. Therefore, integrating subjective self-assessments with objective assessments and HL experience factors is essential for a more holistic and valid understanding of an individual's HL proficiency. This approach not only improves the reliability of proficiency assessments but also ensures that they more accurately reflect the complex, dynamic, context-dependent nature of HL use, ultimately enhancing both the construct validity and overall comprehensiveness of the HL proficiency assessment.

Regarding ecological validity and inclusivity, these findings emphasize the practical applicability of proficiency assessments in real-world language use scenarios. Subjective assessments, being aligned with participants' HL experience factors, suggest that HSs of Spanish have a nuanced understanding of their own HL proficiency. This understanding may not be fully captured by objective assessments alone, highlighting the importance of inclusivity in assessment approaches. Additionally, these results point to the likely limitations of commonly used proficiency assessments in capturing the full realities of bilingual individuals' experiences, given the inconsistent and relatively weak relationships between objective and subjective assessments and language experience factors, and to the next steps of evaluating other proficiency assessments that have been used widely in research.

At the broadest level, our findings highlight the fact that the results of any study are directly dependent on the tools used to assess and operationalize a given variable and, as such, the choice of which tool(s) to use and the interpretation of the data obtained should be undertaken with great care. Incorporating subjective self-assessments alongside objective assessments can offer a fuller and complementary picture of an individual's HL proficiency, capturing both their actual objective performance, self-perceptions, and experience factors. This approach ensures that the diverse experiences and self-perceived abilities of HSs are recognized, promoting inclusivity in language assessment. Thus, we acknowledge that objective proficiency assessments are helpful in establishing a baseline to allow for comparisons across groups and/or studies. At the same time, researchers should be aware that these tasks cannot fully capture the multidimensional nature of HL proficiency.

6 Limitations, future directions, and implications

As with any research endeavor, this study has its limitations. First, we acknowledge that the sample size and, therefore, the statistical power of the current study are modest. A larger sample size may have been able to detect stronger correlations between proficiency assessments and offer a more comprehensive understanding of the questions addressed here. Second, because the present study examined data collected as part of a larger project, which investigated Spanish-English code-switching, the specific objective and subjective assessments were limited to those used in the larger study, which are indeed commonly used in the field. Naturally, countless other proficiency assessments could be analyzed similarly, but here, we were limited by the data available. As for HL experience factors included in our analyses, these were chosen to match prior research practices (e.g., years of exposure and years of formal education in the HL) as well as to address calls in the literature to better capture bilingual experiences (e.g., language entropy; Gullifer et al., 2021) but were limited. Indeed, the overall inconsistent pattern of relationships between our Spanish proficiency assessments and HL experience factors underscores the possibility that a different pattern of results might emerge if different experience factors were analyzed.

Additionally, it is crucial to acknowledge that the tools and tasks examined here likely do not fully encompass the broad spectrum of HSs' language abilities, particularly in oral and aural domains. Therefore, while our study contributes to a line of research aiming to examine and improve our research tools, it focuses on understanding how two commonly used written tasks relate to each other, to self-perceptions, and to language use and experience. The following steps in this line of research should thus involve investigating more holistic, aural, and oral tasks to complement these written assessments, ultimately providing a more comprehensive picture of HSs' communication abilities.

Furthermore, while our study provides insights into how compartmentalization and integration of language use in different contexts related to proficiency assessments among HSs of Spanish, it is important to note the absence of naturalistic data that could enrich our findings. Observational studies or experimental simulations using narrative tasks could provide deeper insights into how such HSs actually communicate in everyday settings, enhancing the ecological validity of our conclusions.

These limitations suggest a clear avenue for future research to more comprehensively explore the nuanced dynamics of HL bilingual communicative practices beyond what can be inferred from subjective self-reports and objective, standardized proficiency assessments. Future work could address these limitations by increasing the sample size and investigating relationships between different proficiency assessments and/or HL experience factors. We further note, as discussed above, that there are multiple ways to score the objective assessments used here. This study represents a first step in better understanding how different scoring techniques

for these objective Spanish proficiency assessments may impact the representation of HSs' diverse HL abilities. Additional research that more directly examines differences in the scoring of these tasks with different speaker populations would be beneficial for refining the use of these tasks in HL and L2 research more broadly.

A key goal of the present study was to explore applications of the results for future bilingualism research. In this respect, we recognize that researchers must make decisions about which tool(s) to employ in their studies based on theoretical, methodological, and practical considerations. The ongoing discussions within academic circles and on social media platforms like the Hispanic and Lusophone Linguistics Facebook group page about technical aspects related to different proficiency measures, such as the "Modified DELE," have highlighted the inherent challenges and complexities of characterizing bilingual language abilities accurately, underscoring the need for continued exploration and refinement of bilingual language assessment tools.[7]

Reflecting on our findings, we first advocate using a combination of objective and subjective assessments, as well as experience factors, to characterize bilingualism in research. Second, a concrete takeaway for researchers, based on our results, is that while Lextale-Esp and VGT (often referred to as the "Modified DELE" in prior research) appear to largely tap into similar skills, there are a few methodological and practical advantages for the Lextale-Esp. In our study, the Lextale-Esp was more correlated with self-ratings of productive language skills, highlighting its ecological validity, whereas the VGT, as a whole, was not. Moreover, it is freely available with open access, is quick and easy to administer, is self-scoring, and, even though the LexTale-Esp also seems to be drawn largely from Peninsular Spanish, it is less culturally bound than the VGT. Furthermore, the existence of forms designed to be parallel across different languages (e.g., Brysbaert, 2013 for French; Lemhöfer and Broersma, 2012 for English; https://lextale.com for Dutch and German) fosters comparability across studies in different languages. However, we acknowledge that further work is needed to provide more external validity evidence for the Lextale-Esp. Even so, researchers wanting to include a receptive, written assessment of heritage Spanish proficiency in their study design may prefer the Lextale-Esp over the VGT if they want something that more closely aligns with HSs' own perceptions. Acknowledging the foundational work of researchers who developed tools to assess language proficiency in the context of bilingualism, our research underscores the critical importance of continued investigation and collaboration among researchers and practitioners for the advancement of bilingual language assessment methods. Such research and dialogue can push the field toward robust and ecologically valid solutions to assessing proficiency and a deeper understanding of bilingual language proficiency. By embracing diverse perspectives and engaging in constructive debates, the research community can more effectively scrutinize, validate, and refine assessment tools. This collective effort is essential not only for ensuring the rigor and relevance of our research methodologies but also for making them more equitable, inclusive, and reflective of the broad spectrum of bilingual experiences. Such collaborative engagement allows our approaches to adapt and evolve in response to new challenges and insights, keeping our research methodologies at the cutting edge of bilingualism studies.

The complex relationships found between proficiency assessments and language experience factors in this study also have broader implications for how bilingualism is assessed and understood in various settings and the inclusivity of assessments. With regard to proficiency assessment, our findings suggest the need for a multifaceted approach that encompasses both the dynamic and integrative aspects of language use, which are vital to the lived experiences of bilingual individuals. In educational settings, these results have the potential to influence how HL programs are conceived. Based on our findings, we discourage curriculum developers and educators from relying solely on one measure to determine HL proficiency; indeed, at least for the assessments examined here, our results indicate that scores across proficiency assessments appear to diverge, and so, using a single measure may miss important information. Instead, we encourage the integration of both objective and subjective assessments into HL assessment practices, recognizing that neither alone will fully capture a HS's language ability or the subtleties of their bilingual experience, in an effort to use assessments that have greater ecological validity and are more inclusive. Thinking a bit further afield, there is an opportunity for educational curricula to be more reflective of the diverse language experiences of HSs, incorporating perspectives and activities that validate the adaptive nature of bilingualism that HSs often experience, and foster positive cultural and identity associations. On a policy level, these findings could inform how language proficiency is conceptualized within official standards. By moving away from a one-size-fits-all approach and adopting a more nuanced understanding of bilingualism, policymakers can create guidelines that support diverse educational pathways for HSs, ensuring that both language education and assessments are accessible, equitable, and inclusive.

7 Conclusion

This study evaluated the reliability and validity of Spanish proficiency assessments among HSs of Spanish in the US as part of a larger effort to assess proficiency in a more inclusive and ecologically valid way. Our findings revealed that both the VGT and the Lextale-Esp are reliable objective assessments of Spanish HL proficiency, showing strong internal consistency. However, while these assessments seem to reliably measure vocabulary-related skills, our findings revealed that their construct validity is limited; they do not seem to fully capture the multifaceted nature of HSs' language proficiency as perceived by the individuals themselves, as revealed by the variability in correlations between the objective and subjective proficiency assessments that were found.

In terms of construct validity, although the VGT and Lextale-Esp showed overlap, likely in assessing vocabulary knowledge, they failed to fully encompass the diverse and nuanced aspects of HL proficiency. The lack of consistent alignment with self-ratings suggests a gap between these assessments–especially the VGT–and HSs' perceived HL proficiency, which should be taken seriously. Furthermore, correlations between objective assessments and HL experience factors were not as robust as those with subjective

7 To access discussion of the origin of this task, see: https://www.facebook.com/groups/75113154059/permalink/10159170476054060/.

self-reports, underscoring the limitations of these tools in capturing the full complexity of HL proficiency and use.

Regarding ecological validity, our findings underscore that subjective assessments align more closely with the real-world experiences of HSs, as reflected by experience factors. Self-ratings of proficiency and language entropy showed complex patterns of relationships with objective proficiency assessments, indicating that HL use and experiences meaningfully shape individuals' self-perceptions. For instance, language use patterns, such as compartmentalization in family contexts vs. integration with friends and self-talk, were related to self-reported proficiency, suggesting that subjective insights offer valuable information about HL proficiency in real-world contexts.

Regarding inclusivity, our findings emphasize the need for a comprehensive approach that integrates objective and subjective assessments to capture the inherent diversity of HL proficiency. Objective assessments like the VGT and Lextale-Esp provide valuable data on specific language skills, but subjective self-assessments may be crucial for capturing the broader and more nuanced aspects of language proficiency that reflect HSs' lived experiences. This approach ensures that the diverse and dynamic nature of bilingualism and bilingual individuals' perceptions are recognized and valued, promoting more inclusive and equitable practices in language assessment. This is critical for researchers in the field of HL bilingualism as failure to do so can lead to a misrepresentation of bilingual language proficiency, which can harm HS communities by perpetuating prescriptive narratives about what it means to be bilingual (e.g., Flores et al., 2020; Kircher and Kutlu, 2023; Tseng, 2021). For example, if research continues to use language proficiency assessments that ultimately provide a reductive view of the linguistic abilities of HSs, negative consequences could include exacerbating linguistic insecurity, undervaluing the language skills of HSs, and reducing the maintenance of HLs in bilingual communities (e.g., Bayram et al., 2021b; Driver, 2023; Gonzalez, 2011; Sánchez-Muñoz, 2016). This is especially detrimental for minoritized and racialized communities, where perceptions about language proficiency are often intricately tied to identity and cultural practices (e.g., Flores and Rosa, 2015; Ortega, 2020). Thus, researchers working with HS populations have a collective ethical and social responsibility to be aware of and act sensitively to these issues so as not to perpetuate harm to these communities (e.g., Bayram et al., 2021b; Driver, 2024; Higby et al., 2023; Leeman et al., 2011; Leivada et al., 2023; López et al., 2023).

Finally, our findings underscore the fact that HL proficiency and experience(s) cannot be reduced to a monolithic construct quantifiable by standardized assessments and questionnaires alone; thus, a more comprehensive approach that encompasses both objective assessments of language proficiency and the subjective experiences of HSs is required. Such an approach involves exploring the rich diversity of HL trajectories and outcomes, while also considering the critical role of HSs' confidence and self-perception of their own experiences and HL abilities, and the application of these abilities in real-world contexts. In support of recent calls in the field (e.g., Dass et al., 2024; De Bruin, 2019; Gullifer et al., 2021; López et al., 2023; Rothman et al., 2023; Titone and Tiv, 2023) and in line with recent empirical work (e.g., Gehebe et al., 2023; Tomoschuk et al., 2019), we advocate for the continued investigation and use of multiple, multidimensional proficiency assessments and research methods for assessing and characterizing the diverse and dynamic nature of HL bilingual proficiency and experiences. Using a combination of carefully chosen objective and subjective assessments we may be able to triangulate data and provide a comprehensive and ecologically valid picture of HL bilingualism that appreciates and embraces its inherently diverse and dynamic nature. By doing so, we aim to join the collective effort of researchers, educators, and practitioners dedicated to promoting equitable and holistic practices in reshaping how HL bilingualism and bilingual communities are represented and supported, thereby contributing to a more inclusive society that values linguistic and cultural diversity as a strength in today's multilingual, multicultural world.

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 below: Open Science Framework (OSF) Repository "Ecological Validity and Inclusivity in Heritage Bilingualism Research: Examining Objective and Subjective 'Proficiency' Assessments and Language Experience Measures": https://osf.io/7xktg/?view_only=c74f357943c448dfb12b1bcfb3b51db0.

Ethics statement

The present study was reviewed and approved by the University of Alabama's Institutional Review Board (IRB) (Protocol ID 21-06-4687). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AL: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Validation, Writing – original draft, Writing – review & editing. BK: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Visualization, Writing – original draft, Writing – review & editing. BI: Data curation, Formal analysis, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing, Conceptualization. MF-S: Methodology, Writing – original draft, Writing – review & editing, Conceptualization. HB: Methodology, Writing – original draft, Writing – review & editing, Conceptualization.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The authors express their heartfelt gratitude to their funders for their generous and invaluable support. The College Academy

of Research, Scholarship, and Creative Activity (CARSCA) at the University of Alabama provided funding for participant payment. The César Nombela Talent Attraction Grant (2023-T1/PH-HUM-29098), awarded to Alicia Luque and funded by the General Directorate of Research and Technological Innovation from the Regional Government of Madrid (Spain), supported the dissemination of results at several international conferences. Additionally, the Department of Language and Culture (ISK) at UiT the Arctic University of Norway provided funding to cover open-access publishing costs.

Acknowledgments

We extend our thanks to our colleagues for their invaluable assistance in recruiting participants: Dr. Rodrigo Delgado, Dr. Inma Taboada, Dr. Clara Burgo, Dr. Mariška Bolyanatz Brown, Dr. Jennifer Cabrelli, Dr. Salvatore Callesano, Dr. Claudia Fernández, Dr. David Giancaspro, Dr. Alexandra Gonzenbach Perkins, Dr. Xabi Granja, Dr. Bradley Hoot, Dr. Cristina Lozano Argüelles, Dr. Silvia Perez-Cortes, Dr. Sara Stefanich, Dr. María Turrero García, and Dr. Janire Zalbidea. Our gratitude also extends to the attendees of the HL@Cross 2023 conference in Istanbul, Turkey, for their valuable feedback. Furthermore, we are profoundly grateful to the following individuals and teams for their enriching discussions that have not only inspired and motivated us but have also significantly strengthened our work, including Dr. Belem López, Dr. Brandy Piña-Watson, and Dr. Jennifer Cabrelli, along with our colleagues at UiT The Arctic University of Norway: the Heritage-bilingual Linguistic Proficiency in their Native Grammar (HeLPiNG) team—Dr. Jason Rothman, Dr. Maki Kubota, Dr. Yanina Prystauka, and Dr. Jiuzhou Hao, the members of the Psycholinguistics of Language Representation (PoLaR) Lab and the AcqVa Aurora Research Center, as well as the Redefining our Language Experience (ROLE) Collective, especially Dr. Ethan Kutlu and Dr. Savithry Namboodiripad. We also, thank our funders, The College Academy of Research, Scholarship, and Creative Activity (CARSCA) at the University of Alabama, the General Directorate of Research and Technological Innovation from the Regional Government of Madrid (Spain), and the Department of Language and Culture (ISK) at UiT The Arctic University of Norway. Finally, we also extend our gratitude to the two reviewers and guest editor that reviewed this manuscript for their invaluable feedback and support in refining and strengthening our 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

Abutalebi, J., and Green, D. W. (2016). Neuroimaging of language control in bilinguals: neural adaptation and reserve. *Bilingual.: Lang. Cogn.* 19, 689–698. doi: 10.1017/S1366728916000225

Achugar, M., and Pessoa, S. (2009). Power and place: language attitudes towards Spanish in a bilingual academic community in Southwest Texas. *Spanish in Cont.* 6, 199–223. doi: 10.1075/sic.6.2.03ach

Ali, F. (2023). Code-switching among heritage learners of Spanish: Attitudes, practices, and pedagogical implications. *Crit. Multilingual. Stud.* 10, 1–35.

Amengual, M., and Chamorro, P. (2016). The effects of language dominance in the perception and production of the Galician mid vowel contrasts. *Phonetica* 72, 207–236. doi: 10.1159/000439406

Andrews, S., and Hersch, J. (2010). Lexical precision in skilled readers: Individual differences in masked neighbor priming. *J. Exp. Psychol.: General* 139:299. doi: 10.1037/a0018366

Bachman, L. F., and Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests* (Vol. 1). Oxford: Oxford University Press.

Barrett, R., Cramer, J., and McGowan, K. B. (2023). *English With an Accent: Language, Ideology, and Discrimination in the United States*. Routledge: Taylor & Francis.

Bayram, F., Kubota, M., Luque, A., Pascual y Cabo, D., and Rothman, J. (2021b). You can't fix what is not broken: contextualizing the imbalance of perceptions about heritage language bilingualism. *Front. Educ.-Educ. Psychol.* 6:628311. doi: 10.3389/feduc.2021.628311

Bayram, F., Kupisch, T., y Pascual y Cabo, D., and Rothman, J. (2019). Terminology matters on theoretical grounds too!: Coherent grammars cannot be incomplete. *Stud. Second Lang. Acquisit.* 41, 257–264. doi: 10.1017/S0272263119000287

Bayram, F., Pisa, G. D., Rothman, J., and Slabakova, R. (2021a). "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.

Beaudrie, S., and Fairclough, M. (2012). *Spanish as a Heritage Language in the United* States: *The State of the Field*. Washington, DC: Georgetown University Press.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Royal Stat. Soc.: Series B (Methodological)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

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

Bermúdez-Margaretto, B., and Brysbaert, M. (2022). How efficient is translation in language testing? Deriving valid Spanish tests from established English tests. *PsyArXiv*.

Birdsong, D. (2014). Dominance and age in bilingualism. *Appl. Linguist.* 35, 374–392. doi: 10.1093/applin/amu031

Birdsong, D., Gertken, L. M., and Amengual, M. (2012). *Bilingual Language Profile: An Easy-to-Use Instrument to Assess Bilingualism*. Austin: COERLL, University of Texas at Austin.

Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., and Gillam, R. B. (2010). What you hear and what you say: language performance in Spanish–English bilinguals. *Int. J. Biling. Educ. Biling.* 13, 325–344. doi: 10.1080/13670050903342019

Bowden, H. W. (2016). Assessing second-language oral proficiency for research: the Spanish elicited imitation task. *Stud. Second Lang. Acquisit.* 38, 647–675. doi: 10.1017/S0272263115000443

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *Am. Psychol.* 32:513. doi: 10.1037/0003-066X.32.7.513

Brown, J. D. (2013). Classical theory reliability. *Compan. Lang. Assessm.* 3, 1165–1181. doi: 10.1002/9781118411360.wbcla054

Brysbaert, M. (2013). Lextale_FR. A fast, free, and efficient test to measure language proficiency in French. *Psychol. Belg.* 53, 23–37. doi: 10.5334/pb-53-1-23

Bullock, B. E., and Toribio, A. J. (2009). "Themes in the study of code-switching," in *The Cambridge Handbook of Linguistic Code-Switching*, eds. A. J. Toribio and B. E. Bullock (Cambridge: Cambridge University Press), 1–18.

Cacoullos, R. T., and Travis, C. E. (2018). *Bilingualism in the Community: Code-Switching and Grammars in Contact*. Cambridge: Cambridge University Press.

Canale, M., and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Appl. Linguist.* 1, 1–47. doi: 10.1093/applin/1.1.1

Carreira, M., and Potowski, K. (2011). Commentary: Pedagogical implications of experimental SNS research. *Heritage Lang. J.* 8, 134–151. doi: 10.46538/hlj.8.1.6

Carroll, J. B., and Freedle, R. O. (1972). *Language Comprehension and the Acquisition of Knowledge*. Washington DC: V. H. Winston & Sons.

Chateau, D., and Jared, D. (2000). Exposure to print and word recognition processes. *Mem. Cognit.* 28, 143–153. doi: 10.3758/BF03211582

Christoffersen, K. (2019). Linguistic terrorism in the Borderlands: language ideologies in the narratives of young adults in the Rio Grande Valley. *Int. Multilingual Res. J.* 13, 137–151. doi: 10.1080/19313152.2019.1623637

Crocker, L., and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334. doi: 10.1007/BF02310555

Cuetos, F., Glez-Nosti, M., Barbon, A., and Brysbaert, M. (2011). SUBTLEX-ESP: Frecuencias de las palabras españolas basadas en los subtítulos de las películas. *Psicológica* 32, 133–144.

Cummins, J. (2013). "Immigrant students' academic achievement: understanding the intersection between research, theory, and policy," in *Herausforderung Bildungssprache - und wie man sie meistert*, eds. I. Gogolin, I. Lange, U. Michel, and H. H. Reich (Münster: Waxmann Verlag), 19–41.

Dass, R., Smirnova-Godoy, I., McColl, O., Grundy, J. G., Luk, G., and Anderson, J. A. (2024). A Content Overlap Analysis of bilingualism questionnaires: Considering diversity. *Biling.: Lang. Cogn.* 2024, 1–7. doi: 10.1017/S1366728923000767

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

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. Nat. Acad. Sci.* 116, 7565–7574. doi: 10.1073/pnas.1811513116

Dewaele, J. M., Housen, A., and Wei, L. (2003). "Bilingualism: beyond basic principles," in *Multilingual Matters*.

Diependaele, K., Lemhöfer, K., and Brysbaert, M. (2013). The word frequency effect in first-and second-language word recognition: a lexical entrenchment account. *Quart. J. Exp. Psychol.* 66, 843–863. doi: 10.1080/17470218.2012.720994

Driver, M. (2023). Measuring and understanding linguistic insecurity in heritage and foreign language contexts: design and validation of a novel scale. *J. Multiling. Multicult. Dev.* 2023, 1–20. doi: 10.1080/01434632.2023.2265901

Driver, M. (2024). Realities of comfort and discomfort in the heritage language classroom: Looking to transformative positive psychology for juggling a double-edged sword. *Modern Lang. J.* 108, 147–167. doi: 10.1111/modl.12899

Duffield, N., and White, L. (1999). Assessing L2 knowledge of Spanish clitic placement: converging methodologies. *Second Lang. Res.* 15, 133–160. doi: 10.1191/026765899668237583

Duhachek, A., and Iacobucci, D. (2004). Alpha's standard error (ASE): an accurate and precise confidence interval estimate. *J. Appl. Psychol.* 89, 792. doi: 10.1037/0021-9010.89.5.792

Faretta-Stutenberg, M., Issa, B. I., Bowden, H. W., and Morgan-Short, K. (2023). Parallel forms reliability of two versions of the Spanish Elicited Imitation Task. *Res. Methods in Appl. Linguist.* 2:100070. doi: 10.1016/j.rmal.2023.100070

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

Ferré, P., and Brysbaert, M. (2017). Can Lextale-Esp discriminate between groups of highly proficient Catalan–Spanish bilinguals with different language dominances? *Behav. Res. Methods* 49, 717–723. doi: 10.3758/s13428-016-0728-y

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

Flores, N. (2020). From academic language to language architecture: challenging raciolinguistic ideologies in research and practice. *Theory Pract.* 59, 22–31. doi: 10.1080/00405841.2019.1665411

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., and Rosa, J. (2015). Undoing appropriateness: raciolinguistic ideologies and language diversity in education. *Harv. Educ. Rev.* 85, 149–171. doi: 10.17763/0017-8055.85.2.149

Flores, N., and Rosa, J. (2023). Undoing competence: coloniality, homogeneity, and the overrepresentation of whiteness in applied linguistics. *Lang. Learn.* 73, 268–295. doi: 10.1111/lang.12528

Flores, N., and Schissel, J. L. (2014). Dynamic bilingualism as the norm: envisioning a heteroglossic approach to standards-based reform. *TESOL Quart.* 48, 454–479. doi: 10.1002/tesq.182

Flores, N., Tseng, A., and Subtirelu, N. (2020). "Bilingualism for all or just for the rich and White? Introducing a raciolinguistic perspective to dual-language education," in *Multilingual Matters*.

Fuller, J. M., and Leeman, J. (2020). "Speaking Spanish in the US: the sociopolitics of language (Vol. 16)," in *Multilingual Matters*.

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. Inquiry Lang. Stud.* 18, 203–228. doi: 10.1080/15427587.2021.1935957

García, O., and Solorza, C. R. (2021). Academic language and the minoritization of US bilingual Latinx students. *Lang. Educ.* 35, 505–521. doi: 10.1080/09500782.2020.1825476

Gardner-Chloros, P. (2009). *Code-Switching*. Cambridge: Cambridge University Press.

Gehebe, T., Wadhera, D., and Marton, K. (2023). Interactions between bilingual language proficiency and exposure: comparing subjective and objective measures across modalities in bilingual young adults. *Int. J. Biling. Educ. Biling.* 26, 845–860. doi: 10.1080/13670050.2022.2125285

Gertken, L., 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.

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

Gonzalez, G. (2011). *Spanish Heritage Language Maintenance: The Relationship Between Language Use, Linguistic Insecurity, and Social Networks*. Tucson, AZ: The University of Arizona. Available at: http://hdl.handle.net/10150/144600

Green, D. W., and Abutalebi, J. (2013). Language control in bilinguals: the adaptive control hypothesis. *J. Cognit. Psychol.* 25, 515–530. doi: 10.1080/20445911.2013.796377

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain Language* 36, 3–15. doi: 10.1016/0093-934X(89)90048-5

Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, MA: Harvard University Press.

Gullifer, J. W., Kousaie, S., Gilbert, A. C., Grant, A., Giroud, N., Coulter, K., et al. (2021). Bilingual language experience as a multidimensional spectrum: Associations with objective and subjective language proficiency. *Appl. Psycholinguist.* 42, 245–278. doi: 10.1017/S0142716420000521

Gullifer, J. W., and Titone, D. (2018). *Compute Language Entropy with {languageEntropy}*. Available at: https://github.com/jasongullifer/languageEntropy

Gullifer, J. W., and Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingual. Lang. Cogn.* 23, 283–294. doi: 10.1017/S1470271642000052

Hao, J., Luque, A., Nakamura, M., Rossi, E., and Rothman, J. (2024). *Individual Differences and Event-Related Potentials in Bilingual Processing: Gender Agreement in Heritage Language Spanish*.

Higby, E., Gámez, E., and Mendoza Holguín, C. (2023). Challenging deficit frameworks in research on heritage language bilingualism. *Appl. Psycholinguist.* 44, 417–430. doi: 10.1017/S0142716423000048

Holguín Mendoza, C., Taylor, A., Romero Montaño, L., Lucero, A., and Dorantes, A. (2023). Too Latinx or not Latinx enough? Racial subtexts and subjectivities in a predominantly white university. *J. Latinos Educ.*22, 1138–1153. doi: 10.1080/15348431.2021.1920945

Hoot, B. (2020). "Does anyone know the definitive origin of the 50-question Spanish proficiency test used very often in L2 research... [Post from the Hispanic and Lusophone Linguists Facebook Group]," in *Facebook*. Available at: https://www.facebook.com/groups/75113154059/permalink/10159170476054060/ (accessed February 1, 2024).

Hržica, G., Košutar, S., and Poropat Jeletić, N. (2024). The relationship between self-assessment of language proficiency and measures of lexical diversity and syntactic complexity: Evidence from bilingual speakers of Italian in Croatia. *Front. Commun.* 9:1371126. doi: 10.3389/fcomm.2024.1371126

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

Hulstijn, J. H. (2015). "Explaining phenomena of first and second language acquisition with the constructs of implicit and explicit learning," in *Implicit and Explicit Learning of Languages*, ed. P. Rebuschat (Amsterdam: John Benjamins), 25–46.

Hyltenstam, K. (2016). *Advanced Proficiency and Exceptional Ability in Second Languages* (Vol. 51). Berlin: Walter de Gruyter GmbH & Co KG.

Hymes, D. H. (1972). "On communicative competence," in *Sociolinguistics: Selected Readings*, eds. J. B. Pride, & J. Holmes (New York City: Penguin), 269–293.

Izura, C., Cuetos, F., and Brysbaert, M. (2014). Lextale-Esp: a test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica: Int. J. Methodol. Exp. Psychol.* 35, 49–66.

Kałamała, P., Chuderski, A., Szewczyk, J., Senderecka, M., and Wodniecka, Z. (2023). Bilingualism caught in a net: a new approach to understanding the complexity of bilingual experience. *J. Exp. Psychol.: General* 152, 157–174. doi: 10.1037/xge0001263

Kelly, L. B. (2018). Interest convergence and hegemony in dual language: Bilingual education, but for whom and why? *Lang. Policy* 17, 1–21. doi: 10.1007/s10993-016-9418-y

Kircher, R., and Kutlu, E. (2023). Multilingual realities, monolingual ideologies: Social media representations of Spanish as a heritage language in the United States. *Appl. Linguist.* 44, 1077–1099. doi: 10.1093/applin/amac076

Kline, P. (2013). *Handbook of Psychological Testing*. London: Routledge.

Koronkiewicz, B. (2023). Inalienable possession in Spanish-English code-switching: Acceptability data from US heritage speakers of Spanish. *Spanish Heritage Lang.* 3, 5–42. doi: 10.5744/shl.2023.1949

Kostromitina, M., and Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: a meta-analysis. *Stud. Second Lang. Acquisit.* 44, 886–911. doi: 10.1017/S0272263121000395

Kroll, J. F., and Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *J. Cogn. Psychol.* 25, 497–514. doi: 10.1080/20445911.2013.799170

Kubota, M., Alonso, J. G., Anderssen, M., Jensen, I. N., Luque, A., Pereira Sores, S. M., et al. (2023). Bilectal exposure modulates neural signatures to conflicting grammatical properties: Norway as a natural laboratory. *Lang. Learn.* 74, 436–467. doi: 10.1111/lang.12608

Lee, J. S., and Wright, W. E. (2014). The rediscovery of heritage and community language education in the United States. *Rev. Res. Educ.* 38, 137–165. doi: 10.3102/0091732X13507546

Leeman, J. (2015). Heritage language education and identity in the United States. *Annu. Rev. Appl. Linguist.* 35, 100–119. doi: 10.1017/S0267190514000245

Leeman, J., and Martínez, G. (2007). From identity to commodity: ideologies of Spanish in heritage language textbooks. *Crit. Inquiry Lang. Stud.* 4, 35–65. doi: 10.1080/15427580701340741

Leeman, J., Rabin, L., and Román-Mendoza, E. (2011). Identity and activism in heritage language education. *Modern Lang. J.* 95, 481–495. doi: 10.1111/j.1540-4781.2011.01237.x

Leivada, E., Rodríguez-Ordóñez, I., Parafita Couto, M. C., and Perpiñán, S. (2023). Bilingualism with minority languages: why searching for unicorn language users does not move us forward. *Appl. Psycholinguist.* 44, 384–399. doi: 10.1017/S0142716423000036

Lemhöfer, K., and Broersma, M. (2012). Introducing LexTALE: a quick and valid lexical test for advanced learners of English. *Behav. Res. Methods* 44, 325–343. doi: 10.3758/s13428-011-0146-0

Li, X., Ng, K. K., Wong, J. J. Y., Lee, J. W., Zhou, J. H., and Yow, W. Q. (2021). Bilingual language entropy influences executive functions through functional connectivity and signal variability. *Brain Lang.* 222, 105026. doi: 10.1016/j.bandl.2021.105026

López, B., Luque, A., and Piña-Watson, B. (2023). Context, intersectionality, and resilience: Moving toward a more holistic study of bilingualism in cognitive science. *Cult. Divers. Ethnic Minority Psychol.* 29, 24–33. doi: 10.1037/cdp0000472

Luque, A., Rossi, E., Kubota, M., Nakamura, M., Rosales, C., López-Rojas, C., et al. (2023). Morphological transparency and markedness matter in heritage speaker gender processing: an EEG study. *Front. Psychol.* 14:1114464. doi: 10.3389/fpsyg.2023.1114464

Macmillan, N. A., and Creelman, C. D. (1996). Triangles in ROC space: history and theory of "nonparametric" measures of sensitivity and response bias. *Psychon. Bull. Rev.* 3, 164–170. doi: 10.3758/BF03212415

Mallonee Gertken, S. E. (2013). *Priming of relative clause attachment during comprehension in French as a first and second language* (Unpublished doctoral dissertation). The University of Texas at Austin, Austin, TX, United States.

Menke, M. R., and Malovrh, P. A. (2021). "The (limited) contributions of proficiency assessments in defining advancedness," in *Advancedness in Second Language Spanish: Definitions, Challenges, and Possibilities*, ed. M. R. Menke and P. A. Malovrh (Amsterdam: John Benjamins), 1–16.

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *Am. Psychol.* 50:741. doi: 10.1037/0003-066X.50.9.741

Montrul, S. (2005). Second language acquisition and first language loss in adult early bilinguals: exploring some differences and similarities. *Second Lang. Res.* 21, 199–249. doi: 10.1191/0267658305sr247oa

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. (2012). Dominant language transfer in Spanish heritage speakers and second language learners in the interpretation of definite articles. *Modern Lang. J.* 96, 70–94. doi: 10.1111/j.1540-4781.2012.01278.x

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. Acquisit.* 25, 351–398. doi: 10.1017/S0272263103000159

Olson, D. J. (2023a). A systematic review of proficiency assessment methods in bilingualism research. *Int. J. Bilingual.* 28, 163–187. doi: 10.1177/13670069231153720

Olson, D. J. (2023b). Measuring bilingual language dominance: an examination of the reliability of the Bilingual Language Profile. *Lang. Test.* 40, 521–547. doi: 10.1177/02655322221139162

O'Neill, R., Cornelius, E. T., and Washburn, G. N. (1981). *American Kernel Lessons: Advanced Student's Book*. Harlow: Longman.

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

Paradis, J. (2023). Sources of individual differences in the dual language development of heritage bilinguals. *J. Child Lang.* 50, 793–817. doi: 10.1017/S0305000922000708

Pascual y Cabo, D., and Prada, J. (2015). Understanding the Spanish heritage language speaker/learner. *Euro. Am. J. Appl. Linguist. Lang.* 2, 1–10. doi: 10.21283/2376905X.3.67

Pascual y Cabo, D., and Prada, J. (2018). Redefining Spanish teaching and learning in the United States. *Foreign Lang. Annals* 51, 533–547. doi: 10.1111/flan.12355

Pliatsikas, 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

Plonsky, L., and Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Lang. Learn.* 64, 878–912. doi: 10.1111/lang.12079

Poarch, G. J., Vanhove, J., and Berthele, R. (2019). The effect of bidialectalism on executive function. *Int. J. Bilingual.* 23, 612–628. doi: 10.1177/1367006918763132

Prada, J. (2021). "Translanguaging thinking y el español como lengua de herencia," in *Aproximaciones al estudio del español como lengua de herencia* (London: Routledge), 111–126.

Prada, J. (2022). Articulating translanguaging as pedagogy of empowerment for racialized, language-minoritized bilinguals: from concepto to proyecto through digital storytelling. *TESL Canada J.* 38, 171–185. doi: 10.18806/tesl.v38i2.1353

Puig-Mayenco, E., Chaouch-Orozco, A., Liu, H., and Martín-Villena, F. (2023). The LexTALE as a measure of L2 global proficiency: a cautionary tale based on a partial replication of Lemhöfer and Broersma (2012). *Linguist. Approach. Bilingual.* 13, 299–314. doi: 10.1075/lab.22048.pui

Rhodes, S., Cowan, N., Parra, M. A., and Logie, R. H. (2019). Interaction effects on common measures of sensitivity: choice of measure, type i error, and power. *Behav. Res. Methods* 51, 2209–2227. doi: 10.3758/s13428-018-1081-0

Rosa, J., and Flores, N. (2017). Unsettling race and language: toward a raciolinguistic perspective. *Lang. Soc.* 46, 621–647. doi: 10.1017/S0047404517000562

Rothman, J. (2009). Understanding the nature and outcomes of early bilingualism: Romance languages as heritage languages. *Int. J. Bilingual.* 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

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

Sánchez Walker, N., and Montrul, S. (2020). Language experience affects comprehension of Spanish passive clauses: A study of heritage speakers and second language learners. *Languages* 6:2. doi: 10.3390/languages6010002

Sánchez-Muñoz, A. (2016). "Heritage language healing? Learners' attitudes and damage control in a heritage language classroom," in *Advances in Spanish as a Heritage Language,* ed. D. Pascual y Cabo (Amsterdam: John Benjamins), 205–218.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Sheng, L., Lu, Y., and Gollan, T. H. (2014). Assessing language dominance in Mandarin–English bilinguals: convergence and divergence between subjective and objective measures. *Biling.: Lang. Cogn.* 17, 364–383. doi: 10.1017/S1366728913000424

Silva-Corvalán, C. (2014). *Bilingual Language Acquisition: Spanish and English in the First Six Years.* Cambridge: Cambridge University Press.

Solís-Barroso, C., and Stefanich, S. (2019). Measuring language dominance in early Spanish/English bilinguals. *Languages* 4:62. doi: 10.3390/languages4030062

Solon, M., Park, H. I., Dehghan-Chaleshtori, M., Carver, C., and Long, A. Y. (2022). Exploring an elicited imitation task as a measure of heritage language proficiency. *Studies in Second Lang. Acquis.* 44, 1095–1123. doi: 10.1017/S0272263121000905

Sulpizio, S., Del Maschio, N., Del Mauro, G., Fedeli, D., and Abutalebi, J. (2020). Bilingualism as a gradient measure modulates functional connectivity of language and control networks. *Neuroimage* 205:116306. doi: 10.1016/j.neuroimage.2019.116306

Surrain, 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

Surrain, S., and Luk, G. (2023). The perceived value of bilingualism among *U.S.* parents: The role of language experience and local multilingualism. *Transl. Issues Psychol. Sci.* 9, 460–471. doi: 10.1037/tps0000352

Tavakol, M., and Dennick, R. (2011). Making sense of Cronbach's alpha. *Int. J. Medical Educ.* 2, 53. doi: 10.5116/ijme.4dfb.8dfd

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Lang. Learn.* 44, 307–336. doi: 10.1111/j.1467-1770.1994.tb01104.x

Titone, D. A., and Tiv, M. (2023). Rethinking multilingual experience through a Systems Framework of Bilingualism. *Biling.: Lang. Cogn.* 26, 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

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

Toribio, A. J., and Duran, L. (2018). "Understanding and leveraging Spanish heritage speakers' bilingual practices," in *The Routledge Handbook of Spanish as a Heritage Language* (London: Routledge), 284–298.

Torres, J. (2018). The effects of task complexity on heritage and L2 Spanish development. *Canad. Modern Lang. Rev.* 74, 128–152. doi: 10.3138/cmlr.3770

Treffers-Daller, J. (2019). What defines language dominance in bilinguals? *Annual Rev. Linguistics* 5, 375–393. doi: 10.1146/annurev-linguistics-011817-045554

Tseng, A. (2021). 'Qué barbaridad, son latinos y deberían saber español primero': Language ideology, agency, and heritage language insecurity across immigrant generations. *Appl. Linguist.* 42, 113–135. doi: 10.1093/applin/amaa004

Valdés, G. (2005). Bilingualism, heritage language learners, and SLA research: opportunities lost or seized? *Modern Lang. J.* 89, 410–426. doi: 10.1111/j.1540-4781.2005.00314.x

Vermeiren, H., Vandendaele, A., and Brysbaert, M. (2022). Validated tests for language research with university students: Tests of vocabulary, general knowledge, author recognition, reading comprehension, reading frequency, attention check. *PsyArXiv*. doi: 10.1016/j.stueduc.2022.101124

Wagner, D., Bekas, K., and Bialystok, E. (2023). Does language entropy shape cognitive performance? A tale of two cities. *Bilingual.: Lang. Cognit.* 26, 1–11. doi: 10.1017/S1366728923000202

Yap, M. J., Balota, D. A., Tse, C. S., and Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: evidence for opposing interactive influences revealed by RT distributional analyses. *J. Exp. Psychol.: Learn. Memory Cognit.* 34, 495–513. doi: 10.1037/0278-7393.34.3.495

Zou, L. X., and Cheryan, S. (2017). Two axes of subordination: a new model of racial position. *J. Pers. Soc. Psychol.* 112, 696–717. doi: 10.1037/pspa0000080