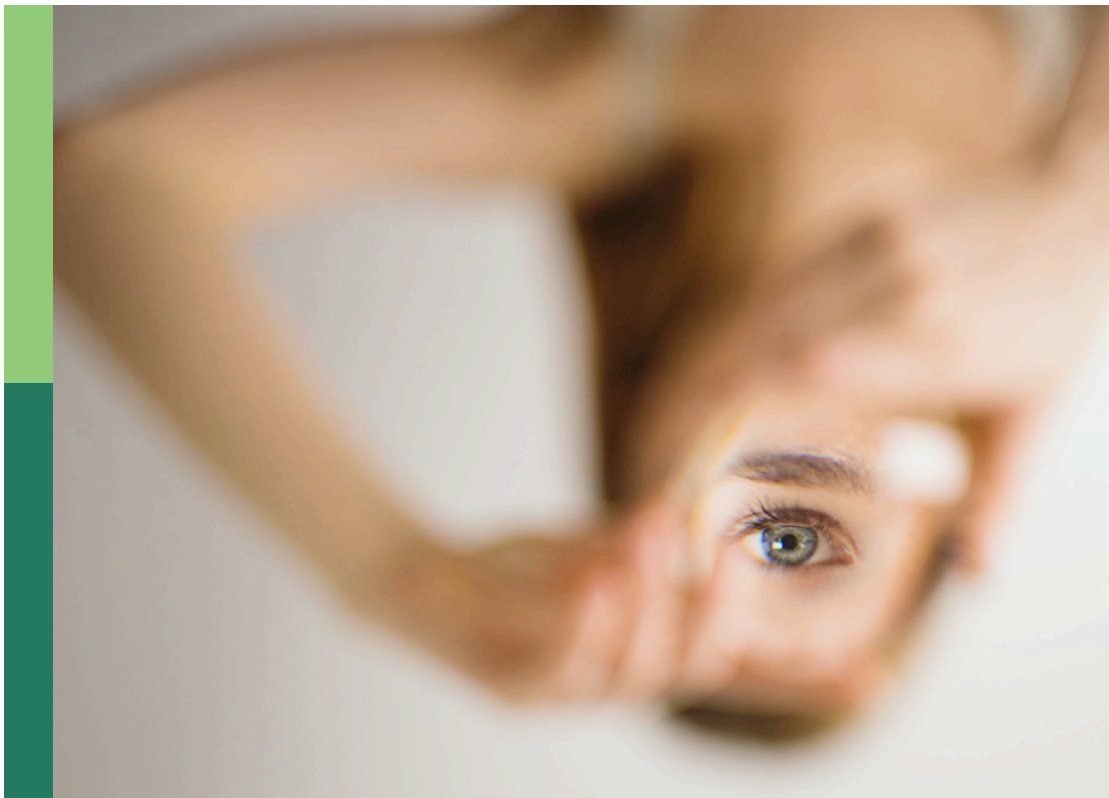
frontiers | Research Topics

Individual differences in second/foreign language speech production: Multidisciplinary approaches and new sounds

Edited by
Peijian Paul Sun, Boping Yuan, Xun Yan and Jimin Kahng

Published in
Frontiers in Psychology
Frontiers in Communication



ISSN 1664-8714
ISBN 978-2-8325-2837-2
DOI 10.3389/978-2-8325-2837-2

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

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

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

Individual differences in second/foreign language speech production: Multidisciplinary approaches and new sounds

Topic editors

Peijian Paul Sun — Zhejiang University, China
Boping Yuan — University of Cambridge, United Kingdom
Xun Yan — University of Illinois at Urbana-Champaign, United States
Jimin Kahng — University of Mississippi, United States

Citation

Sun, P. P., Yuan, B., Yan, X., Kahng, J., eds. (2023). *Individual differences in second/foreign language speech production: Multidisciplinary approaches and new sounds*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2837-2

Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 13 June 2023
DOI 10.3389/fpsyg.2023.1224059

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Peijian Paul Sun
✉ luapnus@zju.edu.cn

RECEIVED 17 May 2023
ACCEPTED 30 May 2023
PUBLISHED 13 June 2023

CITATION
Sun PP, Yuan B, Yan X and Kahng J (2023) Editorial: Individual differences in second/foreign language speech production: multidisciplinary approaches and new sounds. *Front. Psychol.* 14:1224059. doi: 10.3389/fpsyg.2023.1224059

Editorial: Individual differences in second/foreign language speech production: multidisciplinary approaches and new sounds

Peijian Paul Sun[1]*, Boping Yuan[2,3], Xun Yan[4] and Jimin Kahng[5]

[1]Department of Linguistics, Zhejiang University, Hangzhou, China, [2]School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China, [3]Faculty of Asian and Middle Eastern Studies, University of Cambridge, Cambridge, United Kingdom, [4]Department of Linguistics, University of Illinois at Urbana-Champaign, Champaign, IL, United States, [5]Department of Modern Languages, University of Mississippi, Oxford, MS, United States

KEYWORDS

individual difference, L2 speech production, fluency, test performance, gesture

Editorial on the Research Topic
Individual differences in second/foreign language speech production: multidisciplinary approaches and new sounds

Second/foreign language (L2) speech production can be subject to various individual difference factors. A review of the literature indicates that common individual difference factors examined in the field of L2 speaking include, but are not limited to, learning environments, learning strategies, willingness-to-communicate, motivation, and anxiety (Sun and Teng, 2021; Sun, 2022, 2023; Sun and Zhang, 2022). Although there is no consensus on what individual differences encompass, the literature has witnessed a growing interest in research on individual differences in L2 speaking. Nevertheless, our understanding of L2 speaking is still in its infancy compared to L2 writing. We, therefore, call for more research to shed light on the possible influence of individual difference factors on L2 speaking from multidisciplinary approaches.

Our call has resulted in 10 seminal works on individual differences in L2 speech performance and production (9 original research, 1 brief report, and 1 opinion). Specifically, three papers examined individual differences in L2 fluency (Aubrey; Feng; Kahng), two profiled L2 English speakers' oral test performance (Gao; Zhang et al.), and two tapped into elicited imitation in L2 speech production (Lei and Yan; Munro). The rest three papers explored the role of first language (L1, Zhang and Yuan) and gesture (Ma and Jin) in L2 speech production, and the connectedness in L2 speech development (Botezatu et al.).

Individual differences and L2 fluency

To understand the relationship between emotions (i.e., anxiety and enjoyment) and breakdown fluency from an intra-individual level, Aubrey adopted an idiodynamic approach to investigate English-as-a-foreign-language (EFL) university students' L2 speech performance in monolog tasks. The study found a strong positive connection between anxiety and breakdown fluency, but a weak association between enjoyment and breakdown fluency. The study also suggested that task design, task implementation, cognitive-linguistic, and achievement outcome factors could influence the emotion-fluency relationship.

To understand the L2 fluency development from a longitudinal perspective, Kahng examined L1 Chinese EFL learners' L2 utterance fluency change (i.e., speech, pausing, and repair) over 5-month study-abroad. The study showed that there was an improvement in learners' mean syllable duration and end-clause silent pausing. Such a change was the result of learners' increased L2 use rather than their L2 motivation.

However, both Scott's and Kahng's studies focused on within-individuals' rather than between-individuals' L2 fluency. In other words, L2 fluency should be examined not only from an individual perspective but also from a dialogic perspective to take into account the interaction between interlocutors. As Feng pointed out, incorporating both the monadic (intra-individual) and non-monadic (inter-individual) perspectives could enrich the ongoing discussion on factors contributing to L2 speech performance.

Test takers' L2 speech performance

Do high-proficiency EFL speakers differ in oral test performance? If yes, how can we create profiles to capture these differences? Gao's study has provided the audience with detailed steps for profiling high-proficiency EFL speakers' L2 speech performance. Specifically, employing cluster analysis on participants' speech fluency and vocabulary use, the study identified four types of high-proficiency EFL speakers.

In addition to Gao's approach to profiling high-proficiency EFL speakers' oral test performance, Zhang et al. examined Chinese EFL learners' speaking performance in integrated L2 speaking assessment tasks. The study revealed that problem-solving was the most frequently adopted metacognitive strategy in L2 speech production. However, metacognitive strategies were not found to have significant effects on participants' L2 speech performance across the four integrated speaking tasks in the study.

Regardless of test-takers' different profiles of L2 speech performance or different use of metacognitive strategies in L2 speech production, it is necessary to take test contexts into consideration. For example, to what extent can test anxiety and time pressure influence test takers' L2 speech performance? How different learners' L2 speech performance profiles or speaking strategies use will be in testing and non-testing conditions? Future research may consider addressing these issues.

Elicited imitation and L2 speech production

Elicited imitation (EI) is an effective measure of L2 proficiency (Yan et al., 2016, 2020). It requires participants to listen to a series of stimulus sentences, phrases, words, or sounds and then repeat them verbatim (Underhill, 1987). Despite the surge of interest in EI, little is known about the strategies employed by L2 learners in EI tasks and their effect on EI performance. Lei and Yun bridged the gap by examining L2 Chinese learners' strategy use in EI tasks. They found that participants adopted cognitive, metacognitive, communication, approach, and test-wiseness strategies in the process of elicited imitation. Additionally, cognitive strategies were found to have a significant positive influence on EI prompted L2 speech performance, while metacognitive strategies had a significant negative effect.

While differences in strategy use can impact EI performance in L2 speaking, the extent to which differences in ET tasks may influence learners' L2 speech performance in vowel production is under-researched. Munro, therefore, examined L1 Cantonese speakers' production of English high vowels (i.e.,/i/,/ɪ/,/u/,/ʊ/) on two EI tasks: a picture naming task and an interrupted repetition task. The study found that participants' vowel production on the interrupted repetition task was more intelligible than on the picture naming task by over 10 percentage points. However, the task effect on vowel production was inconsistent across speakers.

L1, gesture, and graph structure in L2 speech production

This collection also attracted researchers to explore L2 speech production from the angles of L1, gesture, and graph structure. Specifically, Zhang and Yuan investigated the influence of learners' L1 English and L1 Korean on their L2 Chinese oral production of ellipses. The study revealed that L1 influence on L2 Chinese learners' oral production of ellipses was not observed at the beginner level but observed at the intermediate and advanced levels. Ma and Jin investigated the relationship between co-speech gestures and L2 speech performance. Results showed that there were positive correlations between co-speech gestures and L2 speech measures in meaning and discourse. Last but not least, Botezatu et al.'s graph structure analysis of the relationship between L2 lexical retrieval and the global connectedness of narratives suggested that, in the early phases of L2 oral development, the connectedness of L2 speech can be attributed to individuals' ability in L2 lexical access.

Summing up, this collection of articles provides the audience with a multidisciplinary approach to understanding how individual difference factors may be associated with L2 speech production. The studies in the collection offer valuable insights into the complex nature of L2 speech production and highlight the importance of taking individual differences into account to support learners' L2 speech development.

Author contributions

PS drafted the editorial. All authors contributed to the refinement of the editorial and approved it for publication.

Funding

This work was supported by the National Social Science Foundation of China under Grant 19CYY008.

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

References

Sun, P. P. (2022). Strategic self-regulation for speaking English as a foreign language: scale development and validation. *TESOL Quart.* 56, 1369–1383. doi: 10.1002/tesq.3132

Sun, P. P. (2023). The influence of cognitive, affective, and sociocultural individual differences on L2 Chinese speech performance: A multiple indicators multiple causes approach. *J. Psycholinguist. Res.* 52, 217–239. doi: 10.1007/s10936-022-09847-x

Sun, P. P., and Teng, L. S. (2021). Why so nervous? Revisiting the sources of speech anxiety in Chinese as a second language. *System* 103, 102647. doi: 10.1016/J.SYSTEM.2021.102647

Sun, P. P., and Zhang, L. J. (2022). Speech competence and speech performance: a dual approach to understanding Chinese-as-a-second-language learners' speech production ability. *J. Sec. Lang. Stud.* 3, 2. doi: 10.1075./jsls.22002.jun

Underhill, N. (1987). *Testing Spoken Language: A Handbook of oral Testing Techniques*. Cambridge: Cambridge University Press.

Yan, X., Lei, Y., and Shih, C. (2020). A corpus-driven, curriculum-based Chinese elicited imitation test in US universities. *Fore. Lang. Annals*, 53, 704–732. doi: 10.1111/flan.12492

Yan, X., Maeda, Y., Lv, J., and Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: a narrative review and meta-analysis. *Lang. Test.* 33, 497–528. doi: 10.1177/0265532215594643

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 14 June 2022
doi: 10.3389/fpsyg.2022.876208

Understanding Individual Differences in Metacognitive Strategy Use, Task Demand, and Performance in Integrated L2 Speaking Assessment Tasks

Weiwei Zhang[1], Meijuan Zhao[2] and Ye Zhu[3]**

[1] School of Foreign Languages and International Education, Quzhou University, Quzhou, China, [2] Graduate Institute of Interpretation and Translation, Shanghai International Studies University, Shanghai, China, [3] Institute of Linguistics, Shanghai International Studies University, China Center for Language Planning and Policy Studies, State Language Commission of China, Shanghai, China

This study investigated the concept of individual differences (IDs) in the use of metacognitive strategies (planning, problem-solving, monitoring, and evaluating) and its relationship with task demand and learner performance within Kormos' Bilingual Speech Production Model from the lens of Chinese English-as-foreign-language (EFL) learners in the context of integrated L2 speaking assessment. To measure metacognitive strategies, we administered an inventory on 134 Chinese EFL learners after they completed four integrated L2 speaking assessment tasks. Descriptive analysis and multiple linear regression were adopted for data analysis, and results show that: (a) IDs displayed variance in Chinese EFL learners' metacognitive strategy use; (b) among the four metacognitive strategies under investigation, problem-solving was reported to be used the most frequently in sharp contrast to monitoring, which had the lowest frequency; (c) metacognitive strategies worked interactively, responding to task demands involved in the four integrated L2 speaking assessment tasks; and (d) Chinese EFL learners' use of metacognitive strategies, in individual and interactive working modes, had no relationship with their speaking performance. These results are expected to present some insights into the role of IDs in metacognitive strategy use during L2 speech production under assessment conditions, which will add robust evidence to the existing literature on L2 speaking, in particular on metacognitive strategy use in L2 speaking assessment. In the meantime, the findings will provide some empirical validation support for Kormos' model, which will further provide some implications for L2 speaking instruction and L2 assessment.

Keywords: individual differences in metacognitive strategy use, task demand, speaking performance, integrated L2 speaking assessment tasks, Kormos' Bilingual Speech Production Model

OPEN ACCESS

Edited by:
Peijian Paul Sun,
Zhejiang University, China

Reviewed by:
Mark Feng Teng,
Beijing Normal University, China
Zhengdong Gan,
University of Macau, Macao SAR, China

****Correspondence:***
Meijuan Zhao
zhaomj1@shisu.edu.cn
Ye Zhu
zhuye@shisu.edu.cn

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

Received: *15 February 2022*
Accepted: *29 April 2022*
Published: *14 June 2022*

Citation:
Zhang W, Zhao M and Zhu Y (2022) Understanding Individual Differences in Metacognitive Strategy Use, Task Demand, and Performance in Integrated L2 Speaking Assessment Tasks. Front. Psychol. 13:876208. doi: 10.3389/fpsyg.2022.876208

INTRODUCTION

In the four language skills (reading, listening, writing, and speaking), speaking is acknowledged as the most intricate productive skill to master, and speaking in a foreign language is even more complicated in that speaking is done in real-time, imposing heavy demands on speakers' abilities to use metacognitive strategies, core individual differences (IDs) construct (e.g., Luoma, 2004;

Yahya, 2019; Newton and Nation, 2020; Sun, 2020; Griffiths and Soruç, 2021). Therefore, speaking, among the four language skills, has been proposed to have the closest relationship with foreign and/or second language (L2) speakers' IDs in metacognitive strategy use, empowering L2 speakers to plan the knowledge at hand and to compensate for and facilitate their oral production so as to affect their ultimate speaking performance (Kormos, 2006, 2011; Bygate, 2011; Cohen, 2014). Nonetheless, such a salient role of IDs in metacognitive strategy use in L2 speaking has not been paid sufficient attention, and the available literature along this line of research inquiry primarily focuses on how L2 speakers use metacognitive strategies in non-assessment contexts (Zhang et al., 2021a). Consequently, how IDs in metacognitive strategy use functions in L2 speaking assessment for a smooth speech production still remains unclear, which rationalizes the research context of L2 speaking assessment in our study.

Additionally, there is extensive acknowledgment of the backwash effect of L2 assessment on L2 learning, and an ever-increasing recognition of adopting a holistic approach in L2 instruction through the use of integrated skill tasks involving multiple language skills to familiarize learners with authentic language use tasks for improving their language ability (Newton and Nation, 2020). As such, in formulating the research context of L2 speaking assessment, we embedded our study within the specific context of integrated L2 speaking assessment that involves not only speaking but listening and reading as well.

In the extant literature, although inconsistency still exists in the conceptualization of metacognitive strategies, an agreement has been reached among scholars (e.g., Cohen, 2014; Oxford, 2017) that studies into this concept should be contextualized in accordance with the specific language skill under investigation. Hence, to conceptualize the IDs construct of metacognitive strategies in L2 speaking, we framed our study within Kormos' (2006, 2011) Bilingual Speech Production Model (hereinafter referred to as Kormos' Model where necessary), a model that has been recognized as authentically duplicating the operating mode of metacognitive strategies in L2 speech production (Skehan, 2016, 2018). Furthermore, Kormos' Model has been widely applied in empirical studies on L2 speaking as the major Bilingual Speech Production Model (e.g., Kormos, 2011; Yahya, 2019) and has been accredited as "more elaborate and more targeted" (Wang and Liu, 2018, p. 397), compared with other L2 speech models, due to its solid theoretical grounding and strong empirical support (Wang and Liu, 2018; Yahya, 2019).

Taken together, we investigated IDs in metacognitive strategy use within Kormos' Model in the context of integrated L2 speaking assessment. In the research field of IDs, it is considered that IDs variables, including metacognitive strategies, interact with external factors such as context and tasks, affecting learner performance (e.g., Griffiths and Soruç, 2021). By the same token, in Kormos' Model, metacognitive strategies are also proposed to work, in independent and interactive manners, with tasks, exerting influence on performance. In line with this, to comprehensively study IDs in metacognitive strategy use in our research context, we set our focus on not only the concept *per se* but also its relationship with task and performance as well.

The novelty of our study is to add empirical evidence to the existing literature on IDs in metacognitive strategy use in L2 speech production under assessment conditions, while providing validation support for Kormos' (2006, 2011) model. Simultaneously, our study is expected to enrich the understanding of integrated L2 speaking assessment, an under-explored field (Frost et al., 2020; Zhang et al., 2021a). Moreover, the study is hoped to offer some implications for L2 speaking instruction, in particular, metacognitive instruction on L2 speaking and L2 assessment.

LITERATURE REVIEW

Individual Differences in Metacognitive Strategy Use

Evolving from the research domain of second language acquisition, the concept of individual differences (IDs) has developed into a formal field of scholarship, contributing to a veritable plethora of literature. Despite this, consensus on the taxonomies of IDs is far from absolute (e.g., Griffiths and Soruç, 2021). In a most recent publication, Griffiths and Soruç (2021) defined this concept as "characteristics which make learners different from each other and which affect the way that they behave in the classroom and beyond" (p. 341) based on their extensive review of the literature on IDs. They further proposed 11 learner variables (*viz*. motivation, aptitude, strategies, gender, culture/nationality/ethnicity/race, beliefs, autonomy, personality, style, age and affect) that attribute to IDs in terms of affecting language learning and teaching based on the findings of their empirical study, among which, strategies, especially metacognitive strategies, and motivation have a stronger influence on IDs in comparison with other variables. In addition to Griffiths and Soruç (2021), a large volume of literature has also evidenced the role of metacognitive strategies as a contributing variable to IDs, and hence the concept of metacognitive strategies is also termed IDs construct in the research arena of IDs (e.g., Oxford and Amerstorfer, 2018; Psaltou-Joycey and Gavriilidou, 2018). In accordance with this term, we labeled learners' use of metacognitive strategies as IDs in metacognitive strategy use in this study as shown throughout.

Metacognitive strategies originate from the field of psychology as a pivotal element of metacognition, a concept coined by Flavell (1979) that "refers to one's knowledge concerning one's own cognitive process and products or anything related to them" (p. 32). Since metacognition is multi-faceted, "multidimensional and domain-general in nature" (Teng et al., 2021a, p. 169), a consistent debate has been existing around the definition and components of the concept during its evolution (e.g., Zhang et al., 2021b). Regardless, it is acknowledged that the foundational research on metacognition takes root in two frameworks proposed by Flavell (1979) and Brown (1987) (Sperling et al., 2012; Nazarieh, 2016), in which metacognition is agreed to be comprised of metacognitive knowledge encompassing person/declarative knowledge, task/procedure knowledge and strategy/conditional knowledge, and metacognitive regulation or metacognitive strategies composed of planning (planning individual's learning activities in accordance with their learning objectives prior to L2 learning), monitoring (online monitoring in the individuals' learning process) and evaluating

(post-learning evaluating of the learning process) with the three components working independently and interactively (refer to Teng et al., 2021b; Zhang, 2021, for a review). In the most updated model of metacognition, though focusing on writing, established by Teng et al. (2021a), metacognitive strategies are also proposed to comprise three key components: planning, monitoring, and evaluating.

Due to the crucial role of metacognition in language learning and teaching (e.g., Oxford, 2017), metacognitive strategies are also recognized as one form of language learning strategies (LLSs) and have been reported to be the most important LLSs in a learner's successful learning (Zhang and Zhang, 2018; Gan et al., 2020). Like metacognition, the trajectory of metacognitive strategies is also characterized by debate on the concept's definition and taxonomies in the field of LLSs, which has been manifested by various models, including the widely applied Oxford's (1990) Strategy System Model of Learning Strategies and O'Malley and Chamot's (1990) Strategy Taxonomy Model. In spite of their seeming differences, all these models "reflect relatively the same categorizations of language learning strategies without any fundamental changes" (Zare, 2012, p. 164), and the key elements of metacognitive strategies across these models are consistent: planning, problem-solving, monitoring, and evaluating. Planning refers to L2 learners' learning activities for achieving their learning goals before L2 learning; problem-solving implies the employment of various methods to solve learning problems such as substitution, inferencing, and the use of gap fillers; monitoring denotes L2 learners' online inspection of their learning process; and evaluating images learners' post-learning assessment of their learning process (refer to Zhang, 2021, for an overview). The four metacognitive strategies operate independently and interactively to influence performance through their interactions with external tasks.

It is obvious that there is a great overlap between the research field of metacognition and the LLSs in terms of the components and the working mode of metacognitive strategies (An and Gan, 2021). In fact, such an overlap is also manifested in L2 speech production within Kormos' Model as reviewed in the subsequent section.

Conceptualizing Individual Differences in Metacognitive Strategy Use in Kormos' Model

As stated earlier (refer to "INTRODUCTION" section), the conceptualization of IDs in metacognitive strategy use during L2 speech production in our study was conducted through framing the concept in Kormos' Model. In the research field of speaking, models generated in psycholinguistics are broadly acknowledged and employed (e.g., Kormos, 2006, 2011; Skehan, 2016; Yahya, 2019; Sun, 2022a,b), among which Levelt's (1989) model of monolingual speech production has become "one of the most comprehensive and widely used theoretical frameworks" for monolingual speech production (Sun, 2016, p. 27). Based on this model, De Bot (1992) proposed his L2 speech production model, followed by many similar research efforts (e.g., Poulisse and Bongaerts, 1994; Towell et al., 1996). More recently, integrating Levelt's (1989) L1 model and existing L2 speech models, Kormos (2006, 2011) mapped out the Bilingual Speech Production Model, a substantially influential bilingual speech model employed in L2 speaking studies (Wang and Liu, 2018).

Kormos' Model is modular, and it consists of separate encoding modules: a conceptualizer for planning the message, a formulator for linguistically encoding the message, and an articulator for articulating the encoded message as sounds. In addition, the model also encompasses a large knowledge store labeled as long-term memory which comprises elements such as lexicon and syllabary that provide L2 speakers with the information needed; a speech comprehension system receives L2 speakers' actual discourse for inspection *via* monitoring, and an audition component (an acoustic-phonetic processor) helps the monitor to check the produced utterance. The monitor is based on the conceptualizer, monitoring the outputs of the conceptualizer, the formulator, and the whole process of speech production.

In correspondence to the four modules are the four stages in L2 speech production. They include: conceptualization in which speakers plan, at macro- and microlevels, what to speak or the intended message; formulation where speakers encode linguistically the intended message; articulation through which speakers execute their speech sounds by controlling the articulatory muscles, converting the phonetic plan generated in formulation to overt speech; and monitoring with which speakers check and notice errors for possible modifications and corrections to make their utterance in light of the task demands. Although the role of evaluating is not explicitly emphasized in the model, it should be noted that during monitoring in the different stages of L2 speech production, speakers are assumed to use evaluating in tandem with monitoring (O'Malley and Chamot, 1990) because without evaluating, L2 speakers are unlikely to make comparisons between the preverbal plan produced in conceptualization and the intended messages encoded in the formulation. Similarly, when L2 speakers use monitoring to examine the internal speech in the formulation and the overt speech in articulation, they have to use evaluation; otherwise, they cannot judge whether or not their actual utterances are consistent with task demands (Purpura, 1999). In other words, evaluation plays an equally important part as monitoring in L2 speech production (Zhang, 2021).

In L2 speech production, since speakers' L2 knowledge may not be complete, it is unavoidable that they will encounter problems (Dörnyei and Kormos, 1998), and how speakers solve these problems is demonstrated by problem-solving mechanisms in Kormos' Model. According to Kormos (2006, 2011), there are four types of problems that L2 speakers may tackle in speech production. The first type is resource deficit which normally occurs in conceptualization and formulation, relating to L2 speakers' language knowledge gap that prevents them to verbalize their intended messages. For problems of this type, solutions include substitution. Time pressure is another type of unavoidable problem that L2 speakers frequently encounter in planning and processing their speech. Solutions to the problem are pauses and repetitions, such as the use of gap fillers. The third type of problem is perceived deficiencies

in L2 speakers' language output displayed by the incorrectness or inappropriateness of their utterances, and relevant solutions are self-repair and self-appraising. Finally, the fourth type of problems is the perceived deficiencies in the interlocutor's performance, which are commonly solved by L2 speakers through the use of communicative strategies immediately related to metacognitive strategies (e.g., Bachman and Palmer, 2010; Ellis et al., 2019). It is evident that all the problem-solving mechanisms operating in Kormos' Model essentially replicate L2 speakers' employment of the metacognitive strategy of problem-solving as delineated earlier.

To summarize, the IDs construct of metacognitive strategies works in the forms of planning, problem-solving, monitoring, and evaluating individually and interactively, affecting L2 learners' speaking performance in Kormos' Model. Integrating the taxonomies and working mode of metacognitive strategies in this model with those in the literature on metacognition and LLSs as reviewed above, we conceptualized the IDs construct of metacognitive strategies as four metacognitive strategies: planning, problem-solving, monitoring, and evaluating, and accordingly IDs in metacognitive strategy use are conceptualized as individuals' use of the four metacognitive strategies which function independently and interactively, responding to task demand and impacting performance.

Integrated L2 Speaking Assessment Tasks: Task Demand

Integrated L2 speaking assessment tasks stand for a comparatively new and dynamic assessment/testing format that integrates reading, listening, and speaking to measure L2 learners' speaking performance (Frost et al., 2020). In comparison with other L2 tasks, this type of tasks is closer to real-world speaking tasks, which normally require speakers either to listen or read or to listen and read before speaking. It is agreed that integrated L2 speaking assessment tasks elicit a broad range of strategies from L2 speakers and have an intimate relationship with learners' use of metacognitive strategies (Barkaoui et al., 2013; Cohen, 2014). The close connections to real oral communications have made integrated L2 speaking assessment tasks ideal tasks for L2 speaking classroom instruction (Zhang et al., 2021a,b). Regardless, limited attention has been devoted to this test format in research actuality, which also accounts for the contextualization of our study as noted previously. In the available studies on integrated L2 speaking assessment tasks, the common practice is that researchers adopted the test of English as a foreign language (TOEFL) iBT integrated speaking section composed of four tasks that involve varying degrees of task demand (Barkaoui et al., 2013). Following the spirit of these researchers, we also contextualized our research in this test which concomitantly serves as one of the instruments in this study as described later.

In Kormos' Model, the four TOEFL iBT integrated speaking tasks are proposed to influence L2 speakers' purposeful use of metacognitive strategies on the grounds that conceptualization, formulation, and monitoring are subject to L2 speaker's conscious attentional control determined by task demands. As pointed out by Kormos, an individual's attention resources are limited; hence, the three stages in L2 speech production naturally compete with one another for the attention available. How the limited attention is allocated among the three stages is considerably impacted by task demands. For instance, when task demands are increased, L2 speakers are expected to allocate increasing attention to analyze task characteristics and to plan the conceptualization. As a result, a more complex preverbal plan may be generated. To encode the plan with increased task demands from the perspective of linguistics, L2 speakers are very likely to invest more attention in formulation. After L2 speakers consciously increase the amount of their attention to conceptualization and formulation, the attentional resources controlled by these speakers for monitoring and evaluating will be accordingly reduced, which indicates that more errors may be undetected in various stages of speech production, including the speakers' final speech. Hence, the quality of the speakers' performance will be negatively affected. The working mode of the four metacognitive strategies subject to attentional control caused by task demand variability in Kormos' Model is illustrated in **Figure 1**, which essentially illustrates our study framed in the model.

Empirical Studies

As stated earlier, studies on IDs in metacognitive strategy use and its effect on learner performance contextualized in L2 speaking assessment are quite limited. In these studies, those that were conducted in the specific context of integrated L2 speaking assessment are even fewer. To our knowledge, most of the current literature on IDs in metacognitive strategy use has focused either on the relationship between the concept and performance contextualized in the other three language skills (e.g., listening by Nett et al., 2012) and the relationship between tasks and performance (e.g., Rukthong and Brunfaut, 2020), or on the relationship between the concept, tasks, and performance in the context of independent speaking tests (e.g., Fernandez, 2018). In fact, there are only three studies, Swain et al. (2009), Yi (2012), and Barkaoui et al. (2013), in the literature that have investigated the intricate relationship between IDs in metacognitive strategy use, task demand, and performance contextualized in integrated L2 speaking assessment, as was the case with the present study.

In an exploratory approach, Swain et al. (2009) investigated 30 Chinese EFL learners' metacognitive strategy use in processing the TOEFL iBT integrated speaking tasks and its relationship with their speaking performance reflected by their test scores. The study showed that the participants frequently used metacognitive strategies, and there was no direct relationship between metacognitive strategy use and speaking performance. Barkaoui et al. (2013) re-conducted this study and reached similar findings. In the same research methodology as in Swain et al.'s (2009), Yi (2012) collected data on six Korean EFL university students' metacognitive strategy use and test scores in performing TOEFL-based speaking test tasks in both testing and non-testing conditions. The subsequent data analysis disclosed that the participants used metacognitive strategies frequently, and a weak relationship between metacognitive strategy use and speaking performance under both testing and no-testing conditions was founded.

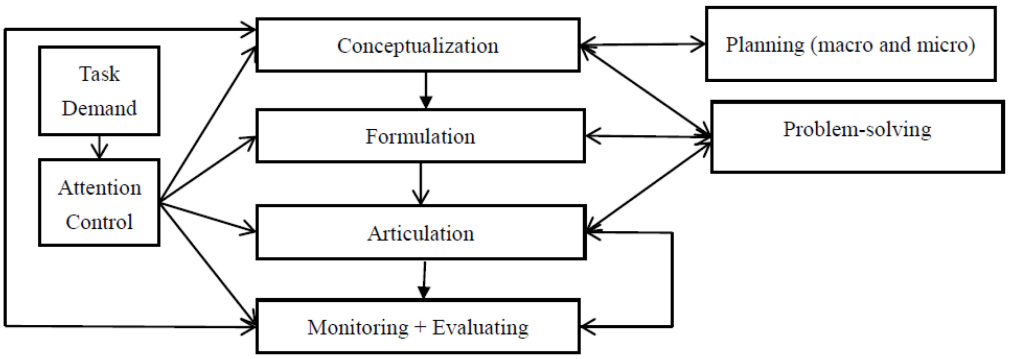


FIGURE 1 | Working mode of metacognitive strategies in Kormos' Model.

However, as Zhang (2021) has pointed out, none of these studies are without limitations, and this indicates the research gaps that we aimed to fill in this study. First, researchers collected data on a small sample (no more than 30), which places the validity and generalizability of the research findings into question (Creswell and Creswell, 2018). Second, since a study into metacognitive strategies is proposed to be contextualized in accordance with the language skill it intends to investigate as noted afore (Cohen, 2014; Oxford, 2017), the exploratory approach to researching L2 learners' metacognitive strategy use deployed in these studies suggests a lack of focus on the IDs construct operating specifically in L2 speech production where metacognitive strategies work in the form of planning, problem-solving, monitoring, and evaluating as delineated in the above literature review. Last but not the least, metacognitive strategies used by the participants in the studies were investigated in an individual manner, and the interactions within the components of metacognitive strategies as well as their response to task demands were not examined, which was not consistent with either the working principle of IDs in metacognitive strategy use or the working mode of metacognitive strategies as reviewed earlier.

METHODOLOGY

Research Questions

To fill the above research gaps, built upon our review of the literature, our study addressed the following research questions (RQ) through an investigation into a rather large sample size formulated by 134 Chinese EFL learners, and our examination of IDs in metacognitive strategy use covered both the independent and the interactive aspects of the working mode of metacognitive strategies:

> **RQ1**: How do Chinese EFL learners' IDs in metacognitive strategy use work in the context of integrated L2 speaking assessment tasks within Kormos' Model?
>
> **RQ2:** What are the relationships among Chinese EFL learners' IDs in metacognitive strategy use, task demand, and their performance in the context of integrated L2 speaking assessment tasks within Kormos' Model?

Participants

As noted above, our study involved a total of 134 Chinese EFL learners by means of convenience sampling. On a voluntary basis, the participants came from two universities situated on the Mainland of the People's Republic of China. The percentages of male and female students are 38 and 62%, and their age range was 18–21 years. All the participants have passed College English Test—Band 4 (CET-4), an authoritative English language proficiency test with high reliability administered specifically to university students in China, which, to a great extent, guarantees the participants' language proficiency to perform the four TOEFL iBT integrated speaking tasks for the smooth progress of this study (Zhang, 2021). Two trained raters, who had the experience in rating the TOEFL iBT integrated speaking section, scored the Chinese EFL learners' speaking performance.

Instruments

The Strategic Competence Inventory for Computer-Assisted Speaking Assessment

We deployed the Mandarin Chinese version of Zhang et al.'s (2021a) Strategic Competence Inventory for Computer-Assisted Speaking Assessment or SCICASA to measure the metacognitive strategies used by Chinese EFL learners in our study. The rationales for doing so are as follows: (a) Our research context, the TOEFL iBT integrated speaking tasks, is one form of computer-assisted speaking assessment. (b) The four operating forms of IDs in metacognitive strategy use under investigation are planning, problem-solving, monitoring, and evaluating, which are consistent with the four dimensions of the strategic competence in the inventory. (c) Inventories are widely applied in exploring L2 learners' internal metacognitive activities. (d) The native language of the Chinese EFL learners in our study is Mandarin Chinese.

The SCICASA has high validity and reliability ($\alpha = 0.941$), and it has 23 items classified into four dimensions: planning, problem-solving, monitoring, and evaluating. Five structured questions on L2 learners' background information (e.g., age,

gender, and language proficiency) are also included in the inventory. A 6-point Likert scale is used for each item: 0 (never or almost never), 1 (rarely), 2 (sometimes), 3 (often), 4 (usually), and 5 (always or almost always) (refer to Zhang et al., 2021a, for the detailed documentation of the inventory).

TOEFL iBT Integrated Speaking Tasks

The TOEFL iBT integrated speaking section served both as the research context of our study and as the instrument that was used to elicit Chinese EFL learners' speaking performance, as noted previously. Because our study was conducted in 2018 before the most recent reform in TOEFL iBT integrated speaking section that took place in 2019, the speaking section that we selected came from the old version of this test. The section comprised of four tasks (Task 1, Task 2, Task 3, and Task 4) that involve varying degrees of task demand. Given the participants' language proficiency reflected in their responses to the SCICASA, we selected one section for L2 learners with an intermediate level of language proficiency from the database for TOEFL iBT integrated speaking. To address RQ2 which concerns the relationship between Chinese EFL learners' metacognitive strategy use and the variability in task demands, we used all of the four different tasks in the section, and to ensure task validity, we did not make any modifications on the four tasks selected.

In the section, Task 1 presents a reading passage on a university's new plan for shuttle route change, which is followed by a discussion between two university students on the plan in the listening section. After that, task-takers are required to state one of the speakers' opinions on the new change. Task 2 provides a reading passage on a psychological concept: audience effect. In the following listening material, a lecture on this topic is delivered, and task-takers are asked to use the examples given in the listening section to explain the concept in the reading material. Task 3 involves a conversation between a professor and a female student on time conflict. To solve the conflict, the professor offers the female student two possible solutions, with neither sounding satisfactory to her. Task-takers are required to recommend one specific solution to the conflict and give the reasons why they believe such a solution might work. Task 4 is a lecture on two definitions of money in the listening section. The broad definition refers to both bills and the barter system. The narrow definition indicates the legal tender or whatever is accepted as payment such as coins in a society. Task-takers are asked to explain the two forms of money with the examples used by the professor in the lecture. Time for preparation before speaking is different, with 30 s for Task 1 and Task 2 and 20 s for Task 3 and Task 4. The varying degrees of task demand involved in the four tasks are displayed in **Table 1**.

TABLE 1 | Variability in task demands in the four integrated speaking test tasks.

Tasks	Preparation time	Topic content	Language skills	Task type
Task 1	30 s	Campus-life situation	R-L-S	Opinion narrating
Task 2	30 s	Academic lectures	R-L-S	Concept-illustrating
Task 3	20 s	Campus-life situation	L-S	Problem-solving
Task 4	20 s	Academic lectures	L-S	Concept-illustrating

s, seconds; R, reading; L, listening; S, speaking.

TOEFL Integrated Speaking Test Rubrics

The TOEFL iBT integrated speaking rubric developed by the Educational Testing Service in 2008 was used by the two raters in scoring L2 learners' performance. The rubric accommodates four criteria: delivery denoted by fluency, clarity, and pronunciation; language use referring to grammatical accuracy and vocabulary use; topic development indicated by cohesion and progression of ideas, and general description (Huang et al., 2018).

Procedures

Data Collection

Chinese EFL learners answered the SCICASA each time they finished one integrated speaking test task. An electronic inventory in the form of word documents was delivered to the learners through a Chinese online survey system named "WenJuanXing",[1] which allowed them to use mobile phones for convenience and for research efficiency. Data collected on the system were automatically saved for our data analysis later. Data collection on the SCICASA for each Chinese EFL learner lasted around 10–20 mins. Chinese EFL learners performed the integrated speaking test tasks on computers with database software packages for TOEFL iBT integrated speaking. The learners' responses to the speaking test tasks were recorded and stored automatically by the software packages as a single file. These files were then named after the learners' codes. The order of those recording files was randomized using a random list generated in Microsoft Excel before they were given to the two raters. All the recording files were backed up in case of data loss (Weir et al., 2006).

By means of analytic scoring before holistic scoring, the two raters first scored independently the four segments of each Chinese EFL learner's responses by referring to the rubric. A score ranging from 0 to 4 points was given to the four segments. Subsequently, the four scores were aggregated to form a composite score for each learner's response to each task. The composite scores from the two raters for each response were then aggregated before they were divided to generate an average score which was used as the holistic score to measure the learners' speaking performance statistically (Huang and Hung, 2013).

Data Analysis

Data preparation yielded 95 valid samples from the initial sampling of 134 participants, and the sample size meets the requirements of statistical testing methods involved in our study, including descriptive analysis, one-way repeated measures MANOVA, ANOVA, and the multiple regression analyses (Pallant, 2016; Frey, 2018). Following some scholars (e.g., Barkaoui et al., 2013; Sun, 2020, 2022a,b), we run a descriptive analysis of the means of L2 learners' use of the four metacognitive strategies across tasks. We then used the line chart generated in Excel *via* the value of the means to illustrate the variance

[1] https://www.wjx.cn/index.aspx

in the Chinese EFL learners' IDs in metacognitive strategy use across tasks. The value of these means in combination with the chart was to address RQ1. Likewise, we used the means of the learners' test scores to represent their speaking performance. To ensure scoring validity, inter-rater reliability was inspected with reference to Cronbach's alpha coefficient. The index was 0.91, larger than the cutoff criterion (>0.70). This suggests the statistical validity of the scores rated by the two raters (Frey, 2018; Sun, 2020).

The subsequent data analysis for answering RQ2 was parsed into three steps. Step one targeted the relationship between Chinese EFL learners' IDs in metacognitive strategy use and task demands involved in the four TOEFL iBT integrated speaking tasks within Kormos' Model. In Step two, we investigated the relationships between the Chinese EFL learners' IDs in metacognitive strategy use and their speaking performance. In Step three, we integrated the results in Step one with those in Step two to answer RQ2. In Step one, one-way repeated measures MANOVA was used, as the variable of IDs in metacognitive strategy use had four individual components and the variable of task demand had four task conditions represented by the four TOEFL iBT tasks. To run the one-way repeated measures MANOVA, a new variable that combined the four individual metacognitive strategies linearly was created to investigate the within-subject variance in the Chinese EFL learners' reported use of the clustering metacognitive strategies across tasks. For identifying variance, values of F ($p < 0.05$) and η^2 were examined, and the rule of thumb-up for these indices was as follows: If η^2 is $\leq$0.01, it suggests a small effect size; a value ranging from 0.01 to 0.06 indicates a moderate effect size, and if η^2 is $\geq$0.14, it indicates a larger effect size. The exact location of the variance in the four individual metacognitive components was further detected *via* the subsequent rounds of ANOVA, which followed the similar above data analysis principle of MANOVA (Pallant, 2016; Frey, 2018).

In Step two, we deployed the statistical procedure of multiple linear regression to assess how the four individual metacognitive strategies clustered to explain the Chinese EFL learners' speaking performance while examining the associations between individual metacognitive strategies and speaking performance. The four subcomponents of the metacognitive strategies were entered into a model simultaneously as the predictor variables, and the Chinese EFL learners' test scores were entered into the model as the outcome variable. Correlation coefficients (r) within the four individual metacognitive strategies were examined first for the appropriateness of the statistical procedure, and for inspecting multicollinearity: When r is $\leq$0.8, the employment of the procedure is suitable. Index regarding model fit was the adjusted R^2, and the rule of thumb for the index is presented as the following:

<0.1: poor fit.
0.11–0.3: modest fit.
0.31–0.5: moderate fit.
>0.5: strong fit.

In addition, as the four strategies were measured on the same units on the SCICASA, the unstandardized coefficients (β) were inspected to investigate the impact of each individual metacognitive strategy on the Chinese EFL learners' speaking performance. The cutoff p-value for β parameters is <0.01, indicating substantive effects of a specific metacognitive strategy on the learners' speaking performance (Pallant, 2016; Frey, 2018). Finally, in Step three, based on the results generated in Step one and Step two, we examined the relationship in Chinese EFL learners' IDs in metacognitive strategy use, task demand, and their speaking performance to answer RQ2.

RESULTS

Chinese English-as-Foreign-Language Learners' Individual Differences in Metacognitive Strategy Use Across Tasks for RQ1

Figure 2 displays the variance manifested in the frequency of the 95 Chinese EFL learners' use of planning, problem-solving, monitoring, and evaluating across the four TOEFL iBT integrated speaking tasks.

Table 2, on the other hand, revealed the descriptive statistics of the frequency by presenting the average means of the four individual metacognitive strategies across the four tasks. It is clear that problem-solving was reported by the Chinese EFL learners as the most frequently used, followed by planning and evaluating, while monitoring was the least frequently used strategy.

The above results addressed **RQ1**: In the context of integrated L2 speaking assessment tasks within Kormos' Model, Chinese EFL speakers' use of the IDs construct of metacognitive strategies displayed variability, and among the four metacognitive strategies under investigation, problem-solving was used by the Chinese EFL speakers the most frequently, which was followed by planning, evaluating, and monitoring.

Chinese English-as-Foreign-Language Learners' Individual Differences in Metacognitive Strategy Use, Task Demand, and Performance for RQ2

Step One: Individual Differences in Metacognitive Strategy Use and Task Demand

In Step one that targeted the relationship between Chinese EFL learners' IDs in metacognitive strategy use and task demand, with reference to the assumption test results for MANOVA, we used the indices of Pillai's trace for the correction test. The more robust Pillai's trace indices pointed out that there was a significant within-subject difference across task demands on the combined dependent variables or the Chinese EFL learners' reported use of the clustering metacognitive strategies: $F(12, 1212) = 12$, $p = 0.01$ (less than the threshold of 0.05), and partial eta squared (η^2) = 0.02. The result demonstrated a significant difference in the synergetic effect of task demands on the clustering metacognitive strategies in the Chinese EFL learners' performance across tasks (Pallant, 2016; Frey, 2018).

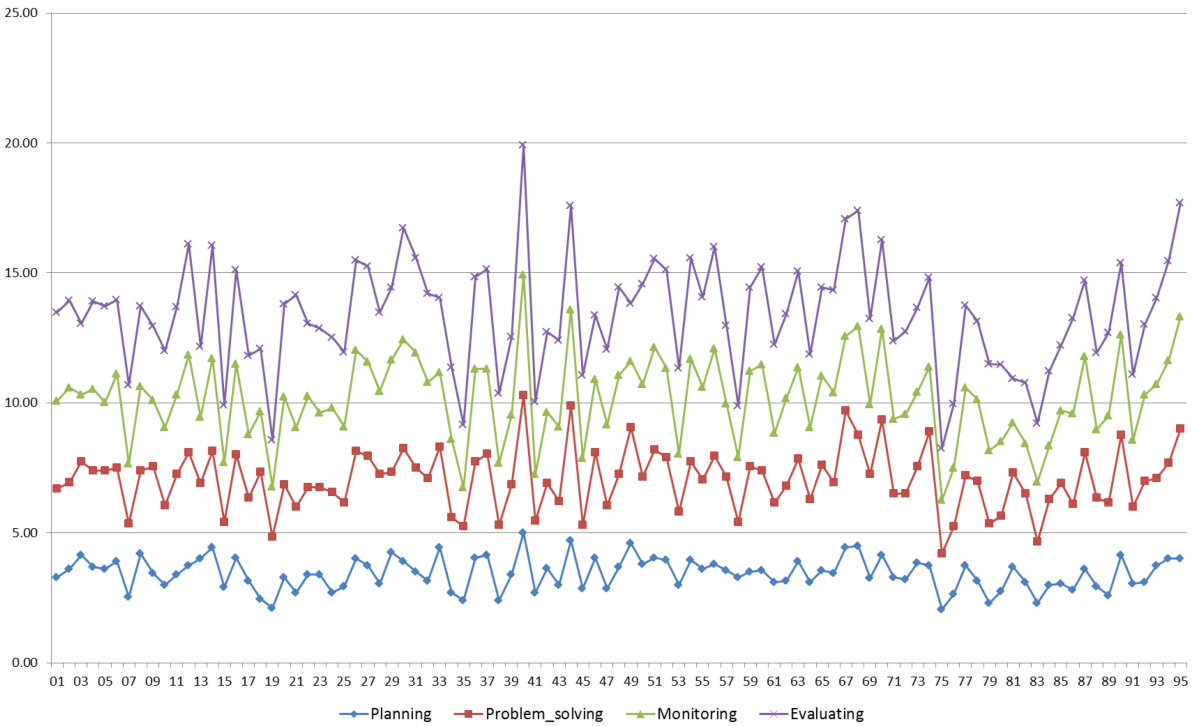


FIGURE 2 | Variance in individual differences (IDs) in metacognitive strategy use among Chinese English-as-foreign-language (EFL) learners.

TABLE 2 | Means of individual metacognitive strategies across tasks.

Tasks	Planning	Problem-solving	Monitoring	Evaluating
Task 1	3.61	3.90	3.17	3.17
Task 2	3.38	3.45	3.21	3.22
Task 3	3.53	3.69	3.18	3.36
Task 4	3.55	3.74	3.30	3.26
Average	3.52	3.70	3.22	3.26

TABLE 3 | Descriptive analysis of oral scores across tasks.

Tasks	Means	SD
Task 1	5.45	2.65
Task 2	4.40	3.15
Task 3	3.51	3.15
Task 4	4.86	2.99

To further locate the diffidence in the four individual metacognitive strategies across tasks, a series of separate ANOVAs were conducted. Each ANOVA was evaluated at an alpha level of 0.25 with Bonferroni adjustment. Results displayed that Chinese EFL learners' reported use of problem-solving demonstrated modest heterogeneity across tasks [F (3, 405) = 3.85, p = 0.01, η^2 = 0.02], whereas substantial variations were not found in the other three individual metacognitive strategies: planning [F(3, 405) = 1. 21, p = 0.38, η^2 = 0.01], monitoring [F(3, 405) = 0.42, p = 0.74, η^2 = 0.003], and evaluating [F(3, 405) = 0.730, p = 0.47, η^2 = 0.01] (Pallant, 2016; Frey, 2018).

Step Two: Individual Differences in Metacognitive Strategy Use and Speaking Performance

Step two focused on the relationships between Chinese EFL learners' IDs in metacognitive strategy use and their speaking performance indicated by their oral scores. As displayed in **Table 3**, with reference to **Table 1**, which illustrates the varying task demands in the four tasks, the means of the Chinese EFL learners' oral scores are as follows: Task 1 (narrating the speakers' opinion on the university's new policy) had the highest value, followed by Task 4 (illustrating a concept on money) and Task 2 (illustrating a concept on audience effect) in contrast to Task 3 (selecting a solution to time conflict), which ranked the lowest.

Results of the subsequent multiple linear regression analysis showed that there were no significant interactive and individual effects of the four metacognitive strategies reported by Chinese EFL learners on their oral scores across tasks.

As shown in **Table 4**, values of the adjusted R^2 on the four tasks were less than 0.1, suggesting a poor model fit. Alternatively stated, the four clustering metacognitive strategies explained a little in the variance of the Chinese EFL learners' oral scores across the tasks. In addition, the p-values of the four tasks were all larger than 0.05, indicating that the four models built on the dataset of the four tasks were not the significant predictors of these learners' speaking performance across tasks. The results implied that no substantial effects of the clustering metacognitive

TABLE 4 | Relationship between the clustering metacognitive strategies and speaking performance across tasks.

Tasks	Adjusted R^2	*df*	*F*	Sig.
Task 1	−0.36	4	0.18	0.95
Task 2	−0.00	4	0.86	0.49
Task 3	0.01	4	1.27	0.29
Task 4	0.01	4	1.19	0.32

TABLE 5 | Relationships between individual metacognitive strategies and speaking performance across tasks.

Tasks	Metacognitive strategies	*β*	*t*	Sig.
Task 1	Planning	0.06	0.22	0.83
	Problem-solving	0.26	0.65	0.52
	Evaluating	0.01	0.13	0.90
	Monitoring	−0.11	−0.27	0.79
Task 2	Planning	0.11	0.67	0.50
	Problem-solving	0.03	0.29	0.77
	Evaluating	0.15	1.11	0.27
	Monitoring	−0.09	−0.53	0.60
Task 3	Planning	−0.02	−0.04	0.97
	Problem-solving	0.95	2.16	0.03
	Evaluating	−0.71	−1.22	0.23
	Monitoring	0.19	0.41	0.68
Task 4	Planning	0.90	1.83	0.070
	Problem-solving	−0.27	−0.55	0.587
	Evaluating	0.30	0.66	0.513
	Monitoring	−0.62	−1.26	0.210

strategies on the Chinese EFL learners' speaking performance across tasks were discovered.

Furthermore, **Table 5** reveals that all the *p*-values of the *β* coefficients for the four subcomponents of the metacognitive strategies on the four test tasks were larger than 0.01. Such results revealed that the four individual metacognitive strategies had no relationships with the Chinese EFL learners' speaking performance across tasks.

Step Three: Individual Differences in Metacognitive Strategy Use, Task Demand, and Performance

By integrating the results from Step one into those from Step two, we answered RQ2: planning, problem-solving, monitoring, and evaluating worked interactively, responding to task demands in the four TOEFL iBT integrated speaking tasks; the four metacognitive strategies, in individual and interactive working modes, had no significant effects on speaking performance.

DISCUSSION

Chinese English-as-Foreign-Language Learners' Individual Differences in Metacognitive Strategy Use Across Tasks

As revealed by the descriptive analysis shown in **Figure 2**, Chinese EFL learners' use of metacognitive strategies differed from one another. This lends empirical evidence to the literature reviewed previously on the concept of IDs, which proposes metacognitive strategy use as one variable accounting for the concept (e.g., Griffiths and Soruç, 2021). Additionally, among the four metacognitive strategies under investigation, problem-solving was reported by Chinese EFL learners as the strategy they used most frequently. Such a result may have to do with how L2 learners performed the integrated speaking testing tasks. According to O'Malley and Chamot (1990), L2 learners tend to use strategies in a problem-solving manner, so it is possible that the Chinese EFL learners in our study considered their use of various strategies as an application of the problem-solving strategy and reported them on the inventory.

Indeed, in line with some scholars (e.g., Flavell, 1979; Zhang, 1999), L2 learners' understanding of problem-solving strategy use reflects their metacognitive knowledge of strategies. As L2 learners' metacognitive knowledge may be fallible or false, it is likely that they believe that they use the problem-solving strategy in performing tasks given, but in fact, they do not use such a metacognitive strategy at all. This may be true with the Chinese EFL learners in our study, which further explains the highest frequency of problem-solving use reported by them (Brown, 1987). The fallibility related to the Chinese EFL learners has been documented by Zhang (1999) whose study revealed the fallibility of Chinese university EFL learners' metacognitive knowledge associated with their reading strategies.

Another possible reason for the highest frequency of the problem-solving strategy use may relate to L2 speech production. As reviewed earlier, in Kormos' Model, unlike the other three metacognitive strategies which either work in a specific stage of the L2 speech process such as planning in conceptualization or work in a covert way during the process such as monitoring, problem-solving operates overtly throughout in L2 speech production, assisting L2 speakers to solve all the possible problems they might encounter in the speaking process. This "throughout" and "overt" characteristic is very likely to result in the highest frequency of problem-solving use reported by Chinese EFL learners in processing L2 speaking tasks.

In contrast, monitoring was reported as the least frequently used metacognitive strategy. This result is possibly due to L2 speech production. In Kormos' Model, monitoring engages in speaking in both covert and overt manners. As the Chinese EFL learners had no prior knowledge of the four metacognitive strategies reported in our initial preparatory survey before the study, it is quite likely that they might not be able to identify monitoring when the strategy was working in a covert manner even though they were using it in the actual task performance. The lowest frequency of monitoring use in the integrated L2 speech assessment has borrowed some support from Swain et al. (2009), Yi (2012), and Barkaoui et al. (2013), the three studies that bear the closest relevance to this research, where metacognitive monitoring was found to be either not used at all or used the least frequently. Since monitoring works in tandem with evaluation as delineated afore, the low frequency of monitoring understandably contributed to the low frequency of evaluation, as was the case with this study.

Finally, the low frequency of planning use may be caused by Chinese EFL learners' lack of prior knowledge of metacognitive strategy use discussed above: When individuals know nothing about how to use metacognitive strategies, it is understandable that they may not have the awareness of using these strategies in performing tasks. Such a relationship between one's knowledge of metacognitive strategies and their use of these strategies has been reported elsewhere (e.g., Fazilatfar, 2010). The lack of motivation is likely to be another cause. According to some scholars (Oxford, 2017; Cohen, 2018), motivation is one of the most important individual factors that affect L2 learners' strategy use, including planning. In this study, because Chinese EFL learners were volunteers and their performance did not affect their credit in the university, they might not be motivated enough to do systematic planning in performing the tasks given as proposed by Kormos and Wilby (2019) that learners' task motivation considerably influences their strategy use.

Chinese English-as-Foreign-Language Learners' Individual Differences in Metacognitive Strategy Use, Task Demand, and Performance

Individual Differences in Metacognitive Strategy Use and Task Demand

In general, the slight statistical variance in Chinese EFL learners' use of the individual metacognitive strategies across tasks illustrates the effect of task demands on these learners' use of individual metacognitive strategies, though not substantial. The result coincides with the finding by Oxford et al. (2004), Swain et al. (2009), Yi (2012), and Barkaoui et al. (2013), in which the types and frequencies of the strategies used by participants were not found to be significantly affected by task demands. The reason for the loose correlation, as Barkaoui et al. (2013) have pointed out, has to do with the integrated L2 speaking tasks: Speaking tasks are highly demanding in terms of strategy use, so it is possible that L2 speakers use whatever metacognitive strategies at hand to tackle the speaking tasks in actuality. As a result, L2 speakers may not purposefully use strategies in response to a specific task demand.

Regarding the synergetic effects of the task demands on metacognitive strategy use, the high correlation index between the four individual metacognitive strategies and the output of the one-way repeated MANOVA suggest that these metacognitive strategies worked in an interactive manner reported by the Chinese EFL learners whose use of these strategies demonstrated substantial variability in response to the changing task demands. The result implies that metacognitive strategies operated in a clustering manner and they were task demand-dependent, which not only corroborates the working principle of IDs variables, including metacognitive strategies, that emphasizes the interactions between these variables with the external contexts or given tasks, but the working mode of metacognitive strategies as illustrated by Kormos' Model and the literature on metacognition and LLSs as reviewed previously. Additionally, the result has been evidenced by a lot of literature on L2 assessment. For instance, Nett et al. (2012) whose longitudinal study with an experience sampling analysis on 70 German students showed that the metacognitive strategies used by the participants worked interactively in test performance. In addition, Fernandez (2018) unfolded the employment of the clusters of the metacognitive strategies reported in International English Language Testing System (IELTS) speaking tests *via* coding participants' discourses. Moreover, the result concurs with Rukthong and Brunfaut's (2020) study where metacognitive strategies were used concurrently by participants in their listening task performance.

Individual Differences in Metacognitive Strategy Use and Speaking Performance

Results of the multiple linear regression analysis indicate that generally, Chinese EFL learners' use of metacognitive strategies, which worked individually and interactively, had no substantial effect on their speaking performance. Such results are not consistent with Kormos' Model in which L2 speakers' metacognitive strategy use is proposed to affect their final oral utterances. The reason for the inconsistency may have to do with the testing condition. Kormos' Model is not formulated for the specific purpose of L2 speech production in testing conditions, while the L2 speaking context in our study is related to testing. It is known that under testing conditions, because of factors such as time limit that may cause learners' anxiety and pressure, learners are unlikely to perform as well as they do under non-testing contexts in terms of strategy use (Huang et al., 2018). Consequently, it is possible that learners' metacognitive strategy use displays no relationship with their speaking test performance, as discovered in our study. In fact, empirical research on the relationships between metacognitive strategy use and test performance, though extensive, has yet been inconclusive (e.g., Phakiti, 2016; Fernandez, 2018). In the current literature on L2 assessment, the weak relationship between L2 learners' metacognitive strategy use and speaking test performance has been discovered in many studies. For example, in examining the relationship between individual metacognitive strategy use and test-takers' integrated speaking test performance, Swain et al. (2009) and Barkaoui et al. (2013) found no significant and positive relations between the two variables. Similarly, Fernandez's (2018) study showed no positive correlation between strategy use and participants' test performance reflected by their test response quality in the IELTS speaking test tasks. These studies additionally lent some support to this study that resulted in a weak relationship between Chinese EFL learners' metacognitive strategy use (either in a clustering manner or in an individual form) and their speaking test performance.

A possible alternative explanation is the instrument or the self-report inventory employed in our study to elicit Chinese EFL learners' metacognitive strategies. Some researchers have pointed out that although self-report inventories have witnessed an extensive application in measuring metacognitive strategies, they may not represent what the participants actually do (e.g., Greene and Azevedo, 2010; Veenman and van Cleef, 2019). In accordance with this view, the metacognitive strategy use reported by the Chinese EFL learners on the SCICASA may not truly reflect their actual use of the strategies, and this may

further affect the result, which we attained only through statistical analysis, on the relationship between the strategy use and the learners' speaking performance.

CONCLUSION AND IMPLICATIONS

Although the results presented by our study were attained from only statistical methods, they will provide empirical evidence for validating Kormos' (2006, 2011) Bilingual Speech Production Model and enrich the literature on the IDs in metacognitive strategy use. Simultaneously, the results will potentially add to the theory regarding the contextualization of Kormos' Model in investigating the process of L2 speaking.

To start with, problem-solving was reported by Chinese EFL learners to be used most frequently in performing the integrated L2 speaking tasks. Such a report validates Kormos' Model where L2 speakers resort to problem-solving to compensate for their incomplete L2 knowledge. Second, although an agreement has been reached on the role of metacognitive strategy use as a variable attributing to IDs, it is still unclear how IDs in metacognitive strategies work in actual language use situations. The salience of problem-solving in L2 speech production founded in our study obviously provides some insights into this research area: Among the various metacognitive strategies that are proposed to manifest IDs, problem-solving plays the most influential role in L2 speech production under authentic language testing conditions. Third, the disagreement between the weak relationship identified in this study concerning metacognitive strategy use and speaking performance within Kormos' Model suggests that to understand IDs in metacognitive strategy use and its relationship with performance in L2 speech production within Kormos' Model, it is necessary to take contexts into account: Are they testing contexts or non-testing contexts? It has been revealed by some studies as discussed earlier that under testing conditions, L2 speaker's metacognitive strategy use may not display the sameness with that under non-testing conditions in which L2 speakers are unlikely to be bothered by test anxiety and pressure caused by testing conditions.

Our study into the concept of L2 speakers' IDs manifested by their metacognitive strategy use and its relationships with task demands and performance has important implications for L2 speaking pedagogy and assessment. First, it indicates that in classroom instruction on metacognitive strategy use to perform L2 speaking tasks, in particular L2 speaking test tasks, teachers can set a special focus on the use of problem-solving since the highest frequency of the strategy has suggested that this strategy is easy for L2 speakers to reach and to use in dealing with L2 speaking test tasks. Such a teaching practice is in light of Oxford (2017), who proposed that EFL teachers' attention be paid to students' cognitive needs based on the students' feedback on strategy use, including metacognitive strategy use. Furthermore, this teaching practice also corroborates Plonsky (2019), who supports a type of metacognitive strategy use instruction in classrooms where students tend to master the target strategies that are narrowed down by their teachers in the most effective way. Second, the relationship between L2 speakers' use of metacognitive strategies and task demands involved in integrated L2 speaking tasks denotes that the holistic integrated L2 speaking tasks can effectively elicit L2 learners' metacognitive strategy use and hence are proposed to be taken into consideration in teachers' syllabus design or task development/design for the purpose of metacognitive strategy instruction on L2 speaking. This is essentially an answer to the call of some scholars (e.g., Frost et al., 2020; Zhang et al., 2021a,b) who advocate the inclusion of integrated skills tasks in classroom instruction for fostering language learners' strategic competence, so that they can achieve learning sustainability. By the same token, regarding test developers, if they aim at testing test-takers' metacognitive strategy use or strategic competence, integrated skill tasks are suggested to be taken into account in the process of test development (Bachman and Palmer, 2010). Finally, the weak relationship between metacognitive strategy use and speaking performance suggests that teachers purposefully create testing conditions in their classroom instruction on metacognitive strategy use by employing L2 speaking assessment/testing tasks. In this way, teachers can teach their students how to use metacognitive strategies effectively and efficiently in tackling testing conditions. Otherwise, if the teachers only teach their students' metacognitive strategy use in non-testing conditions, the students may not know how to use metacognitive strategies in performing testing tasks, a type of tasks they usually take for getting a credit or for being enrolled by an institute (e.g., a university) of a higher level of education to achieve their academic success. As a result, the students may lose motivation in classroom instruction on metacognitive strategy use, which may further lead to the failures of the teachers' pedagogical efforts (Zhang et al., 2021b).

LIMITATIONS AND FURTHER RESEARCH

As delineated earlier, in this study, we only used self-report inventory to collect data. To increase the validity of self-report data, it is postulated that multiple procedures of data collection should be conducted (Creswell and Creswell, 2018). However, due to resource constraints, diverse means were not applied in our study, which may pose a threat to the validity of the research results. Moreover, as we used convenience sampling, the participants had similar backgrounds. Such sampling homogeneity may restrict the generalizability of the research results (Gurven, 2018). These limitations indicate that in future research of relevance diverse methods for data collection are suggested to be administered on heterogeneous sampling for improving the reliability and generalizability of the research findings.

DATA AVAILABILITY STATEMENT

The data are not publicly available due to ethical considerations. The raw data supporting the conclusions of this article will be made available upon request from the corresponding authors.

ETHICS STATEMENT

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

AUTHOR CONTRIBUTIONS

WZ conceived of the initial idea, designed the study, collected and analyzed the data, and drafted the manuscript. MZ and YZ fine-tuned the design, and proofread the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was funded by ETS (Educational Testing Service) based in the United States of America, *via* "TOEFL small research grants for doctoral students" in 2019. This article is also funded by the Philosophy and Social Science Foundation of Shanghai, China (Project Number: 2019BYY009).

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Professor Li Chao from Anhui University of Finance and Economics, China for his constructive supervision on statistical analysis, and to Assistant Professor Anik Nandi from Alliance University, India for his inspiring suggestions on literature review.

REFERENCES

An, Z., and Gan, Z. (2021). "Self-directed technology use for English learning among Chinese university students: contribution of motivational beliefs and strategies to learning achievement," in *Teaching and Researching Chinese EFL/ESL Learners in Higher Education in the 21st Century*, eds Z. Lu, M. Liu, and W. Zhang (Abingdon: Routledge Press).

Bachman, L. F., and Palmer, A. S. (2010). *Language Assessment in Practice: Developing Language Assessments and Justifying Their Use in the Real World*. Oxford: Oxford University Press.

Barkaoui, K., Brooks, L., Swain, M., and Lapkin, S. (2013). Test-takers' strategic behaviours in independent and integrated speaking tasks. *Appl. Linguist.* 34, 304–324. doi: 10.1093/applin/ams046

Brown, A. (1987). "Metacognition, executive control, self-regulation, and other more mysterious mechanisms," in *Metacognition, Motivation, and Understanding*, eds F. E. Weinert and R. H. Kluwe (Hillsdale, NJ: Erlbaum), 65–116.

Bygate, M. (2011). "Teaching and testing speaking," in *The Handbook of Language Teaching*, eds M. H. Long and C. J. Doughty (West Sussex: Wiley), 412–440. doi: 10.1002/9781444315783.ch23

Cohen, A. D. (2014). *Strategies in Learning and Using a Second Language*. London: Routledge. doi: 10.4324/9781315833200

Cohen, A. D. (2018). "Moving from theory to practice: a close look at language learner strategies," in *Language Learning Strategies and Individual Learner Characteristics: Situating Strategy Use in Diverse Contexts*, eds R. L. Oxford and C. M. Amerstorfer (London: Bloomsbury), 31–54.

Creswell, J. W., and Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 5th Edn. Los Angeles, CA: Sage.

De Bot, K. (1992). A bilingual production model: Levelt's 'speaking model' adapted. *Appl. Linguist.* 13, 1–24.

Dörnyei, Z., and Kormos, J. (1998). Problem-solving mechanisms in L2 communication: a psycholinguistic perspective. *Stud. Second Lang. Acquis.* 20, 349–385. doi: 10.1017/S0272263198003039

Ellis, R., Skehan, P., Li, S., Shintani, N., and Lambert, C. (2019). *Task-based Language Teaching: Theory and Practice*. Cambridge: Cambridge University Press. doi: 10.1017/9781108643689

Fazilatfar, A. M. (2010). A study of reading strategies using task-based strategy assessment. *J. Engl. Lang. Teach. Learn.* 53, 20–44.

Fernandez, C. (2018). Behind a spoken performance: test-takers' strategic reactions in a simulated part 3 of the IELTS speaking test. *Lang. Test. Asia* 8, 1–20. doi: 10.1186/s40468-018-0073-4

Flavell, J. H. (1979). Metacognition and cognitive monitoring: a new area of cognitive–developmental inquiry. *Am. Psychol.* 34, 906–911. doi: 10.1037/0003-066X.34.10.906

Frey, B. B. (2018). *The Sage Encyclopaedia of Educational Research, Measurement, and Evaluation*. Los Angeles, CA: Sage. doi: 10.4135/9781506326139

Frost, K., Clothier, J., Huisman, A., and Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: mapping task demands and test takers' use of stimulus content. *Lang. Test.* 37, 133–155. doi: 10.1177/0265532219860750

Gan, Z., Liu, F., and Yang, R. (2020). Student teachers' self-efficacy for instructing self-regulated learning in the classroom. *J. Educ. Teach.* 16, 120–123.

Greene, J. A., and Azevedo, R. (2010). The measurement of learners' self-regulated cognitive and metacognitive processes while using computer-based learning environments. *Educ. Psychol.* 45, 203–209. doi: 10.1080/00461520.2010.515935

Griffiths, C., and Soruç, A. (2021). Individual differences in language learning and teaching: a complex/dynamic/socio-ecological/holisticView. *Engl. Teach. Learn.* 45, 339–353. doi: 10.1007/s42321-021-00085-3

Gurven, M. (2018). Broadening horizons: sample diversity and socioecological theory are essential to the future of psychological science. *Proc. Natl. Acad. Sci. U.S.A.* 115, 11420–11427. doi: 10.1073/pnas.1720433115

Huang, H. D., and Hung, S. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Q.* 47, 244–269. doi: 10.1002/tesq.69

Huang, H. D., Hung, S. A., and Plakans, L. (2018). Topical knowledge in L2 speaking assessment: comparing independent and integrated speaking test tasks. *Lang. Test.* 35, 27–49. doi: 10.1177/0265532216677106

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. Mahwah, NJ: Erlbaum.

Kormos, J. (2011). "Speech production and the cognition hypothesis," in *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance*, ed. P. Robinson (Amsterdam: Benjamins). doi: 10.1075/tblt.2.06ch2

Kormos, J., and Wilby, J. (2019). "Task motivation," in *The Palgrave Handbook of Motivation for Language Learning*, eds M. Lamb, K. Csizér, A. Henry, and S. Ryan (Cham: Palgrave), 267–286. doi: 10.1007/978-3-030-28380-3_13

Levelt, W. J. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511733017

Nazarieh, M. (2016). A brief history of metacognition and principles of metacognitive instruction in learning. *J. Humanit. Arts Med. Sci.* 2, 61–64.

Nett, U. E., Goetz, T., Hall, N. C., and Frenzel, A. C. (2012). Metacognitive strategies and test performance: an experience sampling analysis of students' learning behaviour. *Educ. Res. Int.* 2012:958319. doi: 10.1155/2012/958319

Newton, J. M., and Nation, I. S. P. (2020). *Teaching ESL/EFL Listening and Speaking*, 2nd. Edn. New York, NY: Routledge. doi: 10.4324/9780429203114

O'Malley, J. M., and Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139524490

Oxford, R. L. (1990). *Language Learning Strategies: What Every Teacher Should Know*. Boston, MA: Heinle.

Oxford, R. L. (2017). *Teaching and Researching Language Learning Strategies*. New York, NY: Routledge. doi: 10.4324/9781315719146

Oxford, R. L., and Amerstorfer, C. M. (2018). "The state of the art in language learning strategies and individual learner characteristics," in *Language Learning Strategies and Individual Learner Characteristics: Situating Strategy Use in Diverse Contexts*, eds R. L. Oxford and C. M. Amerstorfer (London: Bloomsbury), 23–33.

Oxford, R. L., Cho, Y., Leung, S., and Hae-Jin, K. (2004). Effect of the presence and difficulty of task on strategy use: an exploratory study. *Int. Rev. Appl. Linguist. Lang. Teach.* 42, 1–47. doi: 10.1515/iral.2004.001

Pallant, J. (2016). *SPSS Survival Manual: A Step by Step Guide to Data Analysis Using IBM SPSS*, 6th Edn. Sydney, NSW: Allen & Unwin.

Phakiti, A. (2016). Test-takers' performance appraisals, appraisal calibration, and cognitive and metacognitive strategy use. *Lang. Assess. Q.* 13, 75–108. doi: 10.1080/15434303.2016.1154555

Plonsky, L. (2019). "Language learning strategy instructing: recent research and future directions," in *Learning Strategy Instruction in the Language Classroom: Issues and Implementation*, eds A. U. Chamot and V. Harris (Bristol: Multilingual Matters), 3–21. doi: 10.21832/9781788892346 15-007

Poulisse, N., and Bongaerts, T. (1994). First language use in second language production. *Appl. Linguist.* 15, 36–57. doi: 10.1093/applin/15.1.36

Psaltou-Joycey, A., and Gavriilidou, Z. (2018). "Language learning strategies in Greek primary and secondary school learners: how individual characteristics affect strategy use," in *Language Learning Strategies and Individual Learner Characteristics: Situating Strategy Use in Diverse Contexts*, eds R. L. Oxford and C. M. Amerstorfer (London: Bloomsbury), 176–195.

Purpura, J. E. (1999). *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeming Approach*. Cambridge: Cambridge University Press.

Rukthong, A., and Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Lang. Test.* 37, 31–53. doi: 10.1177/0265532219871470

Skehan, P. (2016). "Foreign language aptitude, acquisitional sequences, and psycholinguistic processes," in *Cognitive Individual Differences in Second Language Processing and Acquisition*, eds G. Granena, D. Jackson, and Y. Yilmaz (Amsterdam: Benjamins), 7–40. doi: 10.1075/bpa.3.02ske

Skehan, P. (2018). *Second Language Task-Based Performance: Theory, Research, Assessment*. New York, NY: Routledge.

Sperling, R. A., Richmond, A. S., Ramsay, C. M., and Klapp, M. (2012). The measurement and predictive ability of metacognition in middle school learners. *J. Educ. Res.* 105, 1–7. doi: 10.1080/00220671.2010.514690

Sun, P. P. (2016). *Chinese as a Second Language Learners' Speech Competence and Speech Performance in Classroom Contexts: Cognitive, Affective, and Socio-Cultural Perspectives.* Unpublished Ph.D. thesis. Auckland: The University of Auckland.

Sun, P. P. (2020). *Chinese as a Second Language Multilinguals' Speech Competence and Speech Performance: Cognitive, Affective, and Sociocultural Perspectives*. Singapore: Springer Nature.

Sun, P. P. (2022a). Strategic self-regulation for speaking English as a foreign language: scale development and validation. *TESOL Q.*

Sun, P. P. (2022b). The influence of cognitive, affective, and sociocultural individual differences on L2 Chinese speech performance: a multiple indicators multiple causes approach. *J. Psycholinguist. Res.* 1–23. [Epub ahead of print].

Swain, M., Huang, L., Barkaoui, K., Brooks, L., and Lapkin, S. (2009). *The Speaking Section of the TOEFL iBT^{TM} (sstibt): Test-Takers' Reported Strategic Behaviours* (Research Report No. RR-09-02). Princeton, NJ: Educational Testing Service. doi: 10.1002/j.2333-8504.2009.tb02187.x

Teng, F., Qin, C., and Wang, C. (2021a). Validation of metacognitive academic writing strategies and the predictive effects on academic writing performance in a foreign language context. *Metacogn. Learn.* 17, 167–190. doi: 10.1007/s11409-021-09278-4

Teng, F., Wang, C., and Zhang, L. J. (2021b). Assessing self-regulatory writing strategies and their predictive effects on young EFL learners' writing performance. *Assessing Writ.* 51:100573. doi: 10.1016/j.asw.2021.100573

Towell, R., Hawkins, R., and Bazergui, N. (1996). The development of fluency in advanced learners of French. *Appl. Linguist.* 17, 84–119. doi: 10.1093/applin/17.1.84

Veenman, M. V., and van Cleef, D. (2019). Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM Int. J. Math. Educ.* 51, 691–701. doi: 10.1007/s11858-018-1006-5

Wang, R., and Liu, Y. (2018). Literature review of syntactic priming experiment methods and bilingual speech production models. *Adv. Soc. Sci. Educ. Humanit. Res.* 283, 394–398. doi: 10.2991/cesses-18.2018.88

Weir, C. J., O'Sullivan, B., and Horai, T. (2006). *Exploring Difficulty in Speaking Tasks: An Intra-task Perspective. IELTS Research Report No.6*. London: British Council and IDP Australia, 119–160.

Yahya, M. Y. (2019). *Improving Speaking Performance and L2 Motivation Through Task-Based Language Teaching on Malaysian Undergraduate Students.* Unpublished Ph.D. thesis. Reading: The University of Reading.

Yi, J. I. (2012). *Comparing Strategic Processes in the iBT Speaking Test and in the Academic Classroom*. Unpublished Ph.D. thesis. Leicester: The University of Leicester.

Zare, P. (2012). Language learning strategies among EFL/ESL learners: a review of literature. *Int. J. Humanit. Soc. Sci.* 2, 162–169.

Zhang, L. J. (1999). *Metacognition, Cognition and L2 Reading: A Study of Chinese University EFL Readers' Metacognitive Knowledge and Strategy Deployment*. Ph.D. thesis. Nanyang Ave: Nanyang Technological University.

Zhang, L. J., and Zhang, D. (2018). "Metacognition in TESOL: theory and practice," in *The TESOL Encyclopaedia of English Language Teaching*, Vol. II, eds J. I. Liontas and A. Shehadeh (Hoboken, NJ: Wiley-Blackwell), 682–792.

Zhang, W. W. (2021). *Strategic Competence, Task Complexity, and Learner Performance in Computer-Delivered Integrated Speaking Test Tasks: A Study of English-as-a-Foreign-Language (EFL) Learners in China*. Unpublished. Ph.D. Thesis. Auckland: The University of Auckland.

Zhang, W., Zhang, L. J., and Wilson, A. J. (2021a). Supporting learner success: revisiting strategic competence through developing an inventory for computer-assisted speaking assessment. *Front. Psychol.* 12:689581. doi: 10.3389/fpsyg.2021.689581

Zhang, W., Zhang, D., and Zhang, L. J. (2021b). Metacognitive instruction for sustainable learning: learners' perceptions of task difficulty and use of metacognitive strategies in completing integrated speaking tasks. *Sustainability* 13:6275. doi: 10.3390/su13116275

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.

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 17 June 2022
doi: 10.3389/fpsyg.2022.917168

An Exploratory Study of Strategy Use on Elicited Imitation Tasks

Yuyun Lei[1]* *and Xun Yan*[2,3]

[1]Department of East Asian Languages and Cultures, Wake Forest University, Winston-Salem, NC, United States, [2]Department of Linguistics, University of Illinois Urbana-Champaign, Urbana-Champaign, IL, United States, [3]Beckman Institute for Advanced Science and Technology, Urbana, IL, United States

Elicited imitation (EI) has gained popularity with recent interests in the quest for efficient and flexible measures of second language (L2) proficiency. Despite the surge of interests, questions remain as to what specific linguistic knowledge, skills, and strategies EI measures. To contribute to this line of inquiry, this study explored the nature of strategy use and its effect on EI performance to elucidate the constructs of EI. Twenty-four L2 learners and eight native speakers of Chinese completed an EI test of Chinese and a strategy use questionnaire after the test. Qualitative analyses of the questionnaire responses revealed that participants mainly employed five types of strategies, including approach strategies, cognitive strategies, metacognitive strategies, communication strategies, and test-wiseness strategies. While native speakers reported the least number of strategies, higher-proficiency L2 learners reported more strategies than lower-proficiency L2 learners. We further subjected strategy use, along with participant proficiency level, item length, and item complexity level, to linear mixed-effects regression analyses. The results showed that participant proficiency level, item length, and item complexity level explained the largest test score variance; in contrast, strategy use of different types only accounted for a smaller proportion. The total number of cognitive strategies had a significant, positive effect on EI performance whereas the total number of metacognitive strategies had a significant, negative effect. These findings offer some insights into the nature of speech comprehension and production on EI and provide validity evidence for the use of EI as a language proficiency measure.

Keywords: strategy use, elicited imitation, validity, proficiency, Chinese

OPEN ACCESS

Edited by:
Xiaofei Lu,
The Pennsylvania State University, United States

Reviewed by:
Haomin Zhang,
East China Normal University, China
Weiying Huang,
East China University of Technology, China

***Correspondence:**
Yuyun Lei
ylei1academic@hotmail.com

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

Received: *10 April 2022*
Accepted: *09 May 2022*
Published: *17 June 2022*

Citation:
Lei Y and Yan X (2022) An Exploratory Study of Strategy Use on Elicited Imitation Tasks. Front. Psychol. 13:917168. doi: 10.3389/fpsyg.2022.917168

INTRODUCTION

With the COVID-19 global pandemic and ongoing quest for efficient, flexible, and accessible instruments to measure second language (L2) proficiency, traditional tasks that elicit constrained responses but can be easily administered online have regained popularity. One of them is the elicited imitation (EI) task. EI, also known as sentence repetition, requires test takers to listen to a series of stimulus sentences and repeat them verbatim (Underhill, 1987). EI has been frequently used as an instrument to measure participants' proficiency levels in second language acquisition (SLA) research (e.g., Ortega et al., 2002; Gaillard and Tremblay, 2016; Kim et al., 2016). Recently, EI has also appeared in high-stakes, large-scale language

proficiency tests, such as the *Versant* tests by Pearson (Bernstein et al., 2010), the *Duolingo English Test* (Ishikawa et al., 2016), and the newly developed *TOEFL Essentials* test (Davis and Norris, 2021). Despite the resurgence of interests in EI, scholars remain dubious about its construct validity (i.e., what it measures) and question the impact of individual differences on EI performance. Previous studies have addressed this concern by examining the role of learners' memory capacity in EI performance (Okura and Lonsdale, 2012; Kim et al., 2016; Park et al., 2020), finding mostly weak correlations between working memory and EI performance (r ranges from 0.25 to 0.31). In these studies, the strongest predictors of learners' EI performance were shown to be language-related, either their course levels or their performance on another language task. While these studies provide supportive validity evidence for EI as a measure of L2 proficiency, research in this area is still in its infancy to uncover the linguistic constructs measured by EI. According to the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014, p. 15), "Questioning test takers from various groups making up the intended test-taking population about their performance strategies or responses to particular items can yield evidence that enriches the definition of a construct". The use of strategies as part of test taker individual characteristics form an important source of validity evidence for understanding the construct of a task. Inspired by this line of inquiry, this study explored individuals' strategy use and its effect on the language performance of a Chinese EI test. In doing so, we attempt to provide construct validity evidence for EI by elucidating what linguistic knowledge, skills, and strategies EI elicits in speech comprehension and production.

LITERATURE REVIEW

Elicited Imitation as a Measure of Second Language Proficiency

EI has been used as a measure of language proficiency for decades in different research domains. It was first used in first language acquisition research (Fraser et al., 1963; Slobin and Welsh, 1973) and later applied to SLA research (Naiman, 1974; Markman et al., 1975). Despite its popularity among SLA researchers, the exact linguistic knowledge and skills measured by EI tasks have not always been clear in the literature. Depending on research design, EI has been claimed to measure L2 listening (Jensen and Vinther, 2003), L2 grammatical knowledge (e.g., Ellis, 2005; Bowles, 2011), L2 lexical development (West, 2012), L2 pronunciation (Trofimovich and Baker, 2007; Yoon, 2010), and L2 oral proficiency (e.g., Ortega et al., 2002). This is largely because EI allows researchers and test developers to construct sentences flexibly to target specific linguistic elements. That said, there is a general consensus among SLA researchers that EI is a measure of general L2 proficiency or L2 oral proficiency. Because EI is dependent on oral production and presented under time pressure, many SLA researchers have further argued that EI is a measure of implicit grammatical knowledge (Ellis, 2005; Erlam, 2006; Bowles, 2011; Spada et al., 2015). Empirical evidence from factor analysis also tends to group EI with other types of tasks that are commonly used to measure implicit knowledge (e.g., timed grammaticality judgment tasks). However, this argument has not been accepted by all scholars. For example, Suzuki and Dekeyser (2015) recently argued that EI does not measure implicit knowledge but instead measures automatized explicit knowledge. Thus, whether EI is a measure of implicit knowledge still requires further investigation. Nevertheless, EI, as a language proficiency measure, has been widely accepted and used in SLA research since the 1970s (Naiman, 1974). Thus far, EI tests have been developed and validated as a measure of L2 proficiency in Spanish (Bowden, 2016), German, Japanese, English (Ortega et al., 2002), French (Tracy-Ventura et al., 2014; Gaillard and Tremblay, 2016), Korean (Kim et al., 2016), Mandarin Chinese (Wu and Ortega, 2013), Russian (Mozgalina, 2015), and Vietnamese (Chaudron et al., 2005).

The theoretical rationale behind EI as an L2 proficiency measure is that in order to repeat a sentence correctly, one has to understand the meaning of the sentence. Since a sentence exceeding the capacity of short-term memory would be difficult to imitate without actual comprehension, a speaker has to decode the sentence using their linguistic knowledge from long-term memory and then reproduce it (Bley-Vroman and Chaudron, 1994). As such, an EI test typically consists of a number of stimulus sentences, ranging in length (measured in syllables, morphemes, or words) and containing a variety of target features (e.g., specific syntactic structures). For L2 learners, when the stimulus sentence goes beyond their current level of the target language, such as containing unfamiliar vocabulary or presenting new grammatical structures, they are likely to repeat only part of the sentence (the part that they can understand) correctly at best or, at worst, fail to repeat the entire sentence. Therefore, EI responses can reveal L2 learners' strengths and weaknesses in linguistic knowledge and skills, which can facilitate teaching, learning, and other test score uses (Yan et al., 2020). Meta-analysis studies demonstrate that EI, as a general L2 proficiency measure, has a strong ability to discriminate speakers across proficiency levels (Yan et al., 2016; Kostromitina and Plonsky, 2021) and higher reliability compared to other speaking tasks (Henning, 1983).

Although EI as a useful L2 proficiency measure has been accepted by many SLA researchers, there is another concern about its construct validity, that is, whether EI measures language comprehension and production or elicits rote memorization of sounds (Vinther, 2002). Early research observed that when EI stimuli were short enough to be retained as an acoustic representation in short-term memory, it was possible for someone without the knowledge of the target language to "parrot" the stimuli (Fraser et al., 1963; Prutting et al., 1975). This led to the criticism of EI that it measures "perceptual-motor skill" rather than language ability (Fraser et al., 1963, p. 483). To provide evidence for the validity argument of EI as an L2 proficiency measure, a large body of research has focused on establishing the close relationship between L2 learners' EI performance and their performances

on other commonly accepted tests of language proficiency. For instance, Ortega et al. (2002) compared L2 learners' EI performance with their scores on simulated Oral Proficiency Interview (OPI) and the Test of English as a Foreign Language (TOEFL). Participants' EI scores showed moderate to strong correlations with the OPI and TOEFL scores ($r = 0.49$ with the TOEFL scores, $r = 0.61$ with the OPI ratings in Japanese, and $r = 0.88$ with the OPI ratings in Spanish). Erlam (2006) found that L2 learners' EI scores were highly correlated with their scores on subcomponents of the International English Language Testing System (IELTS; $r = 0.67$ with the IELTS speaking score and $r = 0.72$ with the IELTS listening score). Kim et al. (2016) observed similar high correlations between EI scores and the Test of Proficiency in Korean (TOPIK) scores ($r = 0.62$ with the listening score, and $r = 0.77$ with the speaking score). All these results demonstrated that EI, like other established proficiency tests, measures language proficiency of L2 learners.

Another line of research approaches the construct validity of EI by directly investigating the relationship between working memory capacity and EI performance. Okura and Lonsdale (2012) administered an English EI test and a non-word repetition test to sixty-seven students learning English as a Second Language (ESL). They found that EI performances by these participants were significantly correlated with their curricular levels ($r = 0.79$, $p < 0.001$), but not with their scores on the non-word repetition test ($r = 0.25$, $p = 0.12$), which was considered as an index of their working memory capacity. Kim et al. (2016) examined how phonological short-term memory capacity was related to EI performances by sixty-six L2 learners of Korean. A similarly weak correlation was observed between EI scores and phonological short-term memory capacity measured by the digit span test ($r = 0.30$, $p > 0.01$). Park et al. (2020) conducted the same Spanish EI test in Ortega et al. (2002) and a non-word repetition test on seventy-eight L2 learners of Spanish. L2 learners' EI performances were predicted by their performances on an oral narrative task more than their scores on the non-word repetition task, but they found memory capacity may have a facilitative effect for beginning L2 learners. While these studies demonstrated that EI measures a construct that is different from the one measured by memory tests, the question remains as to what specific linguistic constructs EI measures, that is, what linguistic knowledge, skills, and strategies EI elicits in speech comprehension and production.

Test-Taking Strategies as Evidence for Construct Validity

In understanding the constructs of a task, one can not only look at "which responses are considered correct" (i.e., product) but also "what process underlies them" (i.e., process; Alderson, 2000, p. 97). As reviewed above, previous research tends to only focus on the products of EI by examining relationships between scores on EI and scores on other tests that measure either similar or distinct constructs. As Cohen (2006) pointed out,

> "*what was missing was the aspect of test validation that related to respondents' behaviours in taking the tests: little was known about what they were actually doing to produce answers to questions and how it corresponds to the abilities one sought to test*" (p. 89).

Analyzing performance strategies or response processes that test takers engage in on test tasks provides another important source of construct validity evidence in validation studies (AERA, APA, and NCME, 2014). More specifically, analyses of test-taking processes or strategies can offer evidence concerning the fit between the elicited processes in the actual performance and the theorized processes tapped by the construct. If the expected processes are elicited, the test is thought to be valid. If alternative processes, irrelevant to the construct, are observed, the validity of the test warrants questions. Following this line of validation research, this study examined test-taking strategies that participants employed on EI tasks to provide construct validity evidence for EI.

Test-taking strategies, as defined by Cohen (2006) are the "consciously selected processes that the respondents used for dealing with both the language issues and the item response demands in the test-taking tasks at hand" (p. 89). There are three types of test-taking strategies—language learner strategies, test management strategies, and test-wiseness strategies. Language learner strategies assist test takers in "operationalizing the targeted language skills" for a task (Cohen, 2013, p. 3). For instance, employing inferencing strategies would be helpful for test takers to respond to some listening comprehension items. Test management strategies allow test takers to respond "meaningfully to the test items and tasks", such as outlining a plan before speaking. These two types of strategies are expected operations and procedures for task completion and are construct-relevant for test tasks. In contrast, test-wiseness strategies enable test takers to use "knowledge of test formats and other peripheral information to answer test items without going through the expected linguistic and cognitive processes" (Cohen, 2006, p. 90) and thus are considered construct-irrelevant. The degree to which construct-relevant and -irrelevant strategies can be used by test takers determines the validity of a test (Cohen, 2013). Regarding the specific functions of the strategies, test-taking strategies can be mainly classified into approach strategies, cognitive strategies, metacognitive strategies, and communication strategies. According to Swain et al. (2009), approach strategies orient test takers to the task; cognitive strategies involve manipulating the target language to understand and produce language; metacognitive strategies involve organizing, planning, and evaluating language performance; and communication strategies are strategies used for solving a linguistic problem in order to reach a communicative goal.

Test-taking processes or strategies are often obtained by the use of verbal reports in strategy use research. Common verbal reports include think-aloud (e.g., Vandergrift, 1997), stimulated recalls (e.g., Swain et al., 2009), and self-report interviews and questionnaires (e.g., Oxford and Ehrman, 1995). There are some concerns about the veridicality of the verbal report data, that is, whether the data actually reflects participants' thought

processes during task completion. However, researchers have suggested that this threat can be minimized if there is only a short delay between task performance and self-report (Bowles, 2011). In addition, while this type of data may not be exhaustive, it offers a window into the cognitive processes of how test takers arrived at their performance, providing valuable information that cannot be easily addressed by other methods (Gass and Mackey, 2013). More advanced technologies, such as eye-tracking techniques (e.g., Storey, 1997) and event-related brain potentials (e.g., Van Hell and Tokowicz, 2010), have been used in recent strategy use research. Nevertheless, verbal reports remain to be the primary research tool as it is less intrusive and can be effectively conducted.

There has been an increase in recent years in the number of studies investigating test-taking strategies to provide new sources of evidence for construct validity of language tests. For example, Storey (1997) obtained think-aloud data from twenty-five female Chinese students when they were completing an English discourse cloze test. The analyses revealed that participants employed strategies to analyze the rhetorical structure of the text, which supported the argument that cloze tests involve the discourse processing ability, providing validity evidence for the use of cloze tests as a measure of general language ability. Utilizing a strategy inventory, Yang and Plakans (2012) investigated ESL learners' strategy use and its relationship to test performance on an integrated reading–listening–writing test task. They found that the task requires not only text comprehension and production abilities, but also regulation skills to coordinate reading, listening, and writing materials. These strategies conformed to the strategies proposed in the literature on integrated writing, therefore supporting the valid use of integrated reading–listening–writing tests for assessing academic writing ability. Brunfaut and McCray (2015) used both eye-tracking and stimulated recalls from ESL learners taking the reading proportion of the Aptis test developed by the British Council. Different patterns were observed in the processes used to complete the different reading tasks, including the common lower-level (e.g., lexical access and syntactic parsing) and higher-level (e.g., inferencing and creating paragraph-level representations) reading processes. They also found that participants employed test-wiseness strategies (e.g., the reliance on guessing and background knowledge) to complete the tasks, but the strategies were only used to a limited extent. They concluded that the Aptis reading component adequately taps into the construct of reading skills and thus was a valid reading test.

Some test-taking strategy research involved test takers at different proficiency levels and examined the relationship between strategy use and test performance. Test takers at different proficiency levels have been found to utilize strategy differently (e.g., Green and Oxford, 1995; Oxford and Burry-Stock, 1995; Bruen, 2001). These differences are argued as the result of different proficiency levels and would contribute to differential test performance (e.g., Gordon, 1987; Chamot et al., 1988). A positive relationship has been reported between proficiency level and the use of certain types of strategies, such as metacognitive strategies (e.g., Purpura, 1999; Phakiti, 2003), cognitive strategies (e.g., Oxford and Ehrman, 1995; Park, 1997), and compensation strategies (e.g., Dreyer and Oxford, 1996; Nakatani, 2006). A number of studies also observed that within the same category of strategy, some strategies had positive effects on test performance, while others had negative effects. For example, Song (2005) found that the cognitive strategy of linking with prior knowledge contributed positively to the prediction of the listening and writing scores of the Michigan English Language Assessment Battery (MELAB), while the cognitive strategy of repeating/confirming information showed a negative impact on the MELAB scores. The effect of particular strategy use on test performance may be dependent on task types and contexts.

Compared to the test-taking strategy studies on listening, reading, and writing assessments, the research on strategy use in the assessment of speaking is limited. Yoshida-Morise (1998) identified the communication strategies twelve ESL learners used in their Oral Proficiency Interviews. She found that overall lower-proficiency learners used a higher number of strategies than higher-proficiency learners to compensate for their insufficient L2 knowledge. However, she also observed that higher-proficiency learners used certain communication strategies more than lower-proficiency learners, such as restructuring and repair strategies. In response to the debate over the use of independent and integrated speaking tasks for the assessment of oral proficiency, Swain et al. (2009) and Barkaoui et al. (2013) examined the reported use of strategies based on stimulated recalls from 30 Chinese-speaking engineering students after performing on the independent and integrated speaking tasks of the TOEFL iBT. They found that the integrated tasks involving more language skills elicited a wider variety of reported strategy use than did the independent tasks. Although the total number of reported strategies had no relationship with the total scores on the TOEFL iBT speaking tasks, they argued for the inclusion of both integrated and independent speaking tasks in the assessment of oral proficiency as they elicited different strategy use that tapped into distinct constructs of communicative performance. Similarly, Huang (2013) explored strategies that 40 test takers used when responding to three tasks on the IELTS speaking test, including answering questions, speaking about a topic, and holding a discussion with an examiner. Analysis of both reported strategies elicited from stimulated recalls and observed strategies in production data showed that participants shared some similarities and differences in strategy use across the three tasks. However, there were no significant differences in the reported strategy use between intermediate-level and advanced-level learners. Fernandez (2018) qualitatively analyzed 12 ESL students' stimulated recalls of the completion of the discussion task on a simulated IELTS speaking test. She found that participants used a great number of metacognitive and cognitive strategies and argued that these strategies are integral to speaking performance. In addition, she observed that some strategies (e.g., analyzing input, planning, and elaborating) contributed positively to the quality of test takers' responses, while some strategies (e.g., linking to previous knowledge/experience and slowing down) negatively impacted test takers' performances. Huang (2016) used a self-designed

strategy inventory to investigate test-taking strategies used by 215 Taiwanese ESL learners on six speaking tasks for the Test of English for International Communication Speaking Test (TOEIC-S). The tasks included text-reading, picture description, integrated read-to-speak tasks, and independent speaking tasks. Using exploratory factor analysis, he identified three major types of test-taking strategies, that is, communication, cognitive, and affective strategies. Among them, the use of communication and cognitive strategies contributed positively to the TOEIC-S performance. These strategies corresponded to those commonly included in the strategies for real-life oral communication; therefore, they argued that the TOEIC-S test assessed oral communication skills in daily life as intended. While these studies focus on open-ended speaking tasks, little is known about the nature of strategy use on EI tasks and how it affects EI performance. The study thus aims to address the following research questions:

1 What is the nature of strategy use employed by L2 learners of Chinese across proficiency levels on a Chinese EI test?
2 How does strategy use affect language performance on the Chinese EI test by L2 learners of Chinese across proficiency levels?

MATERIALS AND METHODS

Participants

A total of 24 L2 learners of Chinese were recruited for the study. Except that one learner was a native speaker of Japanese, all the participants were native speakers of English (two reported additional native languages, including Spanish and Japanese). They were enrolled in the Chinese language program at a U.S. university: ten in beginning-level classes, seven in intermediate-level classes, and seven in advanced-level classes. Based on the ACTFL (2012) Rating Scale, beginning-level students were equivalent to the levels of Novice Mid to Novice High, intermediate-level students were at Intermediate Low to Intermediate Mid, and advanced-level were at Intermediate High to Advanced Low. There were 13 females and 11 males, with an average age of 20.83. Eight native speakers of Chinese (4 females and 4 males, mean age = 25.5) also participated in this study to provide a baseline measure. They were at the Superior level on the ACTFL Rating Scale.

Instruments

All the participants completed an EI test of Chinese, a strategy use questionnaire, and a language background questionnaire. The EI test was designed based on a corpus analysis of the widely used Chinese textbooks, the primary source of the language input and use for L2 learners of Chinese in U.S. universities. According to a survey conducted to over 200 universities by the Chinese Language Teachers Association of the U.S. (Li et al., 2014), we compiled a corpus of 36 widely used Chinese textbooks, amounting to a total of 688 unit texts and 703, 995 characters. The corpus was divided into three sub-corpora, corresponding to the three proficiency levels of a typical university-level Chinese program, that is, beginning, intermediate, and advanced. Using corpus analysis techniques, we identified a list of commonly shared vocabulary and grammar (i.e., occurs the most frequently) across the textbooks at each of the three proficiency levels. Based on this list, we designed three sets of EI sentences, each set targeting one of the three proficiency levels. In addition, to reflect the language that L2 learners of Chinese are exposed to, the actual average length of the sentences at each level in the textbook corpus were adopted in the EI item design, which are 8 syllables (beginning and intermediate levels) and 12 syllables (advanced-level). A total of 72 sentences were developed, evenly distributed across the three lexico-grammatical complexity levels at the two lengths bands. Five Chinese language teachers checked the sentences and ensured their naturalness. Following the common practice of administering EI as a proficiency test (Ortega et al., 2002), a 2-s silence was inserted between the end of each sentence and a 0.5-s ringtone prompting the start of the repetition. The EI test went through an iterative process of development and validation and demonstrated good psychometric qualities (Yan et al., 2020).[1] Sample EI sentences of each level are provided below.

这件衣服非常便宜 (Beginning-level, 8 syllables).

This piece of clothing is very cheap.

我每天都坐公共汽车去学校 (Beginning-level, 12 syllables).

I take the bus to school every day.

他比以前进步多了 (Intermediate-level, 8 syllables).

He made much more progress than before.

他们的房间里挂着一张地图 (Intermediate-level, 12 syllables).

There is a map hanging in their room.

他从事于研究工作 (Advanced-level, 8 syllables).

He does research for his job.

他不但不高兴，反而有点生气 (Advanced-level, 12 syllables).

He is not happy; on the contrary, he is a little angry.

A questionnaire was used in this study to elicit processes and strategies that participants employed for the completion

[1]The full test is available on IRIS (https://www.iris-database.org/iris/app/home/detail?id=york:938753), including test items, test instructions, item recordings, and the grading rubric.

of the EI test. As this is an exploratory study, open-ended questions were designed to elicit participants' thoughts on the processes and strategies when they were completing the EI test. Although stimulated recall is frequently used for such purposes (Gass and Mackey, 2013), it could not conveniently be implemented when this study was conducted due to the physical and technical constraints posed by the COVID-19 pandemic. Cohen (2014) suggested that questionnaires can also gain insights into test takers' strategy use when the questionnaires include information about test takers' thoughts about their behaviors on test tasks. The main questions probing into test-taking strategies in the questionnaire were:

1 What did you think was/were the key(s) to completing this task?
2 How did you complete this task? Please elaborate your thinking and speaking processes.
3 What do you think makes this task difficult for you? And what did you do to deal with these difficulties?
4 What strategies did you use to help you complete the task? Please provide as many as you can.

The language background questionnaire gathered basic information about participant demographics, language background, and language learning experience. All the tasks were delivered online *via* the Gorilla Experiment Builder (Anwyl-Irvine et al., 2019) to accommodate the need of online testing due to the COVID-19 pandemic.

Procedure

Each participant received a website link after they signed up for the study. They completed the tasks at their convenience. During the EI test, each participant first listened to a stimulus sentence, waited for two seconds, and then started to repeat the sentence after hearing a ringtone. They were instructed to repeat as much as possible of what they heard. They had 20 s to repeat. After 20 s, a new sentence was automatically played. Each sentence was played only once. Immediately after taking the EI test, participants responded to the strategy use questionnaire to reflect on the processes and strategies they employed when completing the test. In the end, they filled out the language background questionnaire. The total time lasted about 20 to 30 min for each participant.

Grading and Coding

Two native speakers of Chinese rated all the EI responses using a five-point rating scale (see below, adapted from Ortega et al., 2002). The inter-rater reliability was high, with a value of 0.85 on Cohen's Kappa. Any disagreements between the two raters were resolved through discussion.

4—Exact repetition or synonymous substitutions;
3—Minor deviation with more than half of the sentence repeated;
2—Half repetition;
1—Inadequate repetition with less than half of the sentence repeated;
0—Silence, unintelligible words, minimal repetition.

Descriptive statistics of the participants' test scores on the EI test are shown in **Table 1**. The total test score was 288. Participants performed as expected: native speakers had almost perfect scores; L2 learners from higher curricular levels scored better than learners from lower curricular levels. The level differences were statistically significant [$F(3, 28) = 28.68$, $p < 0.001$, $\eta 2 = 0.75$], indicating that the EI test used in this study was effective at discriminating participants across proficiency levels.

Participants' responses to the strategy use questionnaire were qualitatively analyzed to identify strategies participants employed when completing the EI test. We first compiled a list of strategies reported in the questionnaire responses. Then the strategies were tallied for each participant by the two raters, with 1 representing the use of a particular strategy and 0 being the absent use of such a strategy. Adopting taxonomies of strategies in Swain et al. (2009) and Vandergrift (1997), the reported strategies were classified into five main types of strategies: approach, cognitive, metacognitive, communication, and test-wiseness strategies. The exact agreement between the two coders was 95%. Disagreements were also resolved through discussion.

Statistical Analyses

Descriptive statistics were calculated to examine possible trends of strategy use across proficiency levels. Linear mixed-effects regression analyses were performed to investigate the effects of the use of different strategies on EI performance. The *lm4* package (Bates et al., 2015) in R (version 3.6.1; R Core Team, 2019) was used. The dependent variable was EI score on each item (2,304 data points = 32 participants * 72 items). Items and participants were treated as random intercepts; participant proficiency level, item length, item complexity level (i.e., targeting beginning, intermediate, or advanced lexico-grammar), and the total number of each type of strategies were treated as fixed effects. The sample size exceeded the minimum recommendation for properly powered mixed-effects models by Brysbaert and Stevens (2018).

RESULTS

Strategy Use Across Proficiency Levels

The first research question concerns what strategies were reported using on the EI test by the participants and how these strategies varied across proficiency levels. **Table 2** summarizes the frequencies of different types of strategies employed on the EI test. **Figure 1** represents the mean frequencies of strategy use across proficiency levels. We found a total of 122 instances of strategy use in participants' questionnaire responses. The strategies were classified into five main categories of strategies, including approach strategies, cognitive strategies, metacognitive strategies, communication strategies, and test-wiseness strategies. Approach strategies set the goals for participants to complete the EI test. Cognitive strategies help participants better understand and repeat the stimulus sentences. Metacognitive strategies direct participants' attention to aid the completion of the test. Communication strategies involve conscious plans to deal with linguistic breakdowns. Test-wiseness

TABLE 1 | Descriptive statistics of EI test scores.

Participant level	N	Mean	Median	SD	Lower 95% CI	Upper 95% CI
Beginning	10	116.90	125	46.72	83.49	150.31
Intermediate	7	170.29	161	60.56	114.28	226.30
Advanced	7	229.29	221	27.87	203.52	255.06
Native	8	287.75	288	0.71	287.16	288.34

TABLE 2 | Frequencies of all strategies reported by the participants.

Strategy type	Participant level Raw Frequency (Mean)				Total
	Beginning (*n* = 10)	Intermediate (*n* = 7)	Advanced (*n* = 7)	Native (*n* = 8)	
Approach	12 (1.20)	7 (1.00)	9 (1.29)	11 (1.38)	39
Cognitive	19 (1.90)	13 (1.86)	17 (2.43)	10 (1.25)	59
Metacognitive	3 (0.30)	2 (0.29)	0 (0.00)	0 (0.00)	5
Communication	3 (0.30)	3 (0.43)	3 (0.43)	0 (0.00)	9
Test-wiseness	1 (0.10)	3 (0.43)	6 (0.86)	0 (0.00)	10
Total	38 (3.80)	28 (4.00)	35 (5.00)	21 (2.63)	122

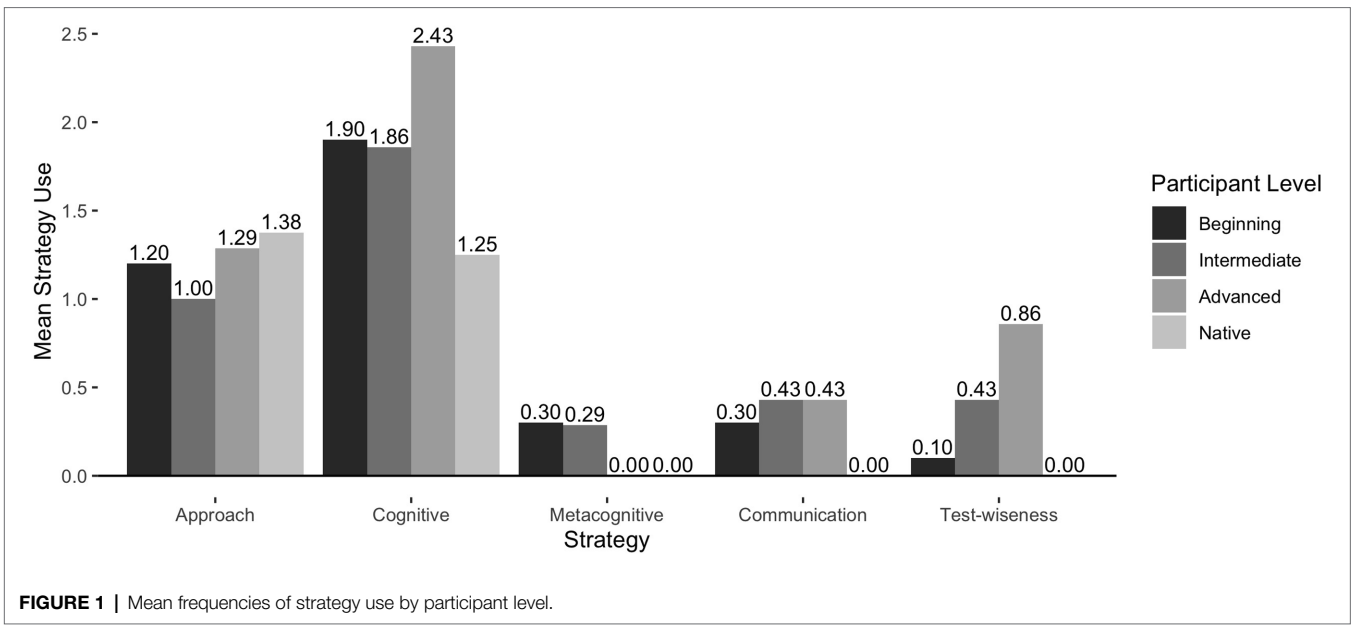


FIGURE 1 | Mean frequencies of strategy use by participant level.

strategies are strategies of completing the test without understanding the stimulus sentences. Overall, participants reported using cognitive strategies the most frequently (n = 59), followed by approach strategies (n = 39), test-wiseness strategies (n = 10), communication strategies (n = 9), and metacognitive strategies (n = 5) being the least frequent. Among the participants, native speakers reported using the least number of all strategies. As some native speakers mentioned that the process of completing the EI test was very easy for them—listening, memorizing, and repeating, it was not necessary for them to intentionally use strategies to complete the test. Native speakers only used approach strategies to set the goals for completing the test and cognitive strategies to help process and retain the information in the stimulus sentences. In contrast, L2 learners used a variety of strategies to complete the EI test. In addition to approach and cognitive strategies, they used metacognitive, communication, and test-wiseness strategies to deal with the linguistic challenges presented by the EI test. Higher-proficiency L2 learners on average employed a greater number of strategies than did lower-proficiency learners. As shown in **Figure 1**, there were positive trends of using approach, cognitive, communication, and test-wiseness strategies as the level of proficiency advances among L2 participants, while a negative trend was observed for the metacognitive strategies. In other

words, higher-proficiency L2 learners used more approach, cognitive, communication, and test-wiseness strategies than did lower-proficiency learners, but less metacognitive strategies.

Table 3 presents the frequencies of subcategories of each type of strategy. There were 15 different subcategories of strategies reported by the participants. In the category of approach strategies, the participants employed three different strategies, that is, *getting as many words repeated as possible*, *getting the main message of the sentence out*, and *controlling the accuracy of the language*. Participants at lower-proficiency levels tended to focus more on repeating as many words as possible, while participants at higher-proficiency levels largely focused on the main message of the sentence. One beginning-level participant commented in the questionnaire that she thought the key to completing the test was to comprehend the main message of the sentence, but in actuality, she tried to repeat as many words as possible. This might have been caused by the fact that she could not understand the sentences, so she resorted to the words that she knew. In addition, more higher-proficiency participants tried to control the accuracy of the language they produced. For instance, one advanced-level participant commented below that he refrained from making up words that he did not know and only repeated the words he was familiar with.

> *For more challenging sentences, I tried to pick up enough familiar words from what the speaker was saying so that I could put together a complete sentence. I tried not to make up any words or assume that I knew what the speaker was saying - instead, for the majority of my responses, I tried to limit what I repeated to the words that I understood* (s19, advanced-level).

As EI involves auditory processing of the stimulus sentences, participants reported using a great number of cognitive strategies. Among the cognitive strategies, *comprehending the meaning of the sentence* was used the most frequently, followed by *recognizing familiar words, phrases, or structures*. Similar to the orientation toward words versus the main message as suggested by the use of approach strategies, lower-proficiency learners reported more strategies on *recognizing familiar words, phrases, or structures*, while higher-proficiency learners focused more on *comprehending the meaning of the sentence*. None of the native speakers employed the strategy of *recognizing familiar words, phrases, or structures*. In addition, more lower-proficiency learners reported using the strategy of *listening for key words, phrases, or structures* to process and understand the information in the sentences. Below are four examples, one from each proficiency level, where participants indicated that they paid attention to key words, phrases, or structures in the sentences while listening. As shown in the examples, higher-proficiency participants were able to focus on larger linguistic units, such as sentence structures, as opposed to lower-proficiency participants who tended to listen for key words or phrases.

> *I tried to…listen for phrases I knew, such as the order of subject, time, to whom, and then verb* (s04, beginning-level).

> *Listen to the person's pauses and when they emphasize certain words and phrases* (s06, intermediate-level).

> *I focused on thinking about the sentence grammatical structures* (s19, advanced-level).

TABLE 3 | Frequencies of individual strategies reported by the participants.

Individual strategy	Participant Level Raw Frequency (Mean)				Total
	Beginning (n = 10)	Intermediate (n = 7)	Advanced (n = 7)	Native (n = 8)	
Approach strategies					
1. Getting as many words repeated as possible	7 (0.70)	3 (0.43)	2 (0.29)	3 (0.38)	15
2. Getting the main message of the sentence out	4 (0.40)	3 (0.43)	5 (0.71)	6 (0.75)	18
3. Controlling the accuracy of the language	1 (0.10)	1 (0.14)	2 (0.29)	2 (0.25)	6
Cognitive strategies					
4. Comprehending the meaning of the sentence	6 (0.60)	3 (0.43)	6 (0.86)	5 (0.63)	20
5. Recognizing familiar words, phrases, or structures	5 (0.50)	4 (0.57)	4 (0.57)	0 (0.00)	13
6. Listening for key words, phrases, or structures	2 (0.20)	3 (0.43)	1 (0.14)	1 (0.13)	7
7. Translating	2 (0.20)	1 (0.14)	3 (0.43)	0 (0.00)	6
8. Rehearsing the sentence before repetition	1 (0.10)	2 (0.29)	2 (0.29)	1 (0.13)	6
9. Connecting to daily scenarios	0 (0.00)	0 (0.00)	1 (0.14)	3 (0.38)	4
10. Chunking the sentence into smaller parts	3 (0.30)	0 (0.00)	0 (0.00)	0 (0.00)	3
Metacognitive strategies					
11. Prioritizing certain parts of the sentence	3 (0.30)	2 (0.29)	0 (0.00)	0 (0.00)	5
Communication strategies					
12. Paraphrasing	1 (0.10)	1 (0.14)	2 (0.29)	0 (0.00)	4
13. Using lexical fillers	1 (0.10)	1 (0.14)	1 (0.14)	0 (0.00)	3
14. Guessing	1 (0.10)	1 (0.14)	0 (0.00)	0 (0.00)	2
Test-wiseness strategies					
15. Imitating the sounds	1 (0.10)	3 (0.43)	6 (0.86)	0 (0.00)	10

I memorized the sentence structure and main verbs (c07, native).

To retain the meaning of the EI sentences, the *translating* strategy was also utilized by the L2 learners. They translated what they heard into English and then translated it back into Chinese when repeating. Advanced-level learners utilized this strategy more than did the other two levels of learners, as advanced-level learners were more likely to understand the meaning of the sentences. Some participants also used the strategy *rehearsing the sentence* before the start of the repetition to help them remember the information. Since rehearsing without comprehension would be difficult, this strategy was used more frequently by higher-proficiency participants. As they were able to comprehend the meaning of the sentences, higher-proficiency participants also used the strategy *connecting to daily scenarios* to help retain the information (see response examples below). For some beginning-level learners, the sentences in the test might have been too long for them to process. Therefore, they reported using the strategy *chunking the sentence into smaller parts* to help themselves understand and memorize the parts they could understand.

I tried to imagine someone saying that to me in real life and tried to listen to the whole sentence to understand it rather than memorizing it (s24, advanced-level).

Sometimes I constructed some scenarios based on what I heard. For example, 这家饭馆的菜不如那家的好 *(The food in this restaurant is not as good as the other restaurant). It feels like a sentence you would say when you go out to eat with your friends* (c03, native).

As regards metacognitive strategies, we found that participants employed one strategy *prioritizing certain parts of the sentence* to complete the test. Only beginning- and intermediate-level L2 participants reported using such a strategy. Since the beginning or the last few words in the sentences are easy to be held in short-term memory, some beginning- and intermediate-level participants prioritized either the beginning or the end part of the sentences in order to complete the test. Two response examples are provided below.

Chunk the info as it was coming in, use beginning only (s13, beginning-level).

If the sentence was long and in multiple parts, especially if I did not know what it meant, I tried to retain the last part of it, because since it was right before I had to repeat it, I found I could remember it better than the first part (s01, intermediate-level).

In the category of communication strategies, we found that L2 participants employed three communication strategies to compensate for areas where they experienced linguistic breakdowns, including *using lexical fillers*, *paraphasing*, and *guessing*. When L2 participants encountered places where they could not understand or remember the words, they used lexical fillers, such as "something" or "什么" ("something" in Chinese), to fill in the gaps (see examples 1 and 2 below). Moreover, some participants could understand the syntactic structures of the sentence and fill the gaps with the words that correspond to the missing part of speech, as shown in example 3.

1. 今年冬天很冷，可是没有下雪 (stimulus sentence)
   Winter *in this year was very cold, but it did not snow.*
   今年something很冷，可是没有下雪 (s06, intermediate-level).
   Something *in this year was very cold, but it did not snow.*
2. 图书馆在公园旁边 (stimulus sentence)
   The library is next to the park.
   图书馆在什么旁边 (s11, advanced-level).
   The library is next to something.
3. 晚上我想跟朋友一起去跑步 *(stimulus sentence)*
   Tonight I want to run together with my friend.
   什么什么时候我和我朋友去跑步 (s21, beginning-level).
   A time *I run with my friend.*

A few L2 participants said that sometimes they would paraphrase in their responses when they understood the meaning of the sentence but could not remember the exact words in the sentence. As shown in the example 4, the L2 participant understood the general gist of the sentence, but probably missed the degree modifier for the adjective. Therefore, he replaced the word "非常" (very) with "真" (really) in his response. Two lower-level L2 participants reported that they guessed either the structure of the sentence based on the phrasing or a few words after the part they recognized.

4. 这件衣服非常便宜 (stimulus sentence)
   This piece of clothing is very *cheap.*
   这件衣服真便宜 (s23, advance-level).
   This piece of clothing is really *cheap.*

We found that L2 participants employed one type of test-wiseness strategy *imitating the sounds* on the EI test. Many L2 learners at higher-proficiency levels reported that they employed two approaches to complete the EI test. They first tried to comprehend the meaning of the sentence. If they could understand it, they would repeat what they heard. If they could not understand it, they would imitate the sounds of the words. A response example is provided below.

If I understood what was said, it was not hard for me to just recreate it in my head and repeat it. However, if there were phrases or words used that I had never heard before, I sometimes had to refer solely to my auditory memory to try and repeat what was said (s11, advanced-level).

Unlike guessing that involves certain levels of L2 processing based on known information, imitation of sounds relies solely

on participants' auditory memory. This strategy is considered to be construct-irrelevant. There were more L2 learners at higher-proficiency levels reported conscious use of this strategy.

Overall, there was a relative increase in frequencies across proficiency levels for the following strategies: approach strategies of *getting the main message of the sentence out* and *controlling the accuracy of the language*, cognitive strategies of *comprehending the meaning of the sentence*, *recognizing familiar words, phrases, or structures*, *translating*, *rehearsing the sentence before repetition*, and *connecting to daily scenarios*, communication strategies of *paraphrasing* and *using lexical fillers,* and the test-wiseness strategy *imitating the sounds*.

Strategy Use and Test Performance

To address the second research question about the effects of strategy use on the EI test performance, we conducted linear mixed-effects regression analyses to examine whether the total number of different types of strategies contributed to the EI scores. The dependent variable was EI score on each individual item, and item and participant served as random effects. **Table 4** presents the statistics of the linear mixed-effects models. First, an empty model was performed to evaluate the appropriateness of treating item and participant as random effects (see Model I statistics in **Table 4**). The empty model included only the random effects. The intercepts of the random effects varied considerably between the items ($\sigma^2 = 0.29$, SD = 0.053) and between the participants ($\sigma^2 = 1.13$, SD = 1.06), suggesting the need to treat these two variables as random effects. Next, participant proficiency level, item length, and item complexity level were added to the empty model to examine the main effects of participants' language ability and task features on the EI scores (Model II). The results of the comparison between the two models showed an improvement in Model II ($\Delta -2LL = 117.79$, $\Delta df = 6$, $p < 0.001$). Lastly, total numbers of the five different types of strategies were added to the model to evaluate the effects of strategy use on the EI performance (Model III). Model III yielded a better fit ($\Delta -2LL = 13.84$, $\Delta df = 5$, $p = 0.02$), thus it was used as the final model. The final model explained 69.2% of the EI score variance. Participant proficiency level, item length, and item complexity level accounted for 51.5% of the variance, whereas total numbers of different types of strategies accounted for 4.8% of the variance. Participants at higher-proficiency levels performed better than lower-proficiency participants: native speakers had on average higher scores ($\beta = 2.37$, $p < 0.001$) than advanced-level learners, advanced-level learners received higher scores ($\beta = 1.56$, $p < 0.001$) than intermediate-level learners, and intermediate-level learners scored higher ($\beta = 0.74$., $p < 0.001$) than beginning-level learners. Overall the participants had lower performance when repeating EI sentences with 12 syllables ($\beta = -0.53$, $p < 0.001$) and targeting intermediate-level ($\beta = -0.39$, $p < 0.001$) and advanced-level lexico-grammar ($\beta = -0.86$, $p < 0.001$). Among the different types of strategies, the total number of cognitive strategies and metacognitive strategies had significant effects on the EI performance, with cognitive strategies having a positive effect ($\beta = 0.27$, $p = 0.002$) and metacognitive strategies having a negative effect ($\beta = -0.62$, $p = 0.008$). In other words, higher-proficiency learners employed significantly more cognitive strategies, whereas lower-proficiency learners employed significantly more metacognitive strategies. Although other strategies did not yield statistical significance, participants who used more communication strategies ($\beta = 0.26$, $p = 0.109$) and approach strategies ($\beta = 0.18$, $p = 0.228$) were likely to have better performance on the EI test, but not if using test-wiseness strategies ($\beta = -0.12$, $p = 0.619$).

TABLE 4 | Statistics for the linear mixed-effects models: Models I, II, and III.

		Model I: Empty			Model II: Main			Model III: Main + Strategy use		
		β	SE	Sig.	*β*	SE	Sig.	*β*	SE	Sig.
Intercept		2.72	0.20	***	2.31	0.18	***	1.69	0.30	***
Fixed effects										
Participant level	Intermediate				0.74	0.26	***	0.78	0.23	***
	Advanced				1.56	0.26	***	1.27	0.28	***
	Native				2.37	0.25	***	2.40	0.23	***
Item length: 12 syllables					−0.53	0.07	***	−0.53	0.07	***
Item level	Intermediate				−0.39	0.10	***	−0.39	0.10	***
	Advanced				−0.86	0.10	***	−0.86	0.10	***
Approach strategies								0.18	0.15	0.228
Cognitive strategies								0.27	0.09	0.002**
Metacognitive strategies								−0.62	0.24	0.008**
Communication strategies								0.26	0.16	0.109
Test-wiseness strategies								−0.12	0.23	0.619
Random effects		Variance	SD		Variance	SD		Variance	SD	
Participant		1.13	1.06		0.27	0.52		0.17	0.41	
Item		0.29	0.53		0.09	0.30		0.09	0.30	
−2*loglikehood		5821.6			5703.7			5689.9		

*Participants n = 32, items n = 72. **p ≤ 0.01 and ***p ≤ 0.001.*

DISCUSSION

This study investigated the nature of strategy use and its effect on the language performance of a Chinese EI test to provide construct validity evidence for EI tasks. Descriptive statistics revealed that participants used 15 different individual strategies on the EI test, representing five main strategy categories, that is, approach strategies, cognitive strategies, metacognitive strategies, communication strategies, and test-wiseness strategies. The cognitive strategy category was used the most frequently by the participants, followed by approach strategies, test-wiseness strategies, communication strategies, and metacognitive strategies. Most participants set the goals of either *getting as many words repeated as possible* or *getting the main message of the sentence out* to complete the EI test. When they were completing the test, they used cognitive strategies (e.g., *understanding the meaning of the sentence*, *recognizing familiar words, phrases, or structures*, or *translating*) to comprehend and retain the information in the sentences. L2 participants also used the metacognitive strategy *prioritizing certain parts of the sentences* to facilitate the completion of the test and utilized communication strategies (e.g., *using lexical fillers* and *paraphrasing*) to deal with any breakdowns. Since the EI test presented no difficulties for native speakers (as shown by their perfect scores on the test), there was no need for them to employ the reported metacognitive and communication strategies. Native speakers only used approach strategies to set the goals for the test and cognitive strategies to help them comprehend and memorize the sentences. In contrast, L2 participants employed a greater and wider range of strategies to compensate for their insufficient L2 knowledge when completing the EI test. Similar to previous research that higher-proficiency L2 learners were more aware of the strategies they used and why they used them (Chamot et al., 1988; Green and Oxford, 1995), higher-proficiency L2 participants in this study also recalled more conscious use of strategies and how they employed the strategies to complete the test.

Among all the reported strategies, many of these strategies could be found in the test-taking strategies employed for completing common listening and speaking tasks. For example, L2 learners used strategies of *translating*, *paraphrasing*, *using lexical fillers*, and *guessing* when responding to independent and integrated speaking tasks (e.g., Swain et al., 2009). They used strategies of *listening for key words*, *connecting to daily scenarios*, *translating* when completing listening comprehension tasks (e.g., Vandergrift, 1997). Other strategies, such as *understanding the meaning of the sentence* and *recognizing familiar words, phrases, or structures*, involve the operation of listening and speaking skills as well as the entailed language processing (e.g., the comprehension of certain vocabulary or grammar); thus, they are expected strategies relevant to the EI construct. In addition, lower-proficiency participants tended to focus on individual words, whereas higher-proficiency participants were able to deal with sentence-level processing. This aligns with previous observations of the different ways of processing by L2 learners at various proficiency levels on listening and speaking tasks (Buck, 2001; Field, 2011). These findings provide evidence to suggest that EI taps into the processes of comprehending and reproducing speech in ways similar to other listening and speech tasks.

That said, due to the repetitive nature of EI, we also observed a test-wiseness strategy unique to EI, especially among higher-proficiency learners. That is, L2 participants tried to imitate the sounds of the words without comprehension after they failed to understand the sentences. Since this strategy does not require linguistic skills and knowledge of the target language, it can be considered construct-irrelevant. This observation seemed to support some researchers' concerns that EI elicit mere rote memorization of individual sounds. However, a negative trend of using the test-wiseness strategy on the EI performance was found in the linear mixed-effects analyses. This indicates that the utilization of the test-wiseness strategy generally does not help participants receive higher scores on the EI test. Based on our observation in the data, many of the imitated sounds were unintelligible. These suggest that test users need not be overly concerned about the effect of this test-wiseness strategy on EI score interpretation.

This study also addressed whether strategy use had effects on EI test performance. Linear mixed-effects models indicate that the total number of cognitive strategies used had a significant, positive effect on the EI scores, while the total number of metacognitive strategies had a significant, negative effect on the EI scores. In other words, more strategies pertaining to the meaningful processing and understanding of the sentence would help participants receive higher scores on the EI test. In contrast, more strategies prone to the reliance on the short-term memory would lead to lower scores on the EI test. As reported by some of the beginning- and intermediate-level participants, they prioritized either the beginning or the end part of the sentences when completing the test. Since the beginning or the last few words in the sentences are easy to be held in short-term memory, it is possible that the participants repeated the words without actually understanding the meaning of those words. However, the negative effect of the metacognitive strategies on EI performance provides evidence in support of the use of EI for assessing general language proficiency, as only strategies require the use of linguistic knowledge and skills in the target language would lead to better performance on the test.

Moreover, we found that the greatest proportion of EI score variance was explained by participants' proficiency levels and item difficulty levels, while strategy use only contributed to a small proportion of EI score variance. Bachman (1990) pointed out that the factors affecting performance on language tests are language ability, individual characteristics of test takers, characteristics of the test method or test tasks, and error of measurement. Among the three types of systematic sources of variability, language ability was the central factor accounting for the variation of test scores. This is consistent with the main effect of proficiency level on EI performance found in the study. This lends support to EI as an indicator of participants' language proficiency. In addition, previous studies suggested that sentence complexity and sentence lengths are two major factors contributing to the difficulty levels of EI items (e.g.,

Ortega et al., 2002; Graham et al., 2010; Yan et al., 2016). L2 participants had lower performances when the sentence length increased and when the sentence contained advanced vocabulary and grammar. This further supports that it requires a sufficient level of language proficiency to be able to perform well on the EI test. Moreover, as strategy use is only one part of the characteristics of test takers, it makes sense that strategy use only accounted for a small proportion of EI score variance.

Taken together, the findings suggest that EI scores are valid indicators of participants' language proficiency. The use of strategies pertaining to the processing and comprehending of the sentences are integral to the successful completion of the EI test, while strategies rely on rote memorization are marginal and detrimental to the test performance.

CONCLUSION

This study examined the nature of strategy use and its effect on the language performance of a Chinese EI test. The results revealed that both L2 and native speakers of Chinese utilized different types of strategies to comprehend and reproduce the stimulus sentences in order to complete the EI test. There are some overlaps between the reported strategies and the common strategies test takers would use for listening and speaking tasks. In addition, cognitive strategies that help the processing of the stimulus sentences contributed positively to the performances on the EI test, whereas metacognitive strategies that allow to focus on only a small chunk of the stimulus sentences contributed negatively to the EI performances. Although some L2 participants imitated the sounds without comprehension, this test-wiseness strategy only had a small effect and generally did not contribute to higher scores on the EI test. These findings provide validity evidence that EI taps into the processing and production of the target language rather than rote memorization of individual sounds.

As an exploratory study, this study utilized a questionnaire to elicit strategy use on EI tasks. Although questionnaires are useful in obtaining a general understanding of test-taking processes and strategies, they could only elicit a partial list of all strategies. It is possible that some strategies participants employed on the EI test were not reported in the questionnaires. To further understand test-taking processes and strategies on EI tasks, other forms of verbal reports, such as stimulated recall, can be conducted to reveal more details on the strategy use for EI tasks. The study observed certain overlaps of strategies between EI and integrated listen-to-speak tasks, both of which requires listening and speaking skills. Future research on the comparison of strategies employed on EI versus integrated listen-to-speak tasks may be helpful to elucidate the linguistic constructs measured by EI. Moreover, this study only examined strategy use by a small number of participants. Although small sample size is common for studies on less commonly taught languages in the US, such as Chinese, a more generalizable conclusion can benefit from a larger number of participants.

The limitations notwithstanding, this study offered a process-oriented perspective to understand the constructs of EI tasks. The findings in this study can provide supporting evidence for the construct validity of EI. That is, EI, as a valid language proficiency measure, assesses speech comprehension and production of the target language.

DATA AVAILABILITY STATEMENT

The elicited imitation test of Chinese used in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: IRIS (https://www.iris-database.org/iris/app/home/detail?id=york:938753).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board Office at the University of Illinois at Urbana-Champaign. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YL designed the study, performed the experiment, analyzed the data, and wrote and edited the manuscript. XY supervised the research, reviewed the results, provided feedback, and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This article was supported by Wake Forest University's Open Access Publishing Fund.

ACKNOWLEDGMENTS

We would like to thank the reviewers for their constructive comments.

REFERENCES

ACTFL (2012). ACTFL Proficiency Guidelines 2012. American Council on the Teaching of Foreign Languages. Available at: https://www.actfl.org/sites/default/files/guidelines/ACTFLProficiencyGuidelines2012.pdf

AERA, APA, and NCME (2014). *Standards for Educational and Psychological Testing*. United States: American Educational Research Association.

Alderson, J. C. (2000). *Assessing reading*. Germany: Ernst Klett Sprachen.

Anwyl-Irvine, A. L., Massonié, J., Flitton, A., Kirkham, N. Z., and Evershed, J. K. (2019). Gorilla in our midst: an online behavioural

experiment builder. *Behav. Res. Methods* 52, 388–407. doi: 10.3758/s13428-019-01237-x

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford, UK: Oxford University Press.

Barkaoui, K., Brooks, L., Swain, M., and Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Appl. Linguis.* 34, 304–324. doi: 10.1093/applin/ams046

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.3929/ethz-b-000105397

Bernstein, J., Van Moere, A., and Cheng, J. (2010). Validating automated speaking tests. *Lang. Test.* 27, 355–377. doi: 10.1177/0265532210364404

Bley-Vroman, R., and Chaudron, C. (1994). "Elicited imitation as a measure of second-language competence," in *Research Methodology in second-Language Acquisition*. eds. A. Mackey and S. Gass (United States: Lawrence Erlbaum), 245–261.

Bowden, H. W. (2016). Assessing second-language oral proficiency for research. *Stud. Second. Lang. Acquis.* 38, 647–675. doi: 10.1017/s0272263115000443

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Stud. Second. Lang. Acquis.* 33, 247–271. doi: 10.1017/s0272263110000756

Bruen, J. (2001). Strategies for success: profiling the effective learner of German. *Foreign Lang. Ann.* 34, 216–225. doi: 10.1111/j.1944-9720.2001.tb02403.x

Brunfaut, T., and McCray, G. (2015). Looking into test-takers' cognitive processes whilst completing reading tasks: a mixed-method eye-tracking and stimulated recall study. ARAGs research reports online (no. AR/2015/001). British Council. Available at: https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf

Brysbaert, M., and Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. Journal of. *J. Cognit.* 1:9. doi: 10.5334/joc.10

Buck, G. (2001). *Assessing Listening*. United Kingdom: Cambridge University Press.

Chamot, A. M., Kupper, L., and Impink-Hernandez, M. (1988). *A Study of Learning Strategies in Foreign Language Instruction: Findings of the Longitudinal Study*. McLean, VA: Interstate Research Associates.

Chaudron, C., Nguyen, H., and Prior, M. (2005). *Manual for the Vietnamese Elicited Imitation test*. Hawaii: National Foreign Language Resource Center, University of Hawaii.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Lang. Assess. Q.* 3, 307–331. doi: 10.1080/15434300701333129

Cohen, A. D. (2013). "Using test-wiseness strategy research in task development," in *The Companion to Language Assessment*. ed. A. J. Kunnan (United States: Wiley and Sons), 893–905.

Cohen, A. D. (2014). *Strategies in Learning and using a second Language*. United States: Routledge.

Davis, L., and Norris, J. (2021). Developing an innovative elicited imitation task for efficient English proficiency assessment. *ETS Res. Rep. Series* 2021, 1–30. doi: 10.1002/ets2.12338

Dreyer, C., and Oxford, R. (1996). "Learning strategies and other predictors of ESL proficiency among Afrikaans speakers in South Africa," in *Language Learning Strategies Around the World: Cross-Cultural Perspectives*. ed. R. Oxford (Hawaii: National Foreign Language Resource Center), 61–74.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Stud. Second. Lang. Acquis.* 27, 141–172. doi: 10.1017/S0272263105050096

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Appl. Linguis.* 27, 464–491. doi: 10.1093/applin/aml001

Fernandez, C. J. (2018). Behind a spoken performance: test takers' strategic reactions in a simu- lated part 3 of the ielts speaking test. *Lang.Testing Asia* 8, 1–20.

Field, J. (2011). "Cognitive validity," in *Examining Speaking: Research and Practice in Assessing second Language Speaking*. ed. L. Taylor (United States: Cambridge University Press), 65–111.

Fraser, C., Bellugi, U., and Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *J. Verbal Learn. Verbal Behav.* 2, 121–135. doi: 10.1016/s0022-5371(63)80076-6

Gaillard, S., and Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Lang. Learn.* 66, 419–447. doi: 10.1111/lang.12157

Gass, S. M., and Mackey, A. (2013). *Stimulated Recall Methodology in second Language Research*. United States: Routledge.

Gordon, C. (1987). The effect of Testing Method on Achievement in reading Comprehension Tests in English as a Foreign Language. MA Thesis. Ramat Aviv, Israel: Tel Aviv University.

Graham, C. R., McGhee, J., Millard, B., Prior, M. T., Watanabe, Y., and Lee, S. (2010). "The role of lexical choice in elicited imitation item difficulty," in *Selected Proceedings of the 2008 Second Language Research Forum*. ed. M. T. Prior (United States: Cascadilla Proceedings Project), 57–72.

Green, J. M., and Oxford, R. (1995). A closer look at learning strategies, L2 proficiency, and gender. *TESOL Q.* 29, 261–297. doi: 10.2307/3587625

Henning, G. (1983). Oral proficiency testing: comparative validities of interview, imitation, and completion methods. *Lang. Learn.* 33, 315–332. doi: 10.1111/j.1467-1770.1983.tb00544.x

Huang, L.-S. (2013). Cognitive processes involved in performing the IELTS speaking test: respondents' strategic behaviours in simulated testing and non-testing contexts. *IELTS Res. Rep. Online Series* 1, 1–51.

Huang, H.-T. D. (2016). Test-taking strategies in L2 assessment: The test of English for international communication speaking test. *Percept. Mot. Skills* 123, 64–90. doi: 10.1177/0031512516660699

Ishikawa, L., Hall, K., and Settles, B. (2016). The Duolingo English test and academic English. Technical report, Duolingo DRR. Available at: englishtest.duolingo.com/resources (January 16, 2016).

Jensen, E. D., and Vinther, T. (2003). Exact repetition as input enhancement in second language acquisition. *Lang. Learn.* 53, 373–428. doi: 10.1111/1467-9922.00230

Kim, Y., Tracy-Ventura, N., and Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *Mod. Lang. J.* 100, 655–673. doi: 10.1111/modl.12346

Kostromitina, M., and Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Stud. Second. Lang. Acquis.* 1, 1–26. doi: 10.1017/s0272263121000395

Li, Y., Wen, X., and Xie, T. (2014). CLTA 2012 survey of college-level Chinese language programs in North America. *J. Chin. Lang. Teach. Assoc.* 49, 1–49.

Markman, B. R., Spilka, I. V., and Tucker, G. R. (1975). The use of elicited imitation in search of an interim French grammar. *Lang. Learn.* 25, 31–41. doi: 10.1111/j.1467-1770.1975.tb00107.x

Mozgalina, A. (2015). Applying an Argument-based Approach for Validating Language Proficiency Assessments in second Language Acquisition Research: The Elicited Imitation test for Russian. Doctoral Dissertation. Washington, DC: Georgetown University.

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Bilingualism* 2, 1–37.

Nakatani, Y. (2006). Developing an oral communication strategy inventory. *Mod. Lang. J.* 90, 151–168. doi: 10.1111/j.1540-4781.2006.00390.x

Okura, E., and Lonsdale, D. (2012). "Working memory's Meager Involvement in Sentence Repetition Tests." in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*. eds. N. Miyake, D. Peebles and R. P. Cooper (United States: Cognitive Science Society), 2132–2137.

Ortega, L., Iwashita, N., Norris, J. M., and Rabie, S. (2002). *An Investigation of Elicited Imitation Tasks in Crosslinguistic SLA Research*. Toronto, Canada: Second Language Research Forum.

Oxford, R. L., and Burry-Stock, J. A. (1995). Assessing the use of language learning strategies worldwide with the ESL/EFL version of the strategy inventory for language learning (SILL). *System* 23, 1–23. doi: 10.1016/0346-251x(94)00047-a

Oxford, R. L., and Ehrman, M. E. (1995). Adults' language learning strategies in an intensive foreign language program in the United States. *System* 23, 359–386. doi: 10.1016/0346-251x(95)00023-d

Park, G. P. (1997). Language learning strategies and English proficiency in Korean university students. *Foreign Lang. Ann.* 30, 211–221. doi: 10.1111/j.1944-9720.1997.tb02343.x

Park, H. I., Solon, M., Henderson, C., and Dehghan-Chaleshtori, M. (2020). The roles of working memory and oral language abilities in elicited imitation performance. *Mod. Lang. J.* 104, 133–151. doi: 10.1111/modl.12618

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use TOEFL reading achievement test performance. *Lang. Test.* 20, 26–56. doi: 10.1191/0265532203lt243oa

Prutting, C. A., Gallagher, T. M., and Mulac, A. (1975). The expressive portion of the nsst compared to a spontaneous language sample. *J. Speech Hear. Disord.* 40, 40–48. doi: 10.1044/jshd.4004.544b

Purpura, J. E. (1999). *Learner Strategy use and Performance on Language Tests: A Structural Equation Modeling Approach*. United States: Cambridge University Press.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: http://www.R-project.org/ (Accessed July 05, 2019).

Slobin, D., and Welsh, C. A. (1973). "Elicited imitation as a research tool in developmental psycholinguistics," in *Studies of child Language Development*. eds. C. Ferguson and D. Slobin (United States: Holt Rinehart and Winston), 485–489.

Song, X. (2005). *Language Learner Strategy use and English Proficiency on the Michigan English Language Assessment Battery. Vol. 3*. Canada: Queen's University.

Spada, N., Shiu, J. L. J., and Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: comparisons with other validation studies. *Lang. Learn.* 65, 723–751. doi: 10.1111/lang.12129

Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Lang. Test.* 14, 214–231. doi: 10.1177/026553229701400205

Suzuki, Y., and DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Lang. Learn.* 65, 860–895. doi: 10.1111/lang.12138

Swain, M., Huang, L. S., Barkaoui, K., Brooks, L., and Lapkin, S. (2009). The speaking section of the TOEFL IBT™(SSTIBT): test-takers' reported strategic behaviors. *ETS Res. Rep. Series* 2009, 1–118. doi: 10.1002/j.2333-8504.2009.tb02187.x

Tracy-Ventura, N., McManus, K., Norris, J. M., and Ortega, L. (2014). ""Repeat as much as you can": elicited imitation as a measure of oral proficiency in L2 French," in *Measuring L2 Proficiency: Perspectives from SLA*. eds. P. Leclercq, H. Hilton and A. Edmonds (United Kingdom: Multilingual Matters), 143–166.

Trofimovich, P., and Baker, W. (2007). Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals. *Appl. Psycholinguist.* 28, 251–276. doi: 10.1017/s0272263106060013

Underhill, N. (1987). *Testing Spoken Language: A Handbook of oral Testing Techniques*, Cambridge, UK: Cambridge University Press.

Van Hell, J. G., and Tokowicz, N. (2010). Event-related brain potentials and second language learning: syntactic processing in late L2 learners at different L2 proficiency levels. *Second. Lang. Res.* 26, 43–74. doi: 10.1177/0267658309337637

Vandergrift, L. (1997). The comprehension strategies of second language (French) listeners: A descriptive study. *Foreign Lang. Ann.* 30, 387–409. doi: 10.1111/j.1944-9720.1997.tb02362.x

Vinther, T. (2002). Elicited imitation: A brief overview. *Int. J. Appl. Linguist.* 12, 54–73. doi: 10.1111/1473-4192.00024

West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from l2 Spanish. *Lang. Cognit.* 4, 203–222. doi: 10.1515/langcog-2012-0011

Wu, S. L., and Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Lang. Ann.* 46, 680–704. doi: 10.1111/flan.12063

Yan, X., Lei, Y., and Shih, C. (2020). A corpus-driven, curriculum-based Chinese elicited imitation test in US universities. *Foreign Lang. Ann.* 53, 704–732. doi: 10.1111/flan.12492

Yan, X., Maeda, Y., Lv, J., and Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Lang. Test.* 33, 497–528. doi: 10.1177/0265532215594643

Yang, H.-C., and Plakans, L. (2012). Second language writers' strategy use and performance on an integrated Reading-listening-writing task. *TESOL Q.* 46, 80–103. doi: 10.1002/tesq.6

Yoon, S. (2010). English Syllable confusion and Imitation in Korean Bilingual and Monolingual children and Adults. Doctoral Dissertation. Champaign, IL: University of Illinois at Urbana-Champaign.

Yoshida-Morise, Y. (1998). "The use of communication strategies in language proficiency interviews," in *Talking and Testing: Discourse Approaches to the Assessment of oral Proficiency*. eds. R. Young and A. W. He (Netherlands: John Benjamins Publishing), 205–238.

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.

frontiers | Frontiers in Psychology

OPINION
published: 30 June 2022
doi: 10.3389/fpsyg.2022.926367

Cognitive Factors Influencing Utterance Fluency in L2 Dialogues: Monadic and Non-monadic Perspectives

Ruiling Feng [1,2]*

[1] Foreign Language College, Tianjin Normal University, Tianjin, China, [2] Department of Foreign Languages and Literatures, Tsinghua University, Beijing, China

Keywords: utterance fluency in L2 dialogues, interlocutor, monadic perspective, non-monadic perspective, cognitive factors

OPEN ACCESS

Edited by:
Peijian Paul Sun, Zhejiang University, China

Reviewed by:
Pauliina Peltonen, University of Turku, Finland

***Correspondence:**
Ruiling Feng
fengruiling@126.com

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

Received: *22 April 2022*
Accepted: *15 June 2022*
Published: *30 June 2022*

Citation:
Feng R (2022) Cognitive Factors Influencing Utterance Fluency in L2 Dialogues: Monadic and Non-monadic Perspectives. Front. Psychol. 13:926367. doi: 10.3389/fpsyg.2022.926367

INTRODUCTION

Studies of second language (L2) speech fluency have largely focused on monologs, while dialogues are rarely studied (McCarthy, 2010; Tavakoli, 2016; Foster, 2020). In dialogues, two or more interlocutors take turns contributing to the flow of interaction. Therefore, utterance fluency (overt fluency performance) in L2 dialogues, including individual/within-turn fluency and interactional/between-turn fluency (Peltonen, 2017a, 2020), requires not only speaker-internal cognitive processing but also between-speaker cognitive cooperation based on shared understanding (Roever and Kasper, 2018; Pickering and Garrod, 2021). Speaker-internal cognitive factors are activated by social interaction (Tavakoli and Wright, 2020). Consequently, examination of cognitive factors influencing utterance fluency in L2 dialogues should include both a monadic perspective that hinges on each individual's private cognitive processing and a non-monadic perspective that analyzes the dialogue as a whole system by considering the relationship between each individual's utterances (Tavakoli and Wright, 2020; Pickering and Garrod, 2021). Incorporating both perspectives could contribute to the ongoing discussion about factors influencing L2 speech production. Especially, the non-monadic view can help reconcile the overreliance on individual fluency performance. Therefore, this paper aims to examine a tentative list of cognitive factors affecting utterance fluency in L2 dialogues from the two perspectives.

MONADIC PERSPECTIVE

The monadic perspective focuses on factors affecting speaker-internal mental activities. From this perspective, cognitive factors influencing fluency in L2 dialogues mainly include L2-specific cognitive fluency (access to L2 knowledge), general cognitive fluency (reflected in personal speaking style), and overall L2 proficiency (linguistic repertoire) (e.g., Segalowitz, 2010, 2016; Kahng, 2014, 2020; Pérez Castillejo, 2018).

L2-Specific Cognitive Fluency

Studies relating L2-specific cognitive fluency to L2 utterance fluency are reviewed in this section. Segalowitz (2010) argued that L2 fluency performance is influenced by both L2-specific cognitive

fluency and language-independent personal speaking style. His argument has been corroborated by later studies (e.g., Kahng, 2014, 2020; Segalowitz, 2016; Suzuki and Kormos, 2022). L2-specific cognitive fluency is gained by partialling out first language (L1) data from equivalent L2 data (Segalowitz, 2010, 2016; Bradlow et al., 2017). Cognitive processes could consist of the four cognitive modules in Levelt's (1989; 1999) speech production model, including conceptualization (preverbal message generation), formulation (grammatical and morpho-phonological encoding), articulation, and monitoring (self-perception) (Tavakoli et al., 2020).

Regarding L2-specific cognitive fluency, the four cognitive modules have received uneven scholarly attention. Conceptualization is regarded as language-independent (e.g., Levelt, 1989, 1999; De Bot, 1992; Segalowitz, 2010); therefore, it is generally excluded from this research strand. Among the other three modules, formulation and articulation are the main foci, formulation in particular. Segalowitz and Freed (2004) measured the L2-specific speed of lexical access and attention control by the reaction time in a semantic classification task and the efficiency of the two measures by the coefficient variation (standard deviation divided by the mean) of the reaction time. Segalowitz (2016) also adopted reaction time and coefficient variation of it, and linguistic attention flexibility. Different from Segalowitz's tests, Kahng (2020) measured lexical retrieval by a picture-naming task and syntactic encoding by a sentence completion task. Besides the aforementioned quantitative measurement, qualitative measurement of formulation fluency is also used. For example, Kahng (2014) adopted stimulated recall to tap thoughts during filled and silent pauses. As for the measurement of articulation, delayed picture naming tasks (e.g., de Jong and Mora, 2017; Kahng, 2020) and controlled speech tasks (Suzuki and Kormos, 2022) have been employed. Monitoring is rarely studied as a dimension of L2-specific cognitive fluency or in terms of its relationship to utterance fluency. However, some monitoring-related features (e.g., repetitions and self-corrections) have been found language-independent rather than L2-specific (Peltonen and Lintunen, 2016; Georgiadou and Roehr-Brackin, 2017; Olkkonen, 2017).

General Cognitive Fluency

Recent studies have found that equivalent L1 fluency performance measures, accounting for general cognitive fluency or stable personal speaking style, can help explain L2 utterance fluency (e.g., Segalowitz, 2010, 2016; Bradlow et al., 2017; Kahng, 2020). For example, mean silent pause duration and filled pause frequency are mainly related to general instead of L2-specific cognitive fluency (de Jong et al., 2013; Kahng, 2020). General cognitive fluency is especially associated with conceptualization as encyclopedia knowledge rather than linguistic knowledge is used in this stage (Segalowitz, 2010; Kahng, 2014, 2020). However, not all L2 fluency performance measures demonstrate a correlation with general cognitive fluency. For instance, L2 speech rate change cannot be predicted by the equivalent L1 measure, and therefore might be an L2-specific feature (Baese-Berk and Morrill, 2015; Baese-Berk and Bradlow, 2021). Besides, overall L2 proficiency could moderate the relationship between general cognitive fluency and L2 utterance fluency, as speakers of higher proficiency demonstrate a stronger correlation between L1 and L2 utterance fluency (Huensch and Tracy–Ventura, 2017; Peltonen, 2018). However, overall L2 proficiency does not mediate the relationship (Duran-Karaoz and Tavakoli, 2020).

Overall L2 Proficiency

Overall L2 proficiency represents L2 linguistic repertoire and influences fluency performance in L2 dialogues (Kahng, 2020; Tavakoli and Wright, 2020). L2 and L1 speakers demonstrate different (dis)fluency patterns. For example, L2 speakers pause markedly more and longer within clauses than L1 speakers do (Kahng, 2014; de Jong, 2016). It could be that L2 speakers are under higher processing time pressure (Baddeley, 2003), due to smaller processing units and lower automaticity (Kroll and de Groot, 1997; Jiang, 2000; Bundgaard-Nielsen et al., 2011; Wang, 2014; Tavakoli and Wright, 2020).

Some fluency-related studies have examined L2 speakers of different proficiency levels. With higher proficiency, reliance on L1 mediation decreases in L2 lexical retrieval (Jiang, 2000), leading to greater processing automaticity with lower switching costs (Costa and Santesteban, 2004; DeKeyser, 2005; Segalowitz, 2010). Williams and Korko (2019) found advanced speakers showed fewer corrections, silent pauses, and filled pauses than lower intermediate speakers in L2 monologs. They attributed the differences to automaticity and the use of formulaic structures. Besides, proficiency affects the length and frequency of turn pauses in dialogues (Peltonen, 2017b; van Os et al., 2020). Lower-proficiency speakers are more hesitant to start turns, resulting in longer and more turn pauses (Peltonen, 2017b), while higher-proficiency speakers could be better ready to take turns with higher automaticity.

NON-MONADIC PERSPECTIVE

The non-monadic perspective analyzes dialogue as a whole system with interaction and interdependence between interlocutors (McCarthy, 2010; Segalowitz, 2016; Tavakoli, 2016; Peltonen, 2017b; Tavakoli and Wright, 2020). As such, speaker stance, interactional competence, and interlocutors' cognitive factors are viewed as potentially contributing to fluency in L2 dialogues.

Speaker Stance

Speaker stance represents an attitude, willingness, or orientation instead of ability and can affect how individuals engage in dialogues. If speakers regard dialogue as a self-performing activity and take a safer speaker stance, they would pay substantial attention to their own production but little to interlocutors' utterances (He and Dai, 2006; Tavakoli and Wright, 2020; Pickering and Garrod, 2021). A safer speaker stance might help achieve higher within-turn fluency, however, at the sacrifice of interactive listening and contingent responses. In contrast, a more other-oriented speaker might consider dialogue as a joint activity and keep both speaker and listener roles active concurrently (Pickering and Garrod, 2021). Therefore, other-oriented speakers are more inclined to incorporate

interactive listening, between-turn responsiveness, and between-speaker alignment and synchrony. Compared to a safer speaker stance, a more other-orientated stance may slow down one's speech production due to more time for comprehension and hence less time for production (Tavakoli and Wright, 2020). The relationship between speaker stance and fluency in L2 dialogues might be moderated by overall L2 proficiency, which affects attentional resources allocated to individual speech and interactional aspects of dialogues (Levelt, 1989; Kormos, 2006).

Introspective and retrospective self-assessment can detect speaker stance, exploring speakers' perceptions of and attitudes toward speech tasks (Alderson, 1985). For example, He and Dai (2006) designed a questionnaire to tap how students viewed and dealt with the group discussion in a high-stakes test. Results showed that most students took a safer speaker stance, largely attending to processing individual turns rather than listening and responding to interlocutors. As such, they could display their most fluent English with long turns but low responsiveness to the just-uttered turn from interlocutors. The safer speaker stance could be associated with factors such as culture and specific task context (e.g., high-stakes tests vs. free discussions).

Interactional Competence

Interactional competence refers to one's ability to adopt different communication strategies, and actively listen and respond to previous speakers' contributions based on proper comprehension (Galaczi, 2014; May et al., 2019). It is important for the co-construction of dialogues (Roever and Kasper, 2018) and affects both individual/within-turn and interactional/between-turn fluency (May et al., 2019; Tavakoli and Wright, 2020). Speakers of higher interactional competence are more likely to respond to and synchronize with interlocutors, while those of lower competence might experience difficulties engaging in dialogues as they cannot guarantee appropriate responsiveness and synchrony (Galaczi, 2008). Synchronization could help keep interlocutors on the same wavelength (Ward and Tsukahara, 2003), and increase fluency in a dialogue as a whole system instead of an individual performance (Pickering and Garrod, 2021). Note that interactional competence might overlap with fluency in dialogues in features like turn pause and breakdown repair (Galaczi and Taylor, 2018; Zhang and Jin, 2021).

Interactional competence is difficult to operationalize due to its multicomponential nature (Galaczi, 2014). Here I propose two dimensions for the measurement of interactional competence, interactive listening and between-turn responsiveness (e.g., May, 2009, 2011; Lam, 2018; Ross, 2018). Interactive listening represents attention to interlocutors' utterances. It aims to show support and comprehension. Responsiveness between adjacent turns could promote predictability of the dialogic flow, and thus fluency (Smith and McMurray, 2018; Pickering and Garrod, 2021). These two dimensions are inevitably related in dialogues, as producing a turn contingent on the just-uttered turn depends on comprehension as a result of interactive listening, though interactive listening cannot guarantee comprehension or responsiveness (Galaczi, 2014).

Interactive listening can be measured by verbal and non-verbal features. Verbal features include listener support moves such as backchannelling and confirmation of comprehension (Galaczi, 2014; Lam, 2018). Non-verbal features refer to paralinguistic features like eye contact and gesticulation (Jenkins and Parra, 2003; Ross, 2018). These features signal listener attentiveness but not necessarily comprehension (Ross, 2018). Sometimes, they are even used to mask insufficient comprehension (Galaczi, 2014; Lam, 2018).

Adequate between-turn responsiveness can demonstrate a link to and extension of the previous speaker's contribution (Galaczi, 2014). A responsive turn (contingent response in Lam, 2018) may include three conversational actions, namely formulation of a just-uttered turn, explaining (dis)agreement with the previous turn, and expanding the topic (Lam, 2018). Based on these actions, Lam proposed three proficiency levels of producing responsive turns (lower, mid, and higher levels).

Interlocutors' Cognitive Factors

In dialogues, interlocutors tend to align and synchronize with each other (Pickering and Garrod, 2021); therefore, the aforementioned cognitive factors of each interlocutor may impact, indirectly *via* their utterances, other interlocutors' fluency performance (Tavakoli and Wright, 2020; Pickering and Garrod, 2021). Speakers' competence and performance in a dialogue can decide, to a large extent, how and what their interlocutors try to comprehend, respond to, and align and synchronize with (Benuš, 2021; Pickering and Garrod, 2021). Previous studies have found dialogue partners converge in some fluency features, for example, inter-word intervals (Himberg et al., 2015) and speech rate (Cohen Priva et al., 2017). The synchronization could facilitate more seamless turn switching (Pickering and Garrod, 2021). Even highly self-oriented speakers have to synchronize somehow with and therefore be influenced by their interlocutors.

CONCLUSIONS

Analysis of cognitive factors affecting utterance fluency in L2 dialogues should incorporate both monadic and non-monadic perspectives. Monadically, L2-specific cognitive fluency, general cognitive fluency, and overall L2 proficiency could affect speakers' fluency performance *via* private cognitive processing. Non-monadically, speaker stance, interactional competence, and interlocutors' cognitive factors influence how speakers listen to, comprehend, and accommodate interlocutors' utterances, and therefore their fluency performance in L2 dialogues. Cautions should be made when predicting utterance fluency in L2 dialogues with a myriad of cognitive factors. For example, some factors might be correlated, which leads to multicollinearity. Also, a linear relationship might not exist between fluency and some factors such as proficiency. This paper focuses on cognitive factors, while affective and sociocultural factors could also affect fluency (Sun and Zhang, 2020; Sun, 2022). These factors warrant future research from monadic

and non-monadic perspectives regarding utterance fluency in L2 dialogues.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This work was supported by the Project of Discipline Innovation and Advancement (PODIA)-Foreign Language Education Studies at Beijing Foreign Studies University (Grant Number: 2020SYLZDXM011), Beijing.

REFERENCES

Alderson, J. C. (1985). "Innovations in language testing?" in *Innovations in Language Testing*, ed M. Portal (Slough: NFER-Nelson), 93–105.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nat. Rev. Neurosci.* 4, 829–839. doi: 10.1038/nrn1201

Baese-Berk, M. M., and Bradlow, A. R. (2021). "Variability in speaking rate of native and nonnative speech," in *Second Language Speech Learning: Theoretical and Empirical Progress*, ed R. Wayland (Cambridge: Cambridge University Press), 312–334.

Baese-Berk, M. M., and Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *J. Acoust. Soc. Am.* 138, EL223–EL228. doi: 10.1121/1.4929622

Benuš, Š. (2021). *Investigating Spoken English: A Practical Guide to Phonetics and Phonology Using Praat, 1st Edn*. Cham: Palgrave Macmillan.

Bradlow, A. R., Kim, M., and Blasingame, M. (2017). Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate. *J. Acoust. Soc. Am.* 141, 886–899. doi: 10.1121/1.4976044

Bundgaard-Nielsen, R. L., Best, C. T., and Tyler, M. D. (2011). Vocabulary size is associated with second-language vowel perception performance in adult learners. *Stud. Second Lang. Acquis.* 33, 433–461. doi: 10.1017/S0272263111000040

Cohen Priva, U., Edelist, L., and Gleason, E. (2017). Converging to the baseline: corpus evidence for convergence in speech rate to interlocutor's baseline. *J. Acoust. Soc. Am.* 141, 2989–2996. doi: 10.1121/1.4982199

Costa, A., and Santesteban, M. (2004). Lexical access in bilingual speech production: evidence from language switching in highly proficient bilinguals and L2 learners. *J. Mem. Lang.* 50, 491–511. doi: 10.1016/j.jml.2004.02.002

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Appl. Linguist.* 13, 1–24.

de Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *Int. Rev. Appl. Linguist. Lang. Teach.* 54,113–132. doi: 10.1515/iral-2016-9993

de Jong, N. H., and Mora, J. C. (2017). Does having good articulatory skills lead to more fluent speech in first and second languages? *Stud. Second Lang. Acquis.* 41, 227–239. doi: 10.1017/S0272263117000389

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., and Hulstijn, J. N. (2013). Linguistic skills and speaking fluency in a second language. *Appl. Psycholinguist.* 34, 893–916. doi: 10.1017/S0142716412000069

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Lang. Learn.* 55, 1–25. doi: 10.1111/j.0023-8333.2005.00294.x

Duran-Karaoz, Z., and Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behavior: the case of L1 Turkish and L2 English speakers. *Stud. Second Lang. Acquis.* 4, 671–695. doi: 10.1017/S0272263119000755

Foster, P. (2020). Oral fluency in a second language: a research agenda for the next ten years. *Lang. Teach.* 53, 446–461. doi: 10.1017/S026144482000018X

Galaczi, E., and Taylor, L. (2018). Interactional competence: conceptualisations, operationalisations, and outstanding questions. *Lang. Assess. Q.* 3, 219–236. doi: 10.1080/15434303.2018.1453816

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: the case of the First Certificate in English examination. *Lang. Assess. Q.* 5, 89–119. doi: 10.1080/15434300801934702

Galaczi, E. D. (2014). Interactional competence across proficiency levels: how do learners manage interaction in paired tests? *Appl. Linguist.* 35, 553–574. doi: 10.1093/applin/amt017

Georgiadou, E., and Roehr-Brackin, K. (2017). Investigating executive working memory and phonological short-term memory in relation to fluency and self-repair behavior in L2 speech. *J. Psycholinguist. Res.* 4, 877–895. doi: 10.1007/s10936-016-9463-x

He, L., and Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Lang. Teach.* 23, 370–401. doi: 10.1191/0265532206lt333oa

Himberg, T., Hirvenkari, L., Mandel, A., and Hari, R. (2015). Word-by-word entrainment of speech rhythm during joint story building. *Front. Psychol.* 6, 797. doi: 10.3389/fpsyg.2015.00797

Huensch, A., and Tracy–Ventura, N. (2017). Understanding second language fluency behavior: the effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Appl. Psycholinguist.* 38, 755–785. doi: 10.1017/S0142716416000424

Jenkins, S., and Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: the complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *Mod. Lang. J.* 87, 90–107. doi: 10.1111/1540-4781.00180

Jiang, N. (2000). Lexical representation and development in a second language. *Appl. Linguist.*21, 47–77. doi: 10.1093/applin/21.1.47

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: temporal measures and stimulated recall. *Lang. Learn.* 64, 809–854. doi: 10.1111/lang.12084

Kahng, J. (2020). Explaining second language utterance fluency: contribution of cognitive fluency and first language utterance fluency. *Appl. Psycholinguist.* 41, 457–480. doi: 10.1017/S0142716420000065

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kroll, J. F., and de Groot, A. M. B. (1997). "Lexical and conceptual memory in the bilingual: mapping form to meaning in two languages," in *Tutorials in Bilingualism: Psycholinguistic Perspectives*, eds A. M. B. de Groot, and J. F. Kroll (Hillsdale, NJ: Erlbaum), 169–199.

Lam, D. M. K. (2018). What counts as "responding"? Contingency on previous speaker contribution as a feature of interactional competence. *Lang. Test.* 35, 377–401. doi: 10.1177/0265532218758126

Levelt, W. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: The MIT Press.

Levelt, W. (1999). "Producing spoken language: a blueprint of the speaker," in *The Neurocognition of Language*, eds C. Brown, and P. Hagoort (Oxford: Oxford University Press), 83–122.

May, L. (2009). Co-constructed interaction in a paired speaking test: the rater's perspective. *Lang. Test.* 26, 397–421. doi: 10.1177/0265532209104668

May, L. (2011). Interactional competence in a paired speaking test: features salient to raters. *Lang. Assess. Q.* 8, 127–145. doi: 10.1080/15434303.2011.565845

May, L., Nakatsuhara, F., Lam, D., and Galaczi, E. (2019). Developing tools for learning oriented assessment of interactional competence: bridging theory and practice. *Lang. Test.* 37, 165–188. doi: 10.1177/0265532219879044

McCarthy, M. (2010). Spoken fluency revisited. *Eng. Prof. J.* 1, 1–15. doi: 10.1017/S2041536210000012

Olkkonen, S. (2017). Processing limitations in L2 fluency: analysis of inaccuracies in lexical access. *Apples J. Appl. Lang. Stud.* 1, 19–41. doi: 10.17011/apples/urn.201704252081

Peltonen, P. (2017a). "L2 fluency in spoken interaction: a case study on the use of other-repetitions and collaborative completions," in *Näkökulmia Toisen Kielen Puheeseen – Insights into Second Language Speech, AFinLA-e: Soveltavan kielitieteen tutkimuksia 10*, eds M. Kuronen, P. Lintunen, and T. Nieminen (Jyväskylä: Finnish Association of Applied Linguistics, AFinLA), 118–138.

Peltonen, P. (2017b). Temporal fluency and problem-solving in interaction: an exploratory study of fluency resources in L2 dialogue. *System* 70, 1–13. doi: 10.1016/j.system.2017.08.009

Peltonen, P. (2018). Exploring connections between first and second language fluency: a mixed methods approach. *Mod. Lang. J.* 4, 676–692. doi: 10.1111/modl.12516

Peltonen, P. (2020). *Individual and Interactional Speech Fluency in L2 English from a Problem-Solving Perspective: A Mixed-methods Approach* (Doctoral dissertation). University of Turku. Available online at: https://urn.fi/URN:ISBN:978-951-29-8137-3 (accessed June 1, 2022).

Peltonen, P., and Lintunen, P. (2016). Integrating quantitative and qualitative approaches in L2 fluency analysis: a study of Finnish-speaking and Swedish-speaking learners of English at two school levels. *Eur. J. Appl. Linguist.* 4, 209–238. doi: 10.1515/eujal-2014-0018

Pérez Castillejo, S. (2018). The role of foreign language anxiety on L2 utterance fluency during a final exam. *Lang. Test.* 36, 327–345. doi: 10.1177/0265532218777783

Pickering, M. J., and Garrod, S. (2021). *Understanding Dialogue: Language Use and Social Interaction.* Cambridge: Cambridge University Press.

Roever, C., and Kasper, G. (2018). Speaking in turns and sequences: interactional competence as a target construct in testing speaking. *Lang. Test.* 35, 331–355. doi: 10.1177/0265532218758128

Ross, S. (2018). Listener response as a facet of interactional competence. *Lang. Test.* 35, 357–375. doi: 10.1177/0265532218758125

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York, NY: Routledge.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *Int. Rev. Appl. Linguist. Lang. Teach.* 54, 79–95. doi: 10.1515/iral-2016-9991

Segalowitz, N., and Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: learning Spanish in at Home and study abroad contexts. *Stud. Second Lang. Acquis.* 26, 173–199. doi: 10.1017/S0272263104262027

Smith, N. A., and McMurray, B. (2018). Temporal responsiveness in mother-child dialogue: a longitudinal analysis of children with normal hearing and hearing loss. *Infancy* 23, 410–431. doi: 10.1111/infa.12227

Sun, P. (2022). The influence of cognitive, affective, and sociocultural individual differences on L2 Chinese speech performance: a multiple indicators multiple causes approach. *J. Psycholinguist. Res. Early access*. doi: 10.1007/s10936-022-09847-x

Sun, P., and Zhang, L. J. (2020). A multidimensional perspective on individual differences in multilingual learners' L2 Chinese speech production. *Front. Psychol.* 11, 59. doi: 10.3389/fpsyg.2020.00059

Suzuki, S., and Kormos, J. (2022). The multidimensionality of second language oral fluency: interfacing cognitive fluency and utterance fluency. *Stud. Second Lang. Acquis.* doi: 10.1017/S0272263121000899

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: challenges in defining and measuring L2 fluency. *Int. Rev. Appl. Linguist. Lang. Teach.* 54, 133–150. doi: 10.1515/iral-2016-9994

Tavakoli, P., Nakatsuhara, F., and Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Mod. Lang. J.* 104, 169–191. doi: 10.1111/modl.12620

Tavakoli, P., and Wright, C. (2020). *Second Language Speech Fluency: From Research to Practice*. Cambridge: Cambridge University Press.

van Os, M., de Jong, N. H., and Bosker, H. R. (2020). Fluency in dialogue: turn taking behavior shapes perceived fluency in native and nonnative speech. *Lang. Learn.* 70, 1183–1217. doi: 10.1111/lang.12416

Wang, Z. (2014). "On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions," In *Processing Perspectives on Task Performance*, ed P. Skehan (Amsterdam: Benjamins), 27–62

Ward, N., and Tsukahara, W. (2003). A study in responsiveness in spoken dialog. *Int. J. Hum.Comput. Stud.* 59, 603–630. doi: 10.1016/S1071-5819(03)00085-5

Williams, S. A., and Korko, M. (2019). Pause behavior within reformulations and the proficiency level of second language learners of English. *Appl. Psycholinguist.* 40, 723–742. doi: 10.1017/S0142716418000802

Zhang, L., and Jin, Y. (2021). Assessing interactional competence in the computer-based CET-SET: an investigation of the use of communication strategies. *Assess. Educ.* 4, 389–410. doi: 10.1080/0969594X.2021.1976107

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

frontiers | Frontiers in Psychology

ORIGINAL RESEARCH
published: 13 July 2022
doi: 10.3389/fpsyg.2022.916736

Variability in L2 Vowel Production: Different Elicitation Methods Affect Individual Speakers Differently

*Murray J. Munro**

Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada

Elicitation methods are known to influence second language speech production. For teachers and language assessors, awareness of such effects is essential to accurate interpretations of testing outcomes. For speech researchers, understanding why one method gives better performance than another may yield insights into how second-language phonological knowledge is acquired, stored, and retrieved. Given these concerns, this investigation compared L2 vowel intelligibility on two elicitation tasks and determined the degree to which differences generalized across vowels, vowels in context, lexical items, and individual speakers. The dependent variable was the intelligibility of Cantonese speakers' productions of English /i ɪ u ʊ/ in varying phonetic environments. In a picture-naming task, the speakers produced responses without an auditory prompt. In a second task–interrupted repetition–they heard exemplars of the same targets without pictures, and repeated each one after counting aloud to 10, a step intended to disrupt their short-term auditory store and therefore prevent simple mimicry. For target words with scores below 80% on picture naming, mean intelligibility was more than 10 points higher on interrupted repetition. However, that difference did not generalize across conditions or across speakers. Thus, although it is technically accurate to say that, on average, interrupted repetition yielded better vowel intelligibility than did picture naming, that observation requires a great deal of qualification, particularly because of individual speaker differences. The outcomes are interpreted in terms of their relevance to language assessment and phonetic learning.

Keywords: Cantonese, second language acquisition, phonetics, ESL, intelligibility, vowels

OPEN ACCESS

Edited by:
Xun Yan,
University of Illinois at Urbana-Champaign, United States

Reviewed by:
Elena Babatsouli,
University of Louisiana at Lafayette, United States
Ji Young Kim,
University of California, Los Angeles, United States

***Correspondence:**
Murray J. Munro
mjmunro@sfu.ca

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

Received: *09 April 2022*
Accepted: *25 May 2022*
Published: *13 July 2022*

Citation:
Munro MJ (2022) Variability in L2 Vowel Production: Different Elicitation Methods Affect Individual Speakers Differently. Front. Psychol. 13:916736. doi: 10.3389/fpsyg.2022.916736

INTRODUCTION

The research presented here is part of a larger investigation of factors influencing vowel production by second-language (L2) speakers of English. The project was not originally motivated by any specific theoretical orientation on L2 production, but instead by pedagogical considerations. However, as will be shown, its relevance extends beyond that domain. Its central concern was the degree of uniformity in L2 vowel acquisition among learners sharing an L1 background. On the one hand, if very similar difficulties are experienced by many learners in a classroom setting, the workload of the pronunciation instructor is considerably lightened. Problem areas ought to be predictable in advance, and difficulties for most or all class members should be addressable with a common set of instructional activities, perhaps carried out in a lock-step fashion. On the other, a lack of uniformity suggests that a "one size fits all" approach to pronunciation teaching

is inadequate and that individual differences require detailed attention if instruction is to be effective. Logically, of course, it might turn out that some aspects of pronunciation learning do show relative uniformity (at least for speakers with a shared L1 background), but that others do not. The project therefore is not aimed at making broad generalizations about L2 segments and prosody, but instead focuses on one specific area of concern: English high vowel acquisition by Cantonese speakers. That focus is appropriate because differences in the two languages' vowel inventories appear to underpin known difficulties for Cantonese speakers (Meng et al., 2007; Wong, 2015). Also, as observed by Cebrian et al. (2021), the English high vowel contrast between /i/ and /ɪ/ has been the subject of much interest in L2 phonetics research, first because it has a high functional load in English, distinguishing many pairs of common words such as *heat* and *hit* (Levis and Cortes, 2008; Sewell, 2017). Second, it poses perceptual and productive difficulties for speakers from diverse L1 backgrounds, including Catalan, Mandarin, Russian, and Spanish (Mora and Fullana, 2007; Kondaurova and Francis, 2008; Munro and Derwing, 2008).

In addition to its pedagogical relevance, this examination of elicitation effects also offers theoretical promise in that it may provide insights into L2 acquisition mechanisms. In particular, evidence that one mode of elicitation yields better performance than the other would raise interesting questions about the processes involved in storing, retrieving, and implementing L2 phonological knowledge. Given the earlier set of findings, one relevant issue is whether individual L2 speakers show differential task effects. A comprehensive model of L2 phonetic learning would need to account for such variability.

The degree to which Cantonese and other learners of English diverge from one another in their success in high vowel production had received little attention until recently. Using speech elicited in a picture-naming task, Munro (2021) observed considerable interspeaker variability in the vowel intelligibility of Cantonese speakers when productions were considered in terms of vowels alone, vowels in rhymes, and even vowels within particular words. In this follow-up study, the Munro (2021) investigation is extended to compare the effect of two speech elicitation techniques, one with and one without an audio prompt, on interspeaker variability. The method of elicitation is important because accurately assessing learner difficulties is fundamental to both pedagogy and theory-building.

Factors Influencing L2 Segmental Production

Pronunciation specialists have devoted considerable attention to the wide range of factors that might predict in advance or explain in *post-hoc* fashion L2 learners' difficulties in producing particular consonants or vowels. Mid-twentieth-century authorities attempted to justify such work by claiming that predicting phonological difficulties can improve pedagogical practices (Moulton, 1962). Although that opinion was disputed long ago (Walz, 1980), interest among teachers in error prediction has persisted (Munro, 2018; Rehman et al., 2020). Recently, Munro et al. (2015) and Munro (2018, 2021) discussed evidence that individual variability in L2 production, even among speakers of a shared L1, is greater than has sometimes been assumed. Although it may be possible to offer broad, probabilistic error hierarchies for groups of learners from particular backgrounds, such predictions often do not apply to all, or even to the majority, of learners.

Influences on L2 pronunciation may be classified as linguistic or non-linguistic. By "linguistic," I mean those that relate specifically to one or both languages at issue. Chief among these is the degree of correspondence between the phonological systems of the languages (Lado, 1957), which is said to trigger "negative transfer" when structures differ. Although transfer effects are clearly an important influence on L2 segmental accuracy, a purely transfer-based account of errors is unsatisfactory, as discovered in early investigations of the Contrastive Analysis Hypothesis (Brière, 1966; Wardaugh, 1970). To some degree, this inadequacy may be due to faulty approaches to comparing language inventories (see Flege and Bohn, 2021; Thomson, 2021). Furthermore, different speakers may differ in their phonetic representations of the sound segments of L1, and may therefore relate L2 sounds to L1 sounds in idiosyncratic ways. Also, the assumption that "what is different is difficult" is undermined by research indicating that similarity, rather than difference, can pose serious problems in L2 phonetic learning (Flege, 1987).

From a theoretical standpoint, invoking "transfer" entails a lack of specificity about the underlying cognitive mechanisms involved in acquiring, storing, and activating phonetic knowledge. Some theorists have attempted to refine the transfer concept to incorporate other linguistic factors predictive of learning (see Archibald, 2021, for a review). A variety of proposals have been offered that integrate such concepts as markedness (Eckman, 1985; Major and Kim, 1996), language-specific constraint rankings (Lombardi, 2003), and feature geometry (Brown, 2000). The extent to which such approaches improve predictive success is not at issue here. Rather, the starting point for the current study is the evidence of considerable inter-speaker variability in phonetic acquisition. This is not generally a focus of linguistic modeling *per se*, and for the most part, such variability must be the result of something other than linguistic factors.

Non-linguistic factors are often highlighted in research on individual differences in L2 phonetic learning. These are independent of the specific languages at issue (for reviews see Piske et al., 2001; Mora, 2022). For instance, the age of L2 learning (AOL) correlates negatively with foreign accent ratings (Flege et al., 1995, 1999; Bylund et al., 2021). Aptitude and motivational factors are also implicated in phonetic learning success (Perrachione et al., 2011; Hu et al., 2013; Kissling, 2014; Nagle, 2018b), as has the quantity of L2 experience (see Flege and Bohn, 2021; Flege et al., 2021). These factors affect global aspects of L2 pronunciation such as accentedness and intelligibility and figure prominently in some theoretical approaches. In particular, the Speech Learning Model (SLM, Flege, 1995) and the Revised Speech Learning Model (SLM-r, Flege and Bohn, 2021) emphasize language experience effects. However, such influences are of much less importance (or are not useful at all) in predicting specific phonetic problems, such as vowel or consonant errors.

Munro et al. (1996), for example, observed a negative relationship between AOL and L2 vowel goodness in Italian speakers of English, with variable effects from vowel to vowel for both accuracy and intelligibility. Moreover, different speakers with approximately the same AOL, and similar L2 experience, varied in the number of the 11 target vowels they produced intelligibly, with some producing as few as six and others producing all 11 intelligibly. This led Munro et al. (1996, p. 332) to observe that "between-vowel effects did not occur uniformly for all, or even for a large majority, of the learners." Such variability, as well as the parallel lack of uniformity in Munro (2021), does not appear explicable in terms of aptitude, motivation, or quantity of general L2 experience. In particular, there is no obvious reason why such characteristics should lead one speaker to produce good exemplars of /ɪ/ in *hit*, but mostly unintelligible productions of the same vowel in *sit*, while another speaker shows the opposite pattern. Yet, just such disparities were seen in the study by Munro (2021), suggesting that learners' knowledge of particular lexical items plays a role in vowel production accuracy. In an investigation of Korean speakers' English productions, Baker and Trofimovich (2008) found an advantage for vowels in words of higher frequency and greater subjective familiarity among adult speakers. Acquired vowel knowledge may therefore depend on the quality, quantity, and timing of learners' encounters with particular words. Hypothetically, for instance, frequent experience with a word very early in the L2 acquisition process, when control over the pronunciation of L2 structures is limited, might yield a different learning outcome from exposure at a later time. Given that L2 phonological knowledge develops rapidly during the hypothesized Window of Maximal Opportunity at the first massive exposure to L2 (Derwing and Munro, 2015), a timing difference of weeks or even days may affect word learning in important ways.

Task Effects in L2 Production

Teachers, assessors, and researchers elicit L2 speech in a variety of ways, depending on their goals. When approximation of real-world language is paramount, extemporaneous and interactive production tasks are preferred, though a drawback of these is limited researcher control over phonetic content. In the assessment of segmental production–a focus of the present study–the elicitation task must yield enough usable exemplars of the target sound to allow satisfactory analysis. Possible tasks include simple repetition (Flege and Eefting, 1988), reading aloud, delayed repetition, in which the speaker reformulates an utterance with a target item (Flege et al., 1995, 1999; Munro and Derwing, 2008), picture naming (Flege and Davidian, 1984; Cebrian et al., 2021), or less-constrained tasks, such as timed picture descriptions (TPD), in which speakers are instructed to use particular target items while giving their descriptions (Saito and Munro, 2014).

Ample evidence shows that differences in elicitation methods can affect L2 production. In comparison with simple repetition, read-aloud tasks can lead to stronger orthographic influences on pronunciation (Bassetti and Atkinson, 2015), even in familiar words produced by experienced L2 users. It is worth noting as well that non-reading tasks do not necessarily eliminate orthographic effects because speakers may have developed internalized representations of words by assuming and practicing pronunciations based on spelling. In fact, even native speakers sometimes use "reading pronunciations" of orthographically opaque words like *epitome* and *blackguard*.

In immediate word repetition, the availability of an aural model appears to facilitate accurate production. Rojczyk (2013), for example, found that Polish speakers produced more native-like English /æ/ formants in immediate imitation than in list reading. Although that seems to suggest that repetition does not require speakers to access their own phonological category representations, Llompart and Reinisch (2019) have argued against such a view. In fact, they observed a close link between speakers' imitation performance and their perceptual capabilities. They also found, however, that imitated and read words containing a difficult vowel distinction for German learners of English differed noticeably in their acoustic properties. They attributed the lower accuracy in reading to "inaccurate non-native lexical representations" (p. 594), which were accessed during reading but not in repetition.

Saito and Munro (2014) performed acoustic analyses on word-initial English /ɹ/ produced by groups of Japanese speakers either living in Japan or residing in Canada for 1–12 months. Targets were elicited in word-reading, sentence-reading, and TPD conditions. In apparent contrast with the findings discussed above, native-like F3 values were found in word-reading, though this was true only for speaker groups with 5 or more months of residency, and the other acoustic dimensions (F2, F1 transition duration) remained unaffected by the task. The authors proposed that the difference was due to the speakers' use of a controlled mode of production during reading.

Also of potential relevance to task effects is research on auditory priming. Trofimovich and Gatbonton (2006) exposed listeners to spoken target words in a priming block followed by a distractor task and a subsequent speeded repetition task. The previously heard targets were produced more quickly than items not heard during the priming block. That outcome suggests that activation during priming persisted even after the distractor task and facilitated access to relevant lexical representations for production. Following on that study, Leong et al. (2021) investigated Mandarin speakers' productions of English tense and lax high vowels, noting that Mandarin makes no tense-lax distinction. Target words were elicited *via* an orthographic presentation on a screen and were primed with a recorded vowel production (three iterations) that either matched (congruent condition) or did not match (incongruent condition) the vowel in the CVC target. A listener-based assessment revealed that /i/ and /ɪ/ were produced with higher intelligibility in the congruent than in the incongruent condition, though a parallel finding was not obtained for the high back vowels. Because the same speakers showed some ability to distinguish tense and lax vowels perceptually, Leong et al. (2021) appealed to a perception-based explanation. In this case, priming with congruent vowels may have facilitated access to the correct perceptual representations for the targets. In fact, most theorizing about L2 speech learning assumes a relationship between perception and production, particularly given evidence that perceptual learning can lead

to more accurate production (Nagle, 2018a). Notably, however, the SLM-r (Flege and Bohn, 2021) has abandoned the original SLM's assumption that accurate perception is a precondition for accurate production, in favor of the view that the two co-evolve through bi-directional processes.

The Present Study

This study extends Munro (2021) with a parallel design involving the same participants, with the addition of a new variable: the elicitation method. The focus is the intelligibility of English high vowel productions of Cantonese speakers under two conditions: a picture-naming task (reported in Munro, 2021) and an interrupted repetition task.

Comparison of Cantonese and English Vowels

As a result of Hong Kong's historical status as a British Overseas Territory, English is one of its official languages, and a Hong Kong variety of English has emerged (Hung, 2000; Hansen Edwards, 2015; Sewell, 2022). The vast majority of Hong Kong residents speak Cantonese natively, with many having a native command of Hong Kong English (HKE) as well. At the same time, some residents are speakers of English as a second language, having grown up with little experience using English for social purposes or work-related communication. Speakers from the latter demographic who had immigrated to Canada were targeted for this investigation.

Whereas Western Canadian English (WCE) has four contrastive high vowels differing in advancement and "tenseness" (/i ɪ u ʊ/), Cantonese has /i/ and /u/, each with tense and lax allophonic variants. It thus differs from Mandarin, the focus of the Leong et al. (2021) study. Front /i/ is produced as [i] in open syllables, and before voiceless labials and alveolars, but as [ɪ] before voiceless velars (Zee, 1991; Chan and Li, 2000). For /u/ the parallel lax variant ([ʊ]) also occurs before velars. Voiced obstruents do not occur in syllable-final positions. Additionally, [ɪ] and [ʊ] are relatively lowered (Zee, 1991). Taken together, these facts indicate that the English rhymes /it/, /ut/, /ɪk/, and /ʊk/ have rough "matches" in Cantonese, but /ɪt/, /ʊt/, /ik/, and /uk/, along with all V+/d/ combinations, do not.

A simple transfer-based analysis would predict that the matching English rhymes should be easier for Cantonese speakers to acquire than the non-matching ones. However, Munro (2021) found extensive evidence to the contrary. First, Cantonese speakers' productions of several non-matching rhymes were, on average, more intelligible than those of the matching rhymes. That outcome was not particularly surprising, since the degree of "match" between L1 and L2 rhymes was by no means exact. Second, the speakers differed from one another in their success in producing the English tense-lax distinction. More intriguingly, they also differed in their success in producing identical VC rhymes in different words. For instance, some speakers produced the vowel in *sit* with high intelligibility, but not the vowel in *hit*; others showed the reverse pattern: high intelligibility for *hit* but not for *sit*. Still others produced both words with full intelligibility, and a fourth subset produced neither word intelligibly. Given the occurrence of all four patterns, the differences between words cannot be attributable to an effect of different initial consonants. An understanding of results such as these requires a close examination of how individuals vary from one another in their production capabilities.

As noted by a reviewer, an additional complication in the interpretation of the Munro (2021) results is that the speakers may have been exposed, to varying degrees, to HKE. The HKE vowel system is described by Hung (2000) as having seven phonemic monophthongs, with a neutralization of the high tense-lax distinction. As a result, speaker-participants may have heard words such as *hit* and *wood* modeled with tense vowels rather than the lax counterparts used in WCE. In such cases, not making a tense-lax distinction could be due partly or mainly to exposure to particular native HKE productions, rather than to L1–L2 transfer.

Design

In the previously published study, 18 Cantonese speakers produced multiple tokens of 31 English target words in a picture elicitation task. Targets were common real words with segmental VC combinations known to pose difficulty for Cantonese speakers. Some of the combinations approximately matched sequences occurring in Cantonese (e.g., /it/ and /ɪk/), while others did not (e.g., /ɪt/ and /ik/). Picture naming was used instead of word reading for two main reasons. First, it verified the speaker's knowledge of the target words and required speakers to access stored phonological knowledge in order to produce them. Second, it was expected to reduce orthographic influences that might be more evident in a reading task. Such effects might be particularly noticeable for the orthographically opaque contrast between /u/ (too, moon, boot) and /ʊ/ (took, look, book). Intelligibility was selected as the independent measure because of its status as the single most important index of communicative ability (Subtelny, 1977; Munro, 2011). Because this dimension of speech cannot be assessed directly with acoustic measures, judgments from trained listeners were obtained.

In this extension of the earlier work, an interrupted repetition task was added, in which speakers first heard an aural exemplar of the target word and were required to count aloud to 10 before producing it. This task was selected instead of immediate repetition so as to minimize the speaker's opportunity to access or "play back" an acoustic image from the short-term auditory store. Like the priming task used by Leong et al. (2021), the speaker heard a good exemplar of the target vowel prior to production. However, in this case, the entire target word was the stimulus, and no other means of elicitation were used (e.g., no on-screen orthographic presentation). It was expected that speakers would have to process the auditory input and recognize the word in order to recall it after counting. Doing so could facilitate the activation of perceptual representations which may be more difficult to access during picture naming without an aural prompt. Past work (Trofimovich and Gatbonton, 2006) indicates that priming effects can persist even after a distractor task. If so, then words elicited *via* interrupted repetition could be expected to have more intelligible vowels than those elicited *via* picture naming. This might indicate that L2 speakers have more phonological knowledge–developed through perceptual

experience–than they are necessarily able to exploit when producing L2 segments without an aural model.

Research Questions

A key research question arose from the practical issue of how elicitation methods affect L2 vowel productions: (1) Does (high) vowel intelligibility differ in word productions elicited with and without a preceding auditory model? A second question follows from the finding of large variability in the intelligibility of vowels elicited *via* picture naming in Munro (2021): (2) To what degree is the effect of the elicitation task consistent across targets and across individual speakers?

METHODS

Speakers

The L2 speakers (10 female; 8 male)–the same as those in Munro (2021) –were 18 Cantonese-speaking adults (M_{age} = 18 years; range: 15–25), who had been born and raised in Hong Kong and were residing in Canada at the time of the investigation [Mean length of residence (M_{LOR}) = 4.9 years; range = 0.75–6.9; Mean age of arrival in Canada (M_{AOA}) = 18 years; range = 15–25 years]. They were recruited *via* email and word-of-mouth, with the requirement that they self-identify as second-language speakers of English. All had grown up speaking Cantonese at home and all had studied English in grade school. However, none reported regular use of English for social purposes before immigration to Canada. More than half (n = 11) had received some ESL instruction in Canada on arrival. At the time of the study, all participants' English skills were advanced enough for them to be studying at English-speaking post-secondary institutions. On average, they reported using English 26% of the time in their day-to-day activities. For comparative purposes in the intelligibility assessment, recordings from two native speakers (1 female; 1 male) of General Canadian English (GCE) were also randomly selected from a database of speakers from a post-secondary student cohort. All speakers passed a pure-tone hearing screen (250–4,000 Hz) at 20dB_{HL}.

Speech Materials

The stimulus items, identical to those used by Munro (2021), were common English CV(C) words representing rhymes with and without "matching" analogs in Cantonese, as shown in **Table 1**. The particular targets were selected because they were likely to be known to the speakers, because they represented VC combinations either corresponding or not corresponding to sequences occurring in Cantonese, and because they could easily be represented visually for elicitation in a picture-naming task. Although several minimal pairs were included, the latter requirement made it impossible to create a fully balanced set.

TABLE 1 | Stimulus items according to syllable structure.

Vowel	Coda	Target words	"Matching" rhyme in Cantonese
/i/	#	*key, see, tea*	Yes
	/t/	*feet*, heat, *seat*	Yes
	/k/	*cheek, speak*	No
	/d/	*feed, read*	No
/ɪ/	/t/	hit, sit	No
	/k/	chick, kick, sick	Yes
	/d/	kid, lid	No
/u/	#	*Sue, two*	Yes
	/t/	boot, suit	Yes
	/k/	*Luke, tuque*	No
	/d/	food	No
/ʊ/	/t/	foot, put	No
	/k/	book, cook, look	Yes
	/d/	*good, wood*	No

Italicized words were excluded from the analyses because intelligibility for those items in PNam exceeded 80%.

Previous Picture Naming Task (PNam)

Details of the PNam task were reported in Munro (2021). The speakers viewed a randomized set of drawings presented individually on letter-size cards, each displaying a stimulus number and the first letter of the target word as a clue. During a practice and familiarization session, a research assistant presented the entire set of cards one at a time as the speaker guessed the target and produced the stimulus number and the target as follows: "Number __. The next word is __." When the guess was wrong, the speaker was instructed to "try again" and to make as many further attempts as necessary until the correct item was named. After the practice round, the assistant shuffled the cards and recorded productions of the full set. This step was repeated twice for a total of three recorded productions of each item. In case of any false starts or hesitations on the part of the speaker (<1% of cases), a further repetition was elicited. Distractor items were included at the beginning and end of each round to minimize the effects of list intonation. Inclusion of the stimulus numbers in the recordings facilitated later sorting and digital extraction of the stimulus words.

Interrupted Repetition Task (IntRep)

In the IntRep task, the stimuli – identical to those in PNam – were presented aurally in the frame "The next word is ___," *via* an audio recording produced by a male native speaker of GCE. As in PNam, items were randomized and presented in three rounds. The speakers were instructed to listen to the model sentence, count to ten orally, and then reformulate the model as "Now I say ___." A short (2-min) practice round was provided immediately before the first round of recording. The stimuli were presented *via* custom playback software, controlled by a research assistant, who monitored the performance of each speaker. In the event, that a production was missed or otherwise unusable (e.g., a false start or hesitation), the stimulus was replayed (< 1% of cases).

Recording Procedures

During the individual sessions, high-quality digital recordings (44.1 kHz; 16 bits quantization) were made in a sound-treated booth. Speakers wore a Shure Beta 54 head-mounted microphone connected to a Symmetrix 302 microphone preamplifier and an HHB Professional digital recorder (CDR-830). They completed

the speaking tasks with PNam preceding IntRep so that they were not exposed to aural models of the target words prior to PNam. As a result of this ordering, in advance of IntRep, the participants were fully familiar with the words, having produced them four times each in PNam, once during a practice round and three times during recording. A break between tasks was given for as long as each speaker desired (typically about 5 min), and drinking water was available as needed.

Listener-Judges

The same four linguistically trained assistants who had evaluated vowel intelligibility in Munro (2021) also judged the IntRep productions in this study. All judges had grown up in Canada in monolingual (Canadian English) households. Two had taught ESL extensively, one had studied Japanese and Korean, and the fourth had extensive experience listening to and measuring non-native speech as a lab assistant. All were familiar with IPA, and all passed the pure-tone hearing screen referenced above.

Token Extraction and Intelligibility Evaluation

After recording, the target word productions were digitally excised and saved as individual peak-normalized audio files. These were evaluated for intelligibility by the listener judges during multiple individual sessions held over several days. Each judge heard a different randomized presentation of the excised words through high-quality headphones and, on a computer screen, selected the symbol for the GCE vowel closest to the one they heard in each production. The available response choices were based on pre-screening of the tokens by the author so as to include vowels that were not actual targets, but were sometimes produced in error: / i ɪ eɪ ɛ u ʊ oʊ /. Inter-judge reliability was assessed on the basis of whether the judge assessed a production as on-target or not. Four-way agreement was found on 72% of items, with at least three of four judges agreeing on 92%. These rates are slightly higher than for Munro (2021) and compare favorably with rates in other L2 vowel studies (Munro and Derwing, 2008).

RESULTS

In Munro (2021), the PNam data were submitted to multiple analyses, including comparisons of mean performance on vowels, rhymes, and words. Because the present study has a more restricted focus than the earlier one–i.e., the effect of elicitation type on vowel intelligibility–the statistical analyses were selected so as to focus on questions relating specifically to that issue. The earlier study, which incorporated self-estimated English use and length of Canadian residence as co-variates, yielded non-significant effects of each, so these were not included here.

Preliminary Analysis

To begin, exploratory probing of the data was carried out to identify suitable directions for more detailed analyses. First, mean intelligibility scores for the two tasks were computed for each speaker by pooling the scores of the four judges over the 31 words for each speaker. A paired samples *t*-test indicated significantly higher scores for IntRep (M = 79.4, SE = 1.78) than for PNam (M = 72.9, SE = 1.99), t (17) = 5.265, $p < 0.001$, $d = 1.241$). The differences between task scores varied considerably across speakers. The maximum difference was 16 percentage points; however, two speakers showed a difference of only 1 point, and two others showed a small reversal, with PNam > IntRep by 2 and 3 percentage points.

TABLE 2 | Mean intelligibility by vowel and task (all targets).

Vowel	PNam mean (%)	SD	COV	IntRep mean (%)	SD	COV	COV diff ratio (%)
i	92.3	9.72	0.105	94.4	6.42	0.068	35
ɪ	50.7	22.67	0.447	59.1	18.87	0.319	29
u	82.2	12.26	0.149	93.1	7.56	0.081	46
ʊ	58.3	11.56	0.198	65	15.21	0.315	−18

Second, because IntRep always followed PNam, it might be proposed that the better performance on IntRep could simply be due to a greater amount of practice with the target items. If so, then one would expect intelligibility to increase over successive recorded trials for one or both tasks. Repeated measures ANOVAs on the intelligibility scores revealed no significant effects of Trial Number for either PNam, F (2, 34) = 0.515, $p = 0.602$ or IntRep, F (2, 34) = 0.024, $p = 0.977$. The fact that there was no evidence of improvement over time *within* tasks suggests that the between-task difference was likely due to the task itself rather than to a practice effect.

Next, a two-factor repeated measures ANOVA yielded significant effects of both Vowel, F (1.451, 24.67) = 45.669, $p < 0.001$(Greenhouse-Geisser, due to a violation of the sphericity assumption), η_p^2= 0.729 and Task F (1, 17) = 29.381, $p < 0.001$, $\eta_p^2 = 0.633$. Relevant means are given in **Table 2**. The interaction of Vowel and Task missed significance, F (2.085, 35.44) = 2.344, $p < 0.109$ (Greenhouse-Geisser), $\eta_p^2 = 0.121$. Despite the lack of significance, the latter effect size falls between medium and large (Cohen, 1988). At the request of a reviewer, coefficients of variation, which can be understood as "standardized" measures of variability are provided in **Table 2**. In general, these are considerably smaller for the tense vowels, suggesting greater precision of estimation of population means for those vowels. However, because of complications arising from inter-speaker differences to be discussed later, it is inadvisable to dwell at length on these outcomes.

An examination of vowel confusion data from the listeners revealed similar patterns to those seen in the study by Munro (2021) in that non-native-like lax targets were not always produced as their tense counterparts. On the one hand, 5% of /i/ targets were judged to be /ɪ/ with only 1% heard as something else. On the other, for /ɪ/ targets, 20% were heard as /i/, 18% as /e/, and 3% as others. Back, tense /u/ was heard as /ʊ/ in 5% of cases, and as /o/ or others in 2%, while /ʊ/ was heard as /u/ 11% of the time and as /o/ 23% of the time.

Finally, it was determined that some stimulus words did not need to be included in the statistical analyses because the performance was at or near the ceiling in the PNam task. In

fact, there was essentially no inter-speaker variability in many of the target items due to 100% accuracy. Given the preliminary analysis above, it was expected that scores on IntRep would also be at or near the ceiling for those items such that the inclusion of these data would grossly violate distributional assumptions for statistical modeling. The criterion for exclusion of words was set at a mean (across all speakers) of 80% or greater intelligibility on the PNam task, the same criterion used by Munro (2021) as an indicator that a particular L2 rhyme had been "acquired." On that basis, 15 words were excluded. For the excluded items, the mean intelligibility difference between tasks (IntRep%–PNam%) was relatively small, only +2.96 percentage points, compared with +10.4 points for the included words. For two excluded words, the IntRep score was slightly lower than the PNam score (−0.9 points for *seat* and −2.3 points for *wood*). Means determined by Task and Word are discussed in a later section. The group-based analyses that follow are based on items remaining after the exclusion, which will be referred to as the "included words." It should be noted that IntRep scores on five included words exceeded 80%. This was not expected to pose problems for the analyses because of considerable interspeaker variability on those items.

Statistical Modeling

Scores for the included words were evaluated with mixed-effects models in *JASP* (JASP Team, 2022), in which Speaker was treated as a random effect. Because some of the fixed effects were not independent, three separate models for fixed factors (Rhyme, Word, and Matching Status) had to be computed (Type III Sum of Squares, Kenward-Roger procedure). *Post-hoc* analyses were Bonferroni-adjusted *t*-tests. In all results, $p < 0.05$ was adopted as the level for significance.

Rhyme Analysis

In 13 of the 14 rhymes in the original PNam recordings, the mean intelligibility of all words associated with the rhyme was either consistently above ($n = 6$) or below ($n = 7$) the "acquired" criterion of 80%. In short, rhyming words tended to "behave" in the same way. In the case of /i/, however, two words reached the criterion (*feet* and *seat*), while the third (*heat*) did not. Because the inclusion of only one of the three items in the modeling would give a misleading picture of the task differences across rhymes, *heat* was omitted from the rhyme analysis, which was therefore based on the six rhymes in which no words reached the intelligibility criterion in PNam. For the mixed-effects ANOVA results in **Table 3**, Rhyme and Task were fixed factors. Including random slopes for speakers led to a singular fit, so these were not modeled. Both main effects reached significance, as did the two-way interaction. *Post-hoc* tests revealed that the Task effect was due to significantly higher vowel intelligibility on IntRep for /ud/ and /ut/, but the between-task differences for /ɪk/, /ɪt/, /ʊk/ and /ʊt/ were not significant despite a general trend toward higher intelligibility on IntRep. Means by rhyme and task are given in **Table 4**.

Word Analysis

Table 5 gives ANOVA results for the Word and Task analysis. Once again, including random slopes for speakers led to a singular fit, so these were removed. Both the Task and Word effects were significant, though the two-way interaction was not. The Task effect was due to higher scores on IntRep than

TABLE 3 | Mixed effects ANOVA (rhyme and task) for included words.

Effect*	*df*	*F*	*p*
Rhyme	6, 509	26.322	<0.001
Task	1, 509	31.091	<0.001
Rhyme * task	6, 509	2.191	0.043

**Speakers were entered as a random effect.*

TABLE 5 | Mixed effects ANOVA (word and task) for included words.

Effect*	*df*	*F*	*p*
Word	15, 527	16.772	<0.001
Task	1, 527	23.879	<0.001
Word * task	15, 527	1.270	0.216

**Speakers were entered as a random effect.*

TABLE 4 | Mean intelligibility by rhyme and task (included rhymes only).

Rhyme	PNam%	SE	95% CI		IntRep%	SE	95% CI		Between-task
			Lower	Upper			Lower	Upper	Difference
/ut/	69.9	5.23	59.7	80.2	84.3	5.23	74	94.5	14.4*
/ʊt/	68.5	5.23	58.3	78.7	78.8	5.23	68.5	89	10.3
/ud/	61.7	6.87	48.2	75.1	93.6	6.87	80.1	107	31.9*
/ɪd/	55.3	5.23	45.1	65.6	65.1	5.23	54.8	75.3	9.8
/ɪk/	52.0	4.55	43.1	60.9	51.9	4.55	42.9	60.8	−0.2
/ɪt/	44.1	5.23	33.9	54.4	63.9	5.23	53.7	74.1	19.8
/ʊk/	33.2	4.55	24.3	42.1	41.2	4.55	32.3	50.1	8.0

**$p_{bonf} < 0.05$.*

TABLE 6 | Mean intelligibility by word and task (included words only).

Word	PNam%	SE	95% CI		IntRep%	SE	95% CI		Between-task
			Lower	Upper			Lower	Upper	Difference*
heat	73.2	0.93	60.3	86.2	75.4	0.80	62.5	88.3	2.2
hit	41.6	0.95	28.6	54.5	61.6	0.84	48.7	74.5	20.1
sit	46.7	1.09	33.8	59.7	66.2	0.88	53.2	79.1	19.4
chick	40.7	0.89	27.8	53.7	47.3	0.78	34.3	60.2	6.6
kick	53.7	0.84	40.8	66.7	40.7	0.61	27.8	53.7	−13
sick	61.6	0.66	48.7	74.5	67.6	0.30	54.6	80.5	5.9
kid	75.4	1.07	62.5	88.4	82.9	0.79	70	95.8	7.4
lid	35.2	1.11	22.2	48.1	47.3	1.10	34.3	60.2	12.1
boot	66.2	0.62	53.3	79.2	81.4	0.70	68.5	94.4	15.2
suit	73.6	0.66	60.7	86.5	87.1	0.42	74.2	100	13.5
food	61.7	0.90	48.7	74.6	93.6	0.26	80.6	106.5	31.9
foot	60.1	0.90	47.2	73	69.9	1.01	57.0	82.9	9.8
put	76.9	0.61	64	89.8	87.6	0.49	74.7	100.5	10.7
book	36.1	0.78	23.1	49	41.2	0.52	28.3	54.2	5.2
cook	26.3	0.71	13.3	39.2	33.4	0.71	20.5	46.3	7.1
look	37.2	0.79	24.2	50.1	48.9	0.63	36	61.9	11.8

**IntRep% – PNam%.*

TABLE 7 | *Post-hoc* comparisons* of %-correct ID for words with identical rhymes.

Rhyme	PNam (Munro, 2021)	PNam and IntRep combined
/ɪt/	hit = sit	hit = sit
/ɪk/	[sick = kick] > chick	sick > [kick = chick]
/ut/	boot = suit	boot = suit
/ʊt/	put > foot	put > foot
/ʊk/	book = cook = look	book = [look > cook]
/ɪd/	kid > lid	kid > lid

**No difference shown by "="; significant difference ($p_{bonf} < 0.05$) shown by ">".*

TABLE 8 | Mixed effects ANOVA (matching status and task) for included words.

Effect*	*df*	*F*	*p*
Matching status	1, 17	5.923	0.026
Task	1, 538	19.915	<0.001
Matching status * task	1, 538	4.024	0.045

**Speakers were entered as a random effect.*

on PNam, as seen for the pairs of means shown in **Table 6**, though one reversal (*kick*) was observed. Given the absence of a significant interaction, *post-hoc* tests were carried out on the combined results of the two tasks to compare scores on words sharing a rhyme. The results, shown in **Table 7**, are similar, but not identical to the results for PNam alone, reported by Munro (2021). In particular, different patterns were seen for /ɪk/, in that for the combined results, the vowel of *sick* was significantly more intelligible than that of *kick* or *chick;* and for /ʊk/, in that the *look* vowel was more intelligible than the *cook* vowel.

Matching Status Analysis of Rhymes

Results of the third analysis, in which Matching Status (of rhymes) and Task were fixed effects, are given in **Table 8**. In this case, by-speaker random slopes were included for each matching condition. The effects of both Matching Status and Task were significant, with the two-way interaction marginally so. **Figure 1** illustrates these outcomes. *Post-hoc* analyses indicated that for IntRep, non-matching items were produced with higher intelligibility than matching ones. For PNam, no statistical difference between matching and non-matching items was observed.

Individual Speaker Performance

As noted earlier, speakers varied in the degree to which their vowel intelligibility differed across tasks. Intelligibility differences according to speaker are given in **Figure 2** (IntRep%–PNam%). Sixteen speakers showed higher mean scores for IntRep, with a difference of 2–28 percentage points, while 2 showed marginally lower scores (≤5 points). There was a small-to-moderate negative correlation (assessed non-parametrically because of uneven data distributions and small sample size) between the differences and the corresponding PNam scores (Spearman's $rho = -0.503$, $p < 0.033$), indicating that speakers with overall lower PNam scores tended to show greater differences on IntRep. This finding is unsurprising, since a lower score on PNam indicates "more room to improve" on the subsequent task. Despite these noteworthy differences between speakers, the mean intelligibility for individual speakers on PNam showed a moderately high correlation with scores on IntRep, $r\ (16) = 0.8$, $p < 0.001$, indicating that PNam scores predicted IntRep scores rather

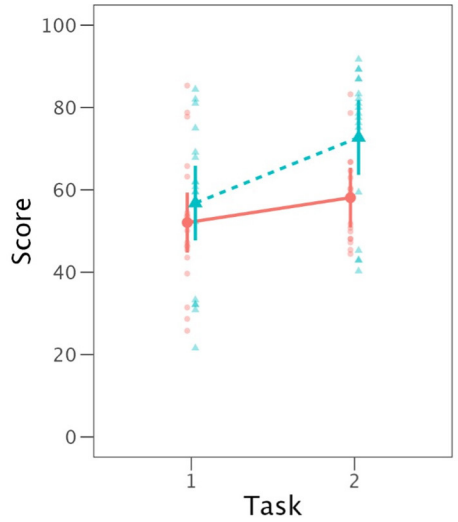


FIGURE 1 | Mean vowel intelligibility (with 95% CI) according to task (1 = PNam, 2 = IntRep) and matching status (solid line = match, dotted line = no match).

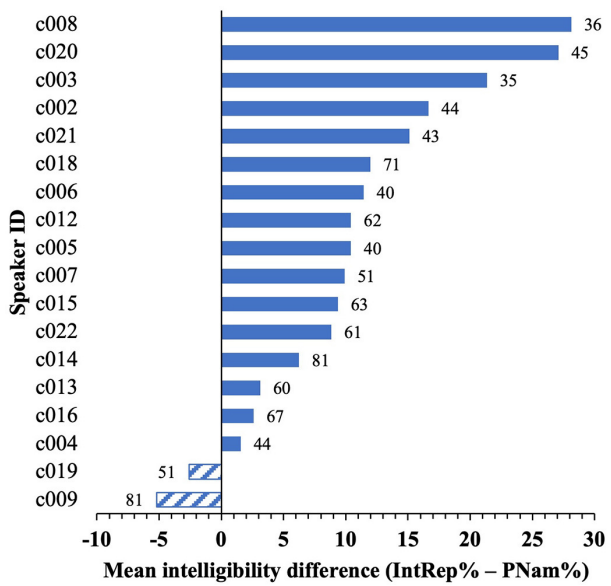


FIGURE 2 | By-speaker differences between mean intelligibility scores on the two tasks for the 16 included items. Data labels are the PNam% scores.

well (See the **Supplementary Materials** for scatterplots showing both patterns reported above). In response to a reviewer query, I stress that my use of the term "improvement" here and elsewhere refers only to the between-task difference. I am not assuming any sort of "learning" or permanent change in performance as a result of the simple exposure provided in IntRep.

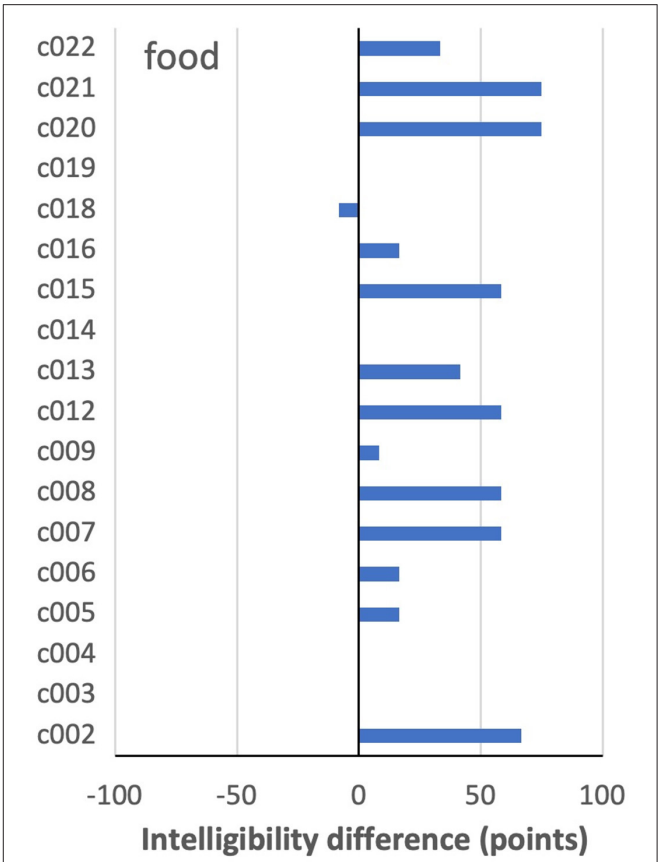


FIGURE 3 | By-speaker differences between mean intelligibility scores on the two tasks for *food*.

Individual Task Effects by Word

To extend Munro's (2021) examination of differences among speakers on words with the same rhyme, the task effect was explored for individual speakers through visual inspection of descriptive data on each word, along with figures comparing speakers. Because the nature of the data did not allow inferential statistics, a conservative criterion of a change of more than ±25 percentage points was adopted to assess whether a between-task difference on a particular target was of importance. Selected cases will be highlighted here because of the interesting findings they illustrate. For completeness, the figures for the full set of included words are provided in the **Supplementary Material**.

In 22% of the 288 word-by-speaker combinations (18 speakers and 16 words), intelligibility improved by more than 25 points in IntRep. In 6%, it dropped by more than 25 points, with the remaining majority (72%) showing <25 points difference.

The word showing the largest overall improvement on IntRep was *food* (62 vs. 94% intelligibility). As shown in **Figure 3**, nine speakers showed a difference of more than +25 points on IntRep. Of the remaining speakers, who showed a smaller between-task difference (<25 points), all but one scored above 83% on PNam, indicating a high level of performance and, therefore, little room for improvement. Thus, in the sense that nearly all who could

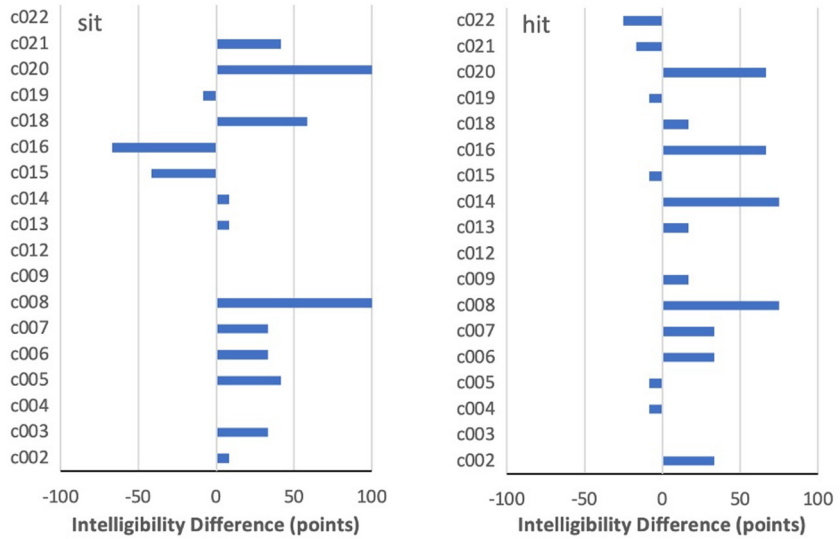


FIGURE 4 | By-speaker differences between mean intelligibility scores on the two tasks for *sit* and *hit*.

potentially show a task effect did so, *food* might be regarded as showing a relatively consistent effect across speakers.

For most other target words, however, there was much less consistency. Two interesting differences between PNam and IntRep are seen for *sit* in **Figure 4**, where c008 and c020 showed a complete change from 0 to 100% intelligibility. The same two speakers had also performed near 0% on *hit* in PNam, and, while both improved on that word in IntRep, the change was somewhat smaller (75 points and 67 points). In contrast, c004 and c019, who also performed at <10% on *sit* in PNam, showed virtually no task difference on *sit*. Moreover, among all speakers, improvement of 25 points or more on either *sit* or *hit* did not necessarily entail any improvement on the other word in the pair. This was true for c021, who improved on *sit* from 33 to 75%, but had surpassed the 80% threshold on *hit* in PNam. Particularly striking is c016, who improved by 67 points on *hit*, but declined by an equal amount on *sit*. Although four of the six speakers (c009, 12, 14, and 22) who performed above 80% on *sit* in PNam showed no meaningful difference between tasks, the two others (c015, 16) declined noticeably in IntRep.

Differences among the three /ʊk/ words are shown in **Figure 5**. These contrast to varying degrees with the comparatively consistent pattern for *food*. In general, the /ʊk/ targets ranked very low on intelligibility in PNam, with only two speakers scoring above 80% on *book*, two on *look*, and not a single speaker on *cook*. For the latter item, three speakers (PNam scores of 8, 50, and 17%) showed an increase of more than 25 points on IntRep, with most of the remainder showing little change. Given that seven of the non-changers had scored <25% on PNam, it is clear that "having room" to improve across tasks was not a good indicator that such improvement actually would occur. Two reversals of more than 25 points occurred for *cook*, with one for *book* and none for *look*. It is also worth noting that c016, who showed an increase of 33 points on *look*, showed a negligible difference on *cook*, despite having plenty of room to improve on that item from a PNam score of only 16%.

As noted earlier, in about 6% of word-by-speaker combinations, a decline in the intelligibility of more than 25 points was seen from PNam to IntRep. For the most part, these cases appeared to be randomly distributed across words, with one or two cases per word. The only exception was *kick*, shown in **Figure 6**. For that item, only one speaker (c005) improved by more than 25 points, with five speakers showing declines of at least that magnitude.

One final illustration of the inconsistency of the task effect can be seen in a comparison between c002 and c004, both of whom scored 44% on PNam and therefore fell into the lower half of the speaker cohort. For c002, there was an overall improvement on IntRep of 17 points, with increases of more than 25 points on three words (*hit*, *food*, *look*) and declines on none. In contrast, c004 had a net improvement of only 2 points, with an improvement of more than 25 points on only one word (*foot*), and a comparable decline on two words (*kick*, *boot*).

DISCUSSION

This study was designed to address two questions relevant to the production of English high vowels by Cantonese L2 speakers: (1) whether vowel intelligibility would differ on two elicitation tasks and (2) whether any observed task effect would be consistent across different production types and across different speakers. In general, the first question can be answered in the affirmative, which means intelligibility on the interrupted repetition task (IntRep) was higher by more than 10 percentage points than on picture naming (PNam). As for the second question, however, substantial evidence of several types of inconsistency was obtained. Consequently, the finding of an intelligibility benefit for IntRep must be interpreted with caution.

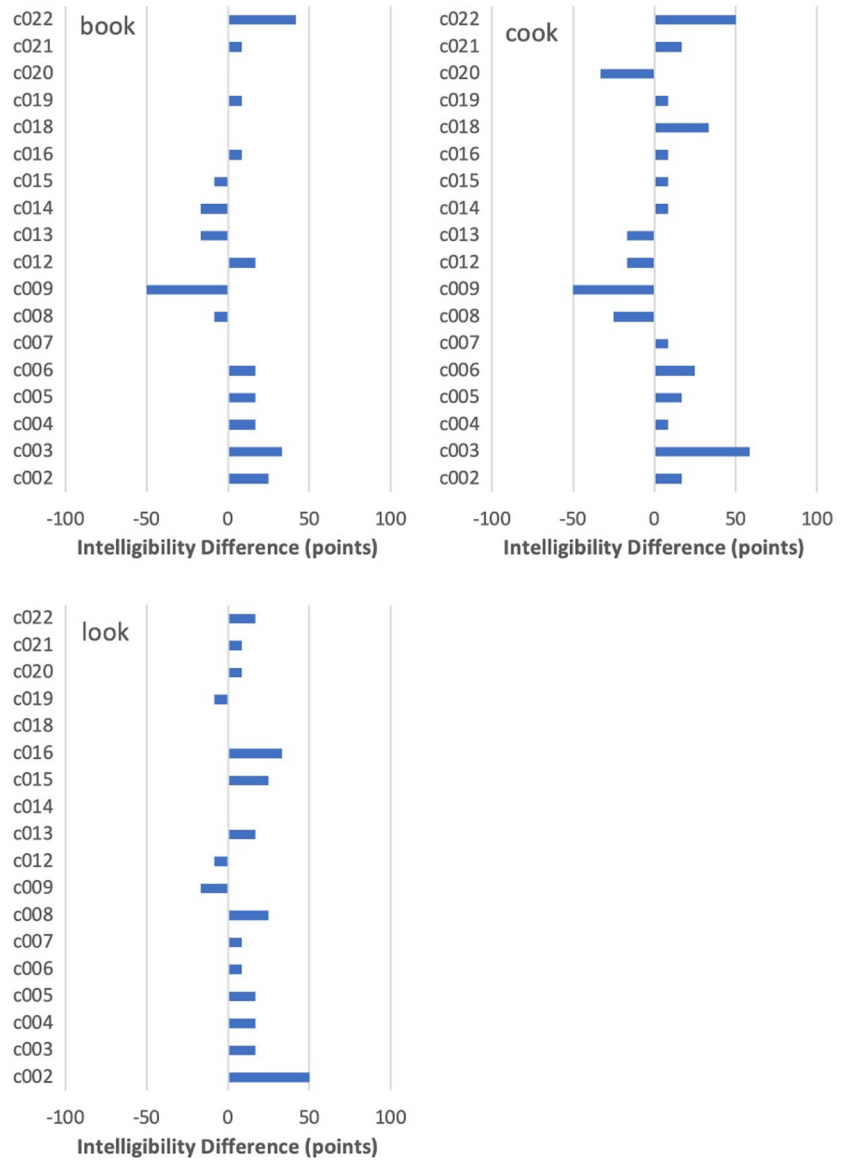


FIGURE 5 | By-speaker differences between mean intelligibility scores on the two tasks for *book, cook,* and *look*.

Why a Task Difference?

Before addressing these inconsistencies, it is appropriate to ask why the IntRep task should offer an intelligibility benefit to begin with. Findings indicate that in Munro (2021) and in the current study, vowel intelligibility was somewhat tied to lexical knowledge. Considering first the PNam task, note that some speakers produced certain targets with an intelligible vowel, but not other targets, even those sharing the same rhyme. Some speakers showed opposite patterns to other speakers, so that one speaker produced the *hit* vowel accurately but not the *sit* vowel, while another performed correctly on *sit*, but not on *hit*. To account for these inconsistencies, it is not possible to appeal exclusively to speakers' knowledge of vowels or rhyme-size units. Rather, the speakers evidently had knowledge of the motoric programs needed to produce /ʊ/ and /ʊk/ which they sometimes employed, but which did not necessarily transfer to all situations. In these cases, many inconsistencies appear to have been due to non-native-like lexical representations the speakers had developed, as in the study by Llompart and Reinisch (2019). However, the results of the PNam task gave no firm indication of how speakers would perform when given support in the form of an aural production of the target. Multiple explanations might be offered for why IntRep sometimes yielded better intelligibility. While this study was not designed to provide a definitive account, one possibility is that speakers sometimes had established more than one lexical representation for a word: one non-native-like, perhaps developed early in the acquisition process, and a competing, but more weakly

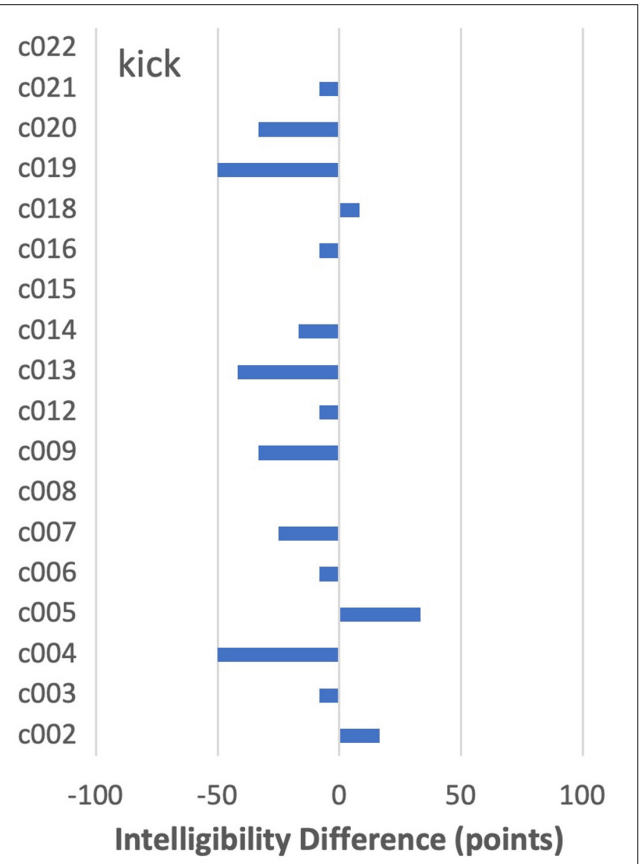


FIGURE 6 | By-speaker differences between mean intelligibility scores on the two tasks for *kick*.

established, representation developed through repeated exposure to a native or native-like target. A correct production might occur in IntRep if the aural prompt (a good exemplar) activated the native-like representation to a greater extent than the incorrect one. This is not implausible. Even native speakers, for instance, sometimes appear to have more than one representation for certain lexical items. Some L1 English speakers might, for instance, establish a non-standard lexical representation for *epitome* based on a "reading pronunciation," viz /'ɛpətoʊm/, and use that pronunciation when speaking extemporaneously. However, they may simultaneously hold another representation developed from hearing the word spoken aloud (/ɪ'pɪtəmi/) without awareness that the two representations correspond to a single word. Likewise, the availability of two different representations for *either*, one beginning with /i/, the other with /aɪ/), may result in varying pronunciations by the same speaker from one time to another. An alternative account is that the Cantonese speakers had only a single non-native-like lexical representation for a word, but the presentation of an aural model activated motoric programs for the correct rhyme in the target word, allowing the speaker to bypass the stored knowledge that was accessed during picture naming. Irrespective of the reasons for the task effect, the results suggest that PNam did not fully capture speakers' production capabilities. They often performed better when support was available in the form of an aural model. As discussed further below, that outcome may be relevant to interactive speaking contexts.

Variability in the Task Effect

One outcome of the study was that scores on 15 of the 16 target words were high (>80%) on both tasks. These items could therefore be classified as "easy" for the L2 speakers, and the importance of any observed difference in scores between tasks would be doubtful. For that reason, there was no reason to examine the production scores in detail, and they were excluded from most statistical analyses. They confirm a conjecture raised in the introduction that some targets would show relatively uniform patterns across speakers (in this case, uniform ease) while others would not.

For the remaining words, several types of variabilities were observed in the task effect. In the first place, the magnitude of the task-related difference was not uniform across different rhymes. For instance, the rhymes /ud/ and /ut/, both containing the tense back vowel, were produced more intelligibly, on average, in the IntRep condition than in PNam, but the same was not true for the rhymes with the lax vowels, /ɪd/, /ɪk/, /ɪt/, /ʊk/, and /ʊt/. This was not simply the result of a ceiling effect due to high levels of performance for PNam on the lax vowel rhymes, as is clear from the means in **Table 4**, which in all but one case fell below the means for the tense vowels. But even the finding of differences between rhymes requires qualification because, as explained below, it did not generalize across speakers.

Second, "matching" vs. "non-matching" between L1 and L2 VC rhymes was tied to different outcomes. In particular, the significant interaction between matching status and rhymes occurred because English rhymes with no analog in Cantonese were produced more intelligibly than matching rhymes *only* in the IntRep condition.

Third, some speakers showed much higher mean intelligibility on IntRep than on PNam–more than a 25-point difference–while many showed little or no difference between the two. Thus, it is clearly not true that all speakers showed a net benefit from the presentation of a spoken model prior to production. Although there was no indication that any speaker performed meaningfully worse on IntRep than on PNam, some word-by-speaker combinations did show sizeable declines, as will be discussed further below.

Fourth, although no significant word-by-task effect emerged, that outcome, which is based on group means, hides interesting between-speaker differences. One word (*food*) showed relative consistency across speakers in that virtually all speakers who had not approached the ceiling in PNam did show increased scores in IntRep; however, that consistency was not common among the target items. Overall, fewer than 25% of word-by-speaker combinations showed an increase of 25 percentage points or more. Not only did the magnitude of the task difference vary from speaker to speaker, but, it also varied on different test items for different speakers. For instance, speakers c008 and c004

both performed poorly on *sit* in PNam, but c008 improved to perfect performance in IntRep, while c004 showed no change at all. Although it is tempting to speculate about why these two speakers patterned differently, no firm conclusions can be drawn. Both speakers had enrolled in ESL classes on arrival in Canada and both reported low daily use of English (<10%). It may be relevant that c008 had been in Canada for somewhat longer than c004 (5.6 vs. 3.8 years for c004). However, LOR did not prove to be a successful predictor of performance in the study by Munro (2021). It was also noteworthy that individual speakers who improved on a particular word (e.g., *look*) did not necessarily improve on other items with the same VC rhyme (e.g., *cook*). This makes it unlikely that any demographic variable such as LOR or L2 use can serve as a straightforward predictor of success in production.

Finally, of the 288 word-by-speaker combinations, about 6% showed a decline of more than 25 points on IntRep. These generally occurred at a rate of once or twice per word, with other smaller declines being more common. In some cases, these reversals might simply reflect random variability in production commonly referred to as regression to the mean. However, an unusually high number of them (5) was observed for *kick*. One possibility is that the nature of the model stimulus played a role. Although model stimuli were screened for categorical accuracy by a research assistant, it is conceivable that, for unknown reasons, some speakers who produced a correct target in PNam did not correctly recognize the model word during the IntRep task and therefore did not match it to their (correct) lexical representation of *kick*.

CONCLUSION

Central to this research and to the study by Munro (2021) is the complex individual differences that emerged in the vowel intelligibility data. Although some individual variability in speech production is always attributable to "noise," many of the patterns seen here are at least partially systematic, and cannot be dismissed as uninteresting simply because no immediate explanation is available. Rather, it is essential to closely examine the nature of this variation to determine what insights it may yield into the L2 speech learning process. In fact, there is a growing awareness of the value of studying individual learning trajectories in other aspects of L2 speaking as evidenced by the increasingly high profile of Complex Dynamic Systems Theory in L2 research (Lowie and Verspoor, 2022).

The findings reported here are likely to be of interest to assessors, language teachers, and researchers because they show that the choice of elicitation method can affect the intelligibility of L2 speakers' vowel productions. On the one hand, picture naming may be a useful way of determining speakers' typical pronunciations in unaided situations. On the other, the outcome of such a task may not capture the full knowledge a speaker possesses about pronouncing target items. The availability of an auditory model, however, seems to facilitate speaker access to knowledge that is not activated in picture naming. The between-task difference in performance raises interesting questions about the benefits of speech learning of interactions in which an L2 speaker hears the productions of an interlocutor and responds using some of the same lexical items as the interlocutor. In such circumstances, the modeled pronunciation may serve a scaffolding function that can facilitate more accurate production by the listener. Although it is possible that this modeling can promote additional learning, the degree to which such learning (if any) actually occurs has not been assessed here and is a topic worthy of further investigation. A more detailed examination of the kinds of effects observed in this study may lead to enhancements of models of L2 speech production (Kormos, 2014) and acquisition (Flege and Bohn, 2021). For example, it may prove useful to probe acoustic data from L2 speakers' productions in IntRep and other types of repetition tasks to determine the degree to which phonetic convergence toward the pronunciation of the model speaker occurs. It is noteworthy as well that the benefit of modeling apparently does not require *immediate* repetition by the speaker. That finding is consistent with research showing that priming effects can persist well after the presentation of the prime itself (Trofimovich and Gatbonton, 2006).

From the standpoint of phonetic learning research, a significant finding of Munro (2021) was that intelligible L2 production of a vowel in a particular word does not predict that the vowel will be produced correctly in other words, even those with the same post-vocalic environment. Rather, accurate pronunciation is somewhat linked to word learning. The present study adds a new complication to that finding in that the effect of presenting an aural model during elicitation is not uniform across targets or speakers. Although some speakers benefit considerably from such an approach, others do not, or they show the benefit on different targets. Consequently, accurate evaluation of segmental difficulties and strengths requires a more sophisticated approach than elicitation of a small number of target words representing the segments of interest.

DATA AVAILABILITY STATEMENT

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

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethics, Simon Fraser University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

This research was funded by a grant from the Social Sciences and Humanities Research Council of

Canada. Open access publication was funded by Simon Fraser University.

ACKNOWLEDGMENTS

I thank Herman Li, Susan Morton, and Natasha Penner for their important contributions to data collection. Also appreciated are many discussions with my colleagues Tracey Derwing and Ron Thomson. All errors are my own.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.916736/full#supplementary-material

REFERENCES

Archibald, J. (2021). Ease and difficulty in L2 phonology: a mini-review. *Front. Commun.* 6, 18. doi: 10.3389/fcomm.2021.626529

Baker, W., and Trofimovich, P. (2008). Lexical and segmental influences on child and adult learners' production of second language vowels. *COPAL Concordia Work. Pap. Appl. Linguist.* 1, 31–54.

Bassetti, B., and Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Appl. Psycholinguist.* 36, 67–91. doi: 10.1017/S0142716414000435

Brière, E. J. (1966). An investigation of phonological interference. *Language* 42, 768–796. doi: 10.2307/411832

Brown, C. (2000). "The interrelation between speech perception and phonological acquisition from infant to adult," in *Second Language Acquisition and Linguistic Theory*, ed J. Archibald (Oxford: Wiley-Blackwell), 4–64.

Bylund, E., Hyltenstam, K., and Abrahamsson, N. (2021). Age of acquisition–not bilingualism–is the primary determinant of less than nativelike L2 ultimate attainment. *Biling. Lang. Cogn.* 24, 18–30. doi: 10.1017/S1366728920000188

Cebrian, J., Gorba, C., and Gavaldà, N. (2021). When the easy becomes difficult: factors affecting the acquisition of the English /i:/-/ɪ/Contrast. *Front. Commun.* 6. doi: 10.3389/fcomm.2021.660917

Chan, A. Y., and Li, D. C. (2000). English and Cantonese phonology in contrast: explaining Cantonese ESL learners' English pronunciation problems. *Lang. Cult. Curric.* 3, 67–85. doi: 10.1080/07908310008666590

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn*, New York: Erlbaum.

Derwing, T. M., and Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research.* Amsterdam: John Benjamins.

Eckman, F. R. (1985). Some theoretical and pedagogical implications of the markedness differential hypothesis. *Stud. Second Lang. Acquis.* 7, 289–307. doi: 10.1017/S0272263100005544

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *J. Phonetics* 15, 47–65. doi: 10.1016/S0095-4470(19)30537-6

Flege, J. E. (1995). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, eds W. Strange (Baltimore: York Press), 233–277.

Flege, J. E., Aoyama, K., and Bohn, O. S. (2021). "The revised speech learning model (SLM-r) applied," in *Second Language Speech Learning: Theoretical and Empirical Progress*, eds N. R. Wayland (Cambridge: Cambridge University Press), 84–118. doi: 10.1017/9781108886901.003

Flege, J. E., and Bohn, O. S. (2021). "The revised speech learning model (SLM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress*, ed R. Wayland (Cambridge: Cambridge University Press), 3–83.

Flege, J. E., and Davidian, R. D. (1984). Transfer and developmental processes in adult foreign language speech production. *Appl. Psycholinguist.* 5, 323–347. doi: 10.1017/S014271640000521X

Flege, J. E., and Eefting, W. (1988). Imitation of a VOT continuum by native speakers of English and Spanish: evidence for phonetic category formation. *J. Acoust. Soc. Am.* 83, 729–740. doi: 10.1121/1.396115

Flege, J. E., Munro, M. J., and MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *J. Acoust. Soc. Am.* 97, 3125–3134. doi: 10.1121/1.413041

Flege, J. E., Yeni-Komshian, G. H., and Liu, S. (1999). Age constraints on second-language acquisition. *J. Mem. Lang.* 41, 78–104. doi: 10.1006/jmla.1999.2638

Hansen Edwards, J. G. (2015). Hong Kong English: attitudes, identity, and use. *Asian Engl.* 17, 184–208. doi: 10.1080/13488678.2015.1049840

Hu, X., Ackermann, H., Martin, J. A., Erb, M., Winkler, S., and Reiterer, S. M. (2013). Language aptitude for pronunciation in advanced second language (L2) learners: behavioural predictors and neural substrates. *Brain Lang.* 127, 366–376. doi: 10.1016/j.bandl.2012.11.006

Hung, T. T. N. (2000). Towards a phonology of Hong Kong English. *World Englishes* 19, 337–356 doi: 10.1111/1467-971X.00183

JASP Team (2022). *JASP* (Version 0.16.1) [Computer software]. Available online at: https://jasp-stats.org.

Kissling, E. M. (2014). What predicts the effectiveness of foreign-language pronunciation instruction? Investigating the role of perception and other individual differences. *Can. Mod. Lang. Rev.* 70, 532–558. doi: 10.3138/cmlr.2161

Kondaurova, M. V., and Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *J. Acoust. Soc. Am.* 124, 3959–3971. doi: 10.1121/1.2999341

Kormos, J. (2014). *Speech Production and Second Language Acquisition*. New York: Routledge.

Lado R. (1957). *Linguistics Across Cultures*. Ann Arbor, MI: University of Michigan Press.

Leong, L., Heift, T., and Wang, Y. (2021). "Auditory priming effects on the pronunciation of second language speech sounds," in *Second Language Speech Learning: Theoretical and Empirical Progress*, ed R. Wayland (Cambridge: Cambridge University Press), 439–462.

Levis, J., and Cortes, V. (2008). "Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching," in *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*, eds C. A. Chappelle, Y. R Ching, and J. Xu (Ames, IA: Iowa State University), 197–208.

Llompart, M., and Reinisch, E. (2019). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Lang. Speech* 62, 594–622. doi: 10.1177/0023830918803978

Lombardi, L. (2003). Second language data and constraints on manner: explaining substitutions for the English interdentals. *Second Lang. Res.* 19, 225–250. doi: 10.1177/026765830301900304

Lowie, W., and Verspoor, M. (2022). "A Complex Dynamic Systems Theory perspective on speaking in second language development," in *The Routledge Handbook of Second Language Acquisition and Speaking*, eds T. M. Derwing, M. J. Munro, and R. I. Thomson (New York: Routledge), 39–53.

Major, R. C., and Kim, E. (1996). The similarity differential rate hypothesis. *Lang. Learn.* 46, 465–496. doi: 10.1111/j.1467-1770.1996.tb01244.x

Meng, H., Zee, E., and Lee, W. S. (2007). *Technical Report: A contrastive phonetic study between Cantonese and English to predict salient mispronunciations by Cantonese learners of English.* Hong Kong, Department of Systems Engineering & Engineering Management, The Chinese University of Hong Kong. Available online at: https://www1.se.cuhk.edu.hk/$\sim$hccl/publications/pub/Meng-Tech-Report-Feb2007.pdf.

Mora, J. C. (2022). "Aptitude and individual differences," in *The Routledge Handbook of Second Language Acquisition and Speaking*, eds T. M. Derwing, M. J. Munro and R. I. Thomson (New York: Routledge), 68–82.

Mora, J. C., and Fullana, N. (2007). "Production and perception of English /i:/-/ɪ/ and/æ/-/ʌ/in a formal setting: investigating the effects of experience and starting age," in *Proceedings of the 16th International Congress of*

Phonetic Sciences, eds J. Trouvain and W. Barry (Saarbrücken: Universität des Saarlandes), 1613–1616.

Moulton, W. G. (1962). Toward a classification of pronunciation errors. *Mod. Lang. J.* 46, 101–109. doi: 10.1111/j.1540-4781.1962.tb01773.x

Munro, M. J. (2011). "Intelligibility: buzzword or buzzworthy?" in *Proceedings of the 2rd Pronunciation in Second Language Learning and Teaching Conference*, eds J. Levis and K. LeVelle (Ames, IA: Iowa State University), 7–16.

Munro, M. J. (2018). How well can we predict L2 learners' pronunciation difficulties? *Catesol J.* 30, 267–281.

Munro, M. J. (2021). On the difficulty of defining "difficult" in second-language vowel acquisition. *Front. Commun.* 6. doi: 10.3389/fcomm.2021.639398

Munro, M. J., and Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: a longitudinal study of vowel production. *Lang. Learn.* 58, 479–502. doi: 10.1111/j.1467-9922.2008.00448.x

Munro, M. J., Derwing, T. M., and Thomson, R. I. (2015). Setting segmental priorities for English learners: evidence from a longitudinal study. Int. *Rev. Appl. Linguist. Lang. Teach.* 53, 39–60. doi: 10.1515/iral-2015-0002

Munro, M. J., Flege, J. E., and MacKay, I. R. (1996). The effects of age of second language learning on the production of English vowels. *Appl. Psycholinguist.* 17, 313–334. doi: 10.1017/S0142716400007967

Nagle, C. (2018b). Motivation, comprehensibility, and accentedness in L2 Spanish: investigating motivation as a time-varying predictor of pronunciation development. *Mod. Lang. J.* 102, 199–217. doi: 10.1111/modl.12461

Nagle, C. L. (2018a). Examining the temporal structure of the perception–production link in second language acquisition: a longitudinal study. *Lang. Learn.* 68, 234–270. doi: 10.1111/lang.12275

Perrachione, T. K., Lee, J., Ha, L. Y., and Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *J. Acoust. Soc. Am.* 130, 461–472. doi: 10.1121/1.3593366

Piske, T., MacKay, I. R., and Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *J. Phon.* 29, 191–215. doi: 10.1006/jpho.2001.0134

Rehman, I., Silpachai, A., Levis, J., Zhao, G., and Gutierrez-Osuna, R. (2020). The English pronunciation of Arabic speakers: a data-driven approach to segmental error identification. *Lang. Teach. Res.* doi: 10.1177/1362168820931888

Rojczyk, A. (2013). "Phonetic imitation of L2 vowels in a rapid shadowing task," in *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*, eds J. Levis and K. LeVelle (Ames, IA: Iowa State University), 66–76.

Saito, K., and Munro, M. J. (2014). The early phase of /ɹ/ production development in adult Japanese learners of English. *Lang Speech.* 57, 451–469. doi: 10.1177/0023830913513206

Sewell, A. (2017). Functional load revisited: reinterpreting the findings of 'lingua franca' intelligibility studies. *J. Second Lang. Pronunciation* 3, 57–79. doi: 10.1075/jslp.3.1.03sew

Sewell, A. (2022). The Hong Kong English accent continuum: insights from implicational scaling. *Asian Englishes*. doi: 10.1080/13488678.2021.2004492

Subtelny, J. D. (1977). "Assessment of speech with implications for training," in *Childhood Deafness: Causation, Assessment, and Management*, ed F. H. Bess (Grune and Stratton), 183–194.

Thomson, R. I. (2021). Commentary: When the easy becomes difficult: factors affecting the acquisition of the English /i:/-/ɪ/ Contrast and On the difficulty of defining "difficult" in second-language vowel acquisition. *Front. Commun.* 6. doi: 10.3389/fcomm.2021.748991

Trofimovich, P., and Gatbonton, E. (2006). Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *Mod. Lang. J.* 90, 519–535. doi: 10.1111/j.1540-4781.2006.00464.x

Walz, J. (1980). An empirical study of pronunciation errors in French. *French Rev.* 53, 424–432.

Wardaugh, R. (1970). The contrastive analysis approach. *TESOL Q.* 4, 123–130. doi: 10.2307/3586182

Wong, J. W. S. (2015). "Comparing the perceptual training effects on the perception and production of English high-front and and high-back vowel contrasts by Cantonese ESL learners," in *Proceedings of the 18th International Congress of Phonetic Sciences*, ed The Scottish Consortium for ICPhS. (Glasgow: University of Glasgow).

Zee, E. (1991). Chinese (Hong Kong Cantonese). *J. Int. Phon. Assoc.* 21, 46–48. doi: 10.1017/S0025100300006058

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

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 16 August 2022
DOI 10.3389/fpsyg.2022.941114

OPEN ACCESS

EDITED BY
Boping Yuan,
University of Cambridge,
United Kingdom

REVIEWED BY
Wei Ren,
Beihang University, China
Natasha Yuhsin Huang,
Tunghai University, Taiwan

*CORRESPONDENCE
Guangsa Jin
g.jin@uibe.edu.cn
Sai Ma
masaibh@126.com

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

RECEIVED 11 May 2022
ACCEPTED 15 July 2022
PUBLISHED 16 August 2022

CITATION
Ma S and Jin G (2022) The relationship between different types of co-speech gestures and L2 speech performance. *Front. Psychol.* 13:941114. doi: 10.3389/fpsyg.2022.941114

The relationship between different types of co-speech gestures and L2 speech performance

Sai Ma[1]* and Guangsa Jin[2]*

[1]Department of English Education, College of Foreign Languages, Capital Normal University, Beijing, China, [2]Department of Linguistics, School of International Studies, University of International Business and Economics, Beijing, China

Co-speech gestures are closely connected to speech, but little attention has been paid to the associations between gesture and L2 speech performance. This study explored the associations between four types of co-speech gestures (namely, iconics, metaphorics, deictics, and beats) and the meaning, form, and discourse dimensions of L2 speech performance. Gesture and speech data were collected by asking 61 lower-intermediate English learners whose first language is Chinese to retell a cartoon clip. Results showed that all the four types of co-speech gestures had positive associations with meaning and discourse L2 speech measures but no association with form-related speech measures, except the positive association between metaphorics and the percentage of error-free clauses. The findings suggest that co-speech gestures may have a tighter connection with meaning construction in producing L2 speech.

KEYWORDS

gesture, L2 speech, meaning, form, discourse, representational, deictic, beat

Introduction

The widely accepted dimensions to evaluate L2 speech performance include complexity, accuracy, and fluency (CAF), among which complexity gauges to what degree speakers can use complex language forms, accuracy measures to what extent speakers produce errorless language forms, and fluency examines to what degree the speech is produced without unnecessary pauses (Skehan, 1998, 2009, 2014; Ellis and Barkhuizen, 2005). Since complexity and accuracy measures are associated with rule-based language knowledge, they represent the quality of language form; in contrast, fluency measures are linked to how quickly speakers convey their ideas, and thus they represent the quality of meaning (Skehan, 1998; Ellis and Barkhuizen, 2005).

L2 speech performance in terms of CAF has been found to interact with individual differences such as personality and anxiety (Oya et al., 2004), working memory (Mota, 2005), willingness to communicate (Nematizadeh and Wood, 2019), and sociocultural attitudes toward target language and culture (Sun, 2022), but few L2 speech studies have examined its association with co-speech gesturing, or specifically, the movements of the hand and arm during speech. Noteworthily, increasingly more researchers agree that speech and gesture are two different but closely related modalities to express thoughts (McNeill, 1992; Goldin-Meadow, 2003; Kendon, 2004), and their close associations have been empirically supported in terms of occurring time (Church et al., 2014), language development (Iverson and Goldin-Meadow, 2005; Vilà-Giménez et al., 2021), semantic content (Kita and Özyürek, 2003), pragmatic functions (Loehr, 2012), etc. Many relevant studies focused on native and bilingual speakers (Nicoladis et al., 2009; Smithson et al., 2011; Laurent et al., 2015), and the associations between speech and gesture use by L2 learners is under-researched.

This study complements the investigation of individual differences linked to L2 speech performance and the relationship between gesture and speech by exploring the associations between four types of co-speech gesture use and meaning, form and discourse aspects of L2 speech performance. In the remaining of this section, we first introduce the cognitive functions of co-speech gestures in speech production, and then review relevant studies on the associations between speech performance and co-speech gestures.

Co-speech gestures include iconics, metaphorics, deictics, and beats (McNeill, 1992). Iconics express concrete concepts by mimicking their size, shape, contour, etc.; metaphorics represent abstract concepts with concrete imageries created by movements of the hand and arm; deictics are pointing gestures that refer to an entity by the extending of the index finger, hand, or arm; and beats are biphasic up-down movements of the finger, hand, or arm (McNeill, 1992; Cartmill and Goldin-Meadow, 2016). Iconics and metaphorics are typical representational gestures that have a referential relationship with speech content.

Co-speech gestures, both on the whole and for each individual type, have been shown to be cognitively beneficial to speech. Working memory is a factor modulating and restricting L2 speech (Kormos, 2006; Weissheimer and Mota, 2009; Skehan, 2014), and co-speech gestures have been found to be able to reduce working memory load. For example, co-speech gestures can maintain mental imageries in memory and thus function to offload working memory burden during speech (Wesp et al., 2001). This function has also been supported by studies using the speech-memory dual task paradigm, in which participants memorized more items when they were allowed to gesture during the speech task than when they were prohibited to do so (Goldin-Meadow et al., 2001; Wagner et al., 2004; Ping and Goldin-Meadow, 2010; Cook et al., 2012), and the effect was especially obvious when participants' working memory capacity was low (Marstaller and Burianová, 2013).

Different types of co-speech gestures are also intimately linked to speech in terms of cognitive functions. Bearing a close relationship with speech content, representational gestures have been found to be beneficial cognitively for speech production. For instance, being embodied and three-dimensional, representational gestures can activate the image of concepts in mind and divide them into expressible units and thus help to scaffold speech content (Alibali et al., 2000; Kita and Davies, 2009; Chu et al., 2014). Meanwhile, using representational gestures can stimulate the linguistic representation of concepts in a cross-modal way, and thus helps speakers to retrieve words (Rauscher et al., 1996; Krauss et al., 2000; Frick-Horbury, 2002). Furthermore, speakers' working memory capacity has been found to be negatively associated with the use of representational gestures, suggesting that such gestures are strategies to compensate for the shortage of cognitive resources (Chu et al., 2014; Gillespie et al., 2014).

With either iconic gestures or iconic and metaphoric gestures together as the study object, the aforementioned studies on representational gestures failed to show the cognitive functions of metaphoric gestures. It has been proposed that both iconic and metaphoric gestures should be equally helpful in constructing concepts and speech (Kita et al., 2017). This proposal is reasonable in that, first, iconics and metaphorics have similar generating mechanisms, that is, both of them are based on the schematization of concepts, with iconics schematizing concrete concepts and metaphorics abstract ones (McNeill, 1992; Cienki and Müller, 2008; Chui, 2011; Burns et al., 2019); second, metaphors are conceptual in essence, so that people are very adept at producing and comprehending metaphorical gestures and speech, and usually employ metaphorical mappings unconsciously and effortlessly (Lakoff and Johnson, 1980; Cienki and Müller, 2008). Therefore, iconics and metaphorics may be similar in helping speech production. However, more empirical studies are in need.

Little research has examined the direct cognitive advantage of using deictics in speech, but deictic gesturing has been found to serve intrapersonal cognitive functions which are very likely to influence speech, such as lowering cognitive load (Ginns and King, 2021; Wang et al., 2022) and regulating attention (Delgado et al., 2009; Korbach et al., 2020). Empirical studies supporting such claims found that performing pointing gestures made learners do better in learning tasks (Hu et al., 2014, 2015; Agostinho et al., 2015; Ginns et al., 2016; Korbach et al., 2020; Ginns and King, 2021; Wang et al., 2022). Furthermore, tasks that require more cognitive effort, such as the verbal improvisation task, made participants generate more deictics than unchallenging tasks, such as the ordinary verbal task, indicating that deictic gesturing is a strategy to lower cognitive

load (Lewis et al., 2015). It is also worth mentioning that in studies that have shown gestures' function of lowering cognitive load using the speech-memory dual task paradigm, a large percentage of gestures used in the gesture-allowed condition were deictic ones (Goldin-Meadow et al., 2001; Wagner et al., 2004; Ping and Goldin-Meadow, 2010; Cook et al., 2012).

Contrary to representational and deictic gestures, beats have little relationship with speech content but are associated with discourse features of the accompanying speech (McNeill, 1992; Dimitrova et al., 2016; Shattuck-Hufnagel and Ren, 2018). Beat gesturing has been found to be beneficial to solving tip-of-the-tongue problems (Ravizza, 2003), retrieving low-frequency words (Lucero et al., 2014), encoding and recalling foreign words (Morett, 2014), and enhancing children's narrative structure and fluency while narrating stories (Vilà-Giménez and Prieto, 2020). These studies show that beats are helpful cognitively in accessing lexicons, structuring narration, and improving fluency.

We next review literature on the associations between co-speech gestures and speech performance. Gesturers tend to do better than non-gesturers in fluency, lexical richness, and speech length for both native and bilingual speakers. For example, gesture restriction worsened fluency measures for English native speakers (Rauscher et al., 1996) and reduced the number of word types, word tokens, and scenes for bilinguals (Laurent and Nicoladis, 2015); in addition, gesturers spoke faster and produced more word tokens than non-gesturers for bilinguals (Smithson et al., 2011).

Individual types of co-speech gestures are associated with speech performance in different ways. Iconics, which have drawn the most scholarly attention, have been observed to relate to speech performance either positively or neutrally. Studies have shown that the iconic gesture rate was positively correlated with the number of word types for native French speakers (Nicoladis et al., 2009), with the number of word tokens for native English speakers (Smithson et al., 2011), and with the number of scenes (Laurent et al., 2015) and the length of utterances (Nicoladis et al., 1999) for bilingual kids; but it had no relationship with the number of word tokens for Chinese-English or French-English bilinguals (Smithson et al., 2011), with the number of word types for English native speakers (Nicoladis et al., 2009) or English-French bilinguals (Nicoladis et al., 2009; Laurent et al., 2015), or with the speech rate for either native or bilingual speakers (Smithson et al., 2011). In addition, the number of word tokens could predict the iconic gesture rate for Spanish, English, and French native speakers (Nicoladis et al., 2016), but the speech rate could not predict it for bilinguals (Aziz and Nicoladis, 2019).

Most of the above studies invited either native speakers or highly proficient L2 speakers to complete a cartoon-retelling task, which might be a reason for the mixed results. Both native and highly proficient speakers are less likely to find speech tasks challenging, in which case their available working memory resources are adequate for the task and their gestures are mainly a reflection of their speaking styles (Nagpal et al., 2011). On the other hand, less proficient L2 speakers' gesture use has closer association with speech since they need strategies like gesturing to compensate for the shortage of working memory resources during speaking (Nicoladis et al., 2007; Aziz and Nicoladis, 2019). For example, Nicoladis et al. (2007) found that the positive correlation between the iconic rate and the number of scenes described only held for intermediate Chinese English learners but not for native Chinese speakers. Unfortunately, little attention has been paid to the relationship between gesture and speech produced by L2 learners except for the study conducted by Nicoladis et al. (2007) and Ma et al. (2021). Ma et al. (2021) found that, for both concrete and abstract speech content, lower-intermediate L2 learners' representational gesture use was positively associated with speech fluency measures and speech length.

Research specifically investigating the relationship between metaphoric gestures and speech performance has been rare. It is necessary to verify the proposal that metaphorics should bear similar cognitive functions to iconics in speech production (Kita et al., 2017) through exploring the association between metaphoric gesturing and speech performance.

The relationship between using deictics and speech performance has been scarcely examined. Nicoladis (2002) found that the number of deictics was positively associated with the number of utterances for French-English bilingual preschoolers. Several studies concluded that the associations of speech with iconics and beats differ from its association with deictics in that the use of iconics and beats develops with speech proficiency whereas deictic gesturing compensates for weak language proficiency (Nicoladis et al., 1999; Mayberry and Nicoladis, 2000; Nicoladis, 2002; Gan and Davison, 2011; Lin, 2019). This conclusion is supported by findings that iconic and beat gesture use bore stronger correlations with speech measures than deictic gesture use did (Nicoladis et al., 1999; Mayberry and Nicoladis, 2000), that bilinguals used more iconics instead of deictics in their dominant language (Nicoladis, 2002), that L2 learners used more deictics rather than iconics in L2 discussion (Gan and Davison, 2011), and that L2 learners with higher proficiency produced more iconics and beats but fewer deictics than less proficient L2 speakers (Lin, 2019).

Beats have been found to be associated positively with some aspects of speech performance. The use of beats was correlated positively with the length of utterances, the number of word types, and the number of scenes for bilingual kids (Nicoladis et al., 1999; Laurent et al., 2015). Furthermore, when explainers explained foreign words to learners, both explainers' and learners' beat gesture use predicted the number of word tokens they produced (Morett, 2014). Beat gesturing has been observed to co-occur with the use of connectives in the discourse (McNeill, 1992; Levy and McNeill, 2013), and as speakers' ability to establish discourse cohesion increased, their use of beats also

went up (Alamillo et al., 2013; Colletta et al., 2015). It can be seen that beats are associated positively with the length of speech units, speech length, and lexical richness, but the quantitative association between beat gesturing and discourse-associated speech measures requires further investigation.

Previous research has provided valuable findings in the associations between co-speech gesturing and speech performance, but there are still important aspects to be examined. First, a large percentage of relevant studies focused on representational gestures, leaving the relationship between individual types of co-speech gestures and speech performance insufficiently studied. Second, most studies on gesture and speech did not situate themselves in the framework of L2 acquisition, so they failed to fully consider how L2 speech performance is measured in the area of applied linguistics and thus could not give a full picture of the associations between gesturing and L2 speech performance. Third, the literature that has explored the associations between gesture use and speech performance largely focused on native speakers and highly proficient speakers, whose gesture use was probably associated more with personal speaking styles than with the speech production process, and thus such studies have yielded mixed results. Less attention has been paid to L2 learners whose gesture use is more likely to facilitate speech, and research on this topic also has pedagogical implications.

In view of the above research gaps, this study aims to explore the associations between different types of co-speech gestures and L2 speech performance. All of the four types of co-speech gestures were considered, and the meaning, form, and discourse aspects of L2 speech performance were measured (see Section "Speech transcription and coding" for details). The specific research question is: To what extent is each type of co-speech gestures associated with the meaning, form, and discourse aspects of L2 speech performance?

Methodology

Participants

Participants were 61 lower-intermediate level English learners in a Chinese university. They were recruited from a spoken English course for first year non-English major undergraduates taught by the second author. The speech task that provided speech and gesture data was one of the class activities. Students who agreed to have their data studied were offered a coupon for cake. All students agreed to participate.

Our participants included 24 male and 37 female students aged from 17 to 19 ($M = 17.98$; $SD = 0.46$). They were all Han Chinese with Chinese as their mother tongue, English the second language and no third language. They started to learn English from primary school and had been enrolled in the undergraduate program for about 8 weeks when the study took place. At the time of enrollment, they were categorized as intermediate-level English learners based on an English proficiency test provided by the i-Test system designed by Foreign Language Teaching and Research Press. The test, with a written component and a spoken component, examines learners' comprehensive English abilities. Intermediate English learners in this university were required to study English for two semesters before taking the nationwide College English Test, band 4 (CET-4). CET-4 matches the fifth level of China's Standards of English Language Ability scale (Wang, 2018), which corresponds to IELTS 5.5 and the lower B2 level of CEFR. Therefore, the English proficiency level of our participants was regarded as lower-intermediate.

Instrument

A cartoon-retelling task was used to elicit speech and gesture data. The task used a 30-s cartoon clip of *Tom and Jerry*, in which Tom played with Jerry by running after him, catching him, and letting him go time after time, and Jerry was trying to escape from Tom with all his might. The cartoon is rich in movements, actions, and trajectories, which is conducive to gesturing. The instructions to the participants were as follows: *You are going to see a clip of a cartoon. After watching it, please describe what happened in the clip.*

Procedures

Participants were told that this classroom activity was also intended as the data source of a study on how to improve L2 learners' spoken English. It was made clear that students who agreed to have their data used would be rewarded with a coupon for cake, and students who chose not to participate would not be affected in any way. After filling the general information sheet and consent form, students were led to a nearby classroom one by one, where the experimenter stated the task requirements and showed each of them the video clip. Participants were asked to do the retelling immediately after watching the clip, and there was no time limit for the retelling. During this process, participants stood opposite the experimenter so that their gestures could be captured clearly, and they were required to keep their hands empty. They were allowed to ask the experimenter to clarify the task requirement. A video camera was used to record participants' performance. A quick post-experiment survey showed that no one had been aware that the study intended to elicit gesture data. Afterwards, the real purpose of the study was revealed to the participants, and they could choose to withdraw from the study within 3 months from the experiment date, and none did so.

Speech transcription and coding

Students' speeches were transcribed verbatim and analyzed. In choosing speech performance measures for this study, we referred to the widely accepted complexity, accuracy, and fluency dimensions in evaluating L2 speech performance and adopted the relevant notion that L2 speech performance includes meaning and form associated aspects. In addition, we also considered the aspects of speech performance that have been reported to be empirically or theoretically associated with gesturing. Based on relevant literature, co-speech gesturing is related to speech length, lexical richness, speaking fluency, the length of certain speech units, and discourse cohesion. Speech length pertains to the productivity in communicating messages, lexical richness reflects the diversity of words used to express thoughts, and speaking fluency represents how quickly the speaker intends to covey ideas; thus, they were all treated as meaning-associated speech measures in this study. The length of certain speech units reflects how complex a clause or utterance is, and thus was taken as a measure related to form. Discourse cohesion reflects speakers' ability to produce text within which there are associated meaning units and establish the underlying meaning structure on the discourse level (Halliday and Hasan, 1976), and it was considered a separate dimension of L2 speech performance, namely, the discourse dimension.

This study employed three meaning-associated speech measures, including the number of word tokens to measure speech length, the root type/token ratio (RTTR) to measure lexical richness, and the speech rate to measure fluency. Although little research has shown significant association between form-associated speech measures and gesturing, we still checked this possibility since gesturing is believed to have the function of lowering working memory load in general. Three widely used form-associated measures were chosen, including the percentage of subordination to measure syntactic complexity (a complexity measure), the mean length of clauses to measure clause complexity (a complexity measure), and the percentage of error-free clauses to measure speech correctness (an accuracy measure). We employed the number of connectives per clause to measure discourse cohesion. In sum, seven speech measures were employed to represent the meaning, form, and discourse dimensions of L2 speech performance.

Table 1 shows how meaning, form, and discourse-associated speech measures were calculated. In calculating form-associated speech measures, the unit adopted is the Analysis of Speech Unit (Foster et al., 2000, p. 365–366), which is defined as "an independent clause or an independent sub-clausal unit together with any associated subordinate clause(s)." Clauses consisted of the Analysis of Speech Units and subordinate clauses.

To calculate the above meaning, form, and discourse-associated speech measures, we first removed repair disfluencies from the verbatim transcription, such as repetitions, self-corrections, hesitations, and speech irrelevant to the cartoon-retelling; we then counted speech data in terms of the number of word tokens, word types, syllables, connectives, the Analysis of Speech Units, subordinate clauses, clauses, and error-free clauses, and we also annotated sample duration. For annotations that could not be generated automatically, i.e., connectives, the Analysis of Speech Units, subordinate clauses, and error-free clauses, the first author did the coding first and a research assistant coded 20% (13 participants) of the data. The agreement was 100% ($N = 117$) for connectives, 92.25% ($N = 142$) for the number of the Analysis of Speech Units, 97.56% ($N = 41$) for subordinate clauses, and 100% ($N = 44$) for error-free clauses.

TABLE 1 The calculations of speech measures.

Meaning-associated speech measures
Number of word tokens: The total number of word tokens
RTTR: The number of word types divided by the square root of the number of word tokens
Speech rate: The number of syllables divided by sample time (in min)
Form-associated speech measures
Percentage of subordination: The number of subordinate clauses divided by the number of clauses
Mean length of clauses: The number of word tokens divided by the number of clauses
Percentage of error-free clauses: The number of error-free clauses divided by the number of clauses
Discourse-associated speech measures
Number of connectives per clause: The total number of connectives including coordinating and subordinating conjunctions, connective adverbs, and other connective expressions, divided by the number of clauses

Gesture coding

We followed McNeill (1992) to categorize co-speech gestures into iconics, metaphorics, deictics, and beats. Gestures were coded in ELAN (Lausberg and Sloetjes, 2009). Gesture types were annotated based on both the gesture form and the accompanying speech content. Participants often used iconics to mimic the moving route and actions of Tom and Jerry; metaphorics were employed to accompany abstract speech content such as process, time, and emotion; deictics were adopted when participants referred to the protagonists of the cartoon, Tom and Jerry, as well as to their body parts associated with movements, such as Jerry's tail and Tom's mouth; beats were the most frequently used type of gestures for all participants.

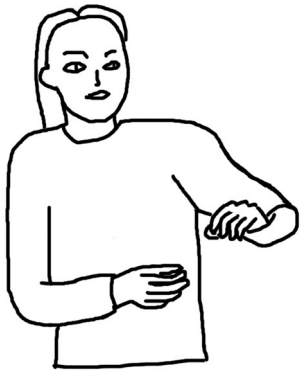


FIGURE 1
An example of an iconic gesture.

FIGURE 2
An example of a metaphoric gesture.

FIGURE 3
An example of a deictic gesture.

FIGURE 4
An example of a beat gesture.

Figures 1–4 illustrate examples of the iconic, metaphoric, deictic, and beat gestures respectively. The gesture in Figure 1 was made when the participant said *the cat use its fingers to kind of catch and hold the tail of the mouse*. Her right hand was kept still; meanwhile, her left elbow lifted, and her left thumb and forefinger reached out and touched each other as if she was pinching something. Obviously, she was mimicking the action of the cat, and thus this gesture was coded as an iconic gesture. The participant in Figure 2 said *and the cat is quite enjoying this process* when he made the gesture. His two hands were in front of him and near each other and he rotated his elbows so that his hands were circling like a wheel. This gesture was coded as a metaphoric one since it mimicked the abstract concept "process" with the circling of both hands. Figure 3 shows a deictic gesture when the participant said *and then he open his mouth and jerry run straight to his mouth*. Her left hand was in a relaxing position, and she pointed her right hand toward her mouth. Figure 4 illustrates the use of beats. The participant kept his left hand still and moved his right hand up and down repeatedly when he said *well several times Jerry went into the Tom's mouth*.

All gesture data were annotated by the first author, and then 20% of the data (data of 13 participants) were coded by a research assistant unaware of the research purpose. The agreement was 71.43% ($N = 42$) for iconics, 84.21% ($N = 5$) for metaphorics, 100% ($N = 18$) for deictics, and 88% ($N = 111$) for beats. We used the gesture rate as the gesture measure, which was calculated by dividing the number of each type of gestures by the number of word tokens and multiplying the result with 100.

Statistical analysis

We used both difference analysis and correlation analysis to show the relationship between co-speech gesturing and L2 speech performance. ANOVA was used to compare L2

TABLE 2 Descriptive statistics of data used in speech measure calculations and those of speech measures.

	Mean	*SD*	*Range*
Descriptive statistics of data used in speech measure calculations			
Number of syllables	73.69	48.11	20–241
Number of the analysis of speech units	8.67	5.57	2–28
Number of subordinations	2.64	2.04	0–9
Number of error-free clauses	3.77	3.18	0–14
Sample duration (in s)	62.46	27.21	17.6–125.74
Descriptive statistics of speech measures			
Meaning-associated speech measures			
Number of word tokens	61.20	38.91	19–180
RTTR	4.056	0.758	2.524–5.859
Speech rate	71.55	28.9	13.67–138.40
Form-associated speech measures			
Percentage of subordination	0.23	0.16	0–0.71
Mean length of clauses	5.43	1.00	3.14–8.25
Percentage of error-free clauses	0.33	0.23	0–1
Discourse-associated speech measures			
Number of connectives per clause	0.795	0.31	0.182–1.571

TABLE 3 Descriptive statistics of gesture tokens and frequency.

	Token			Frequency		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
Iconics	3.21	4.75	0–24	4.31	5.45	0–21.43
Metaphorics	0.56	1.42	0–8	0.58	1.36	0–5.56
Deictics	1.38	2.46	0–12	2.13	3.40	0–12.90
Beats	7.07	10.08	0–41	9.78	11.94	0–45.71

speech measures for gesturers and non-gesturers for each co-speech gesture type. Results of the Mann–Whitney *U*-test were reported for cases involving metaphorics as there was considerable difference in the number of gesturers and non-gesturers (see Table 4). Welch correction was adopted for cases where Levene's test reached significance. Since the gesture frequency data were not normally distributed, Spearman's rho was adopted to explore the correlational relationship between different types of co-speech gesture rate and speech measures.

Results

We first show descriptive statistics of the speech and gesture data, then display the results of difference analysis and correlation analysis, and finally answer the research question by summarizing the calculation results.

Descriptive statistics of speech and gesture data

Table 2 shows the means, standard deviations, and ranges of the data used in the calculations of speech measures and the descriptive statistics of the speech measures.

Table 3 illustrates the descriptive statistics of both the token and frequency of each type of gestures. It can be seen that the most frequently used gesture type was the beat, followed by the iconic, and then by the deictic. Metaphorics were used least, which was probably due to the largely concrete content involved in the cartoon clip. Table 4 shows the number of participants that used and did not use each type of gestures and the descriptive statistics of the gesture rate for gesturers. The descriptive statistics of speech data for gesturers and non-gesturers of each type of gesture are displayed in Tables 5–8. Please note that non-gesturers of a certain type of gesture may have used other types of

TABLE 4 Number of gesturers and non-gesturers of four types of gestures.

	Number of gesturers (*Mean*, *SD*, and *Range* of gesturers' gesture rate)	Number of non-gesturers
Iconics	33 ($M = 7.966$; $SD = 5.067$; *Range* = 0.654–21.429)	28
Metaphorics	13 ($M = 2.744$; $SD = 1.698$; *Range* = 0.654–5.556)	48
Deictics	27 ($M = 4.809$; $SD = 3.654$; *Range* = 1.053–12.903)	34
Beats	33 ($M = 18.081$; $SD = 10.599$; *Range* = 3.061–45.714)	28

TABLE 5 Descriptive statistics of speech measures for gesturers and non-gesturers of iconics.

	Gesturers (iconics)			Non-gesturers (iconics)		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
Meaning-associated speech measures						
Number of word tokens	78.424	40.894	19–180	40.893	24.33	19–103
RTTR	4.409	0.742	2.524–5.859	3.639	0.539	2.772–4.704
Speech rate	83.875	26.136	29.138–138.382	57.018	25.329	13.67–111.111
Form-associated speech measures						
Percentage of subordination	0.202	0.12	0–0.474	0.272	0.188	0–0.714
Mean length of clauses	5.488	0.809	3.923–7.176	5.352	1.196	3.143–8.25
Percentage of error-free clauses	0.312	0.191	0–0.714	0.359	0.27	0–1
Discourse-associated speech measures						
Number of connectives/clause	0.913	0.267	0.421–1.571	0.655	0.302	0.182–1.389

gestures (e.g., a non-gesturer of iconics may be a gesturer of beats).

Results of difference analysis: The presence and absence of co-speech gestures and L2 speech performance

Table 9 shows whether speech measures were different due to the presence and absence of individual types of co-speech gestures. Specifically, participants who used iconics produced speeches that were better in the number of word tokens ($F = 19.614$, $p < 0.001$), RTTR ($F = 20.826$, $p < 0.001$), the speech rate ($F = 16.453$, $p < 0.001$), and the number of connectives per clause ($F = 12.57$, $p < 0.001$), whereas using iconics did not make much difference in the percentage of subordination, the mean length of clauses, and the percentage of error-free clauses.

Similar to the case of iconics, participants who used metaphorics produced speech with more word tokens ($U = 530$, $p < 0.001$), higher RTTR ($U = 517$, $p < 0.001$), and faster speed ($U = 527$, $p < 0.001$) than participants who did not. The number of connectives per clause also trended toward significance ($U = 412$, $p = 0.077$), and the percentage of subordination and the length of clauses did not reach significance. Compared with participants who did not use metaphorics, participants who used them produced a higher percentage of error-free clauses ($U = 448.5$, $p = 0.016$), which is a form-associated measure.

The cases for deictic and beat gesturing resembled those of iconic gesturing, only with weaker effects. Deictic gesturers produced a significantly larger number of word tokens ($F = 5.724$, $p = 0.02$), higher RTTR ($F = 7.517$, $p = 0.008$), higher speech rate ($F = 7.742$, $p = 0.007$), and more connectives per clause ($F = 9.314$, $p = 0.003$) than non-gesturers of deictics, but none of the three form-associated measures showed any significant difference. Likewise, beat gesturers performed better than non-gesturers of beats in the number of word tokens ($F = 7.611$, $p = 0.008$), RTTR ($F = 6.354$, $p = 0.014$), the speech rate ($F = 9.268$, $p = 0.003$), and the number of connectives per clause ($F = 7.345$, $p = 0.009$), but speeches produced by beat gesturers and non-gesturers did not differ in any of the three form-associated speech measures.

Results of correlation analysis: The correlations between co-speech gestures and L2 speech performance

Table 10 shows the correlational relationships between different types of co-speech gesture rate and the speech measures associated with meaning, form, and discourse. The iconic gesture rate was positively correlated with both the three

TABLE 6 Descriptive statistics of speech measures for gesturers and non-gesturers of metaphorics.

	Gesturers (metaphorics)			Non-gesturers (metaphorics)		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
Meaning-associated speech measures						
Number of word tokens	99.615	42.9	45–180	50.792	30.723	19–147
RTTR	4.767	0.653	3.709–5.859	3.863	0.668	2.524–5.379
Speech rate	100.408	25.63	56.621–138.382	63.731	24.609	13.67–111.111
Form-associated speech measures						
Percentage of subordination	0.201	0.107	0–0.4	0.243	0.168	0–0.714
Mean length of clauses	5.645	0.705	4.357–7.176	5.366	1.063	3.143–8.25
Percentage of error-free clauses	0.429	0.109	0.235–0.625	0.308	0.247	0–1
Discourse-associated speech measures						
Number of connectives/clause	0.923	0.243	0.571–1.29	0.76	0.319	0.182–1.571

TABLE 7 Descriptive statistics of speech measures for gesturers and non-gesturers of deictics.

	Gesturers (deictics)			Non-gesturers (deictics)		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
Meaning-associated speech measures						
Number of word tokens	74.074	40.539	19–180	50.971	34.847	19–153
RTTR	4.339	0.841	2.524–5.859	3.831	0.607	2.772–5.114
Speech rate	82.501	27.358	38.182–138.382	62.849	27.433	13.67–116.505
Form-associated speech measures						
Percentage of subordination	0.202	0.136	0–0.5	0.26	0.17	0–0.714
Mean length of clauses	5.545	0.717	3.923–7.176	5.331	1.178	3.143–8.25
Percentage of error-free clauses	0.316	0.202	0–0.714	0.348	0.252	0–1
Discourse-associated speech measures						
Number of connectives/clause	0.922	0.3	0.421–1.571	0.694	0.282	0.182–1.308

meaning-associated speech measures, including the number of word tokens ($r = 0.452$, $p < 0.001$), RTTR ($r = 0.423$, $p < 0.001$), and the speech rate ($r = 0.407$, $p = 0.001$), and the discourse-associated speech measure, i.e., the number of connectives per clause ($r = 0.538$, $p < 0.001$). However, none of the form-associated speech measures were correlated significantly with the iconic gesture rate.

Similar to the iconic gesture rate, the metaphoric gesture rate was correlated positively and significantly with all the meaning-associated speech measures. The Spearman's rho was 0.485 for the number of word tokens ($p < 0.001$), 0.479 for RTTR ($p < 0.001$), and 0.483 for the speech rate ($p < 0.001$). In addition, the metaphoric gesture rate was not significantly correlated with the percentage of subordination and the mean length of clauses, both form-associated measures. Unlike the iconic gesture rate, the metaphoric gesture rate was correlated positively with the percentage of error-free clauses ($r = 0.312$, $p = 0.014$), which is also a speech measure associated with language form, but had no significant correlation with the number of connectives per clause.

Significant correlations were fewer for the deictic gesture rate. It was correlated positively with the speech rate ($r = 0.265$, $p = 0.039$) and the number of connectives per clause ($r = 0.354$, $p = 0.005$); its correlation with the number of word tokens ($r = 0.225$, $p = 0.082$) was close to being significant and with RTTR nonsignificant. The correlations between the deictic gesture rate and form-associated speech measures were not significant.

The correlational relationships for beats were similar to, though weaker than, those involving iconics. The beat gesture rate was correlated positively with two of the meaning-associated speech measures, including the number of word tokens ($r = 0.333$, $p = 0.009$) and the speech rate ($r = 0.303$, $p = 0.018$), and the discourse-associated measure, i.e., the number of connectives per clause ($r = 0.272$, $p = 0.034$). Its correlation with RTTR trended toward significance ($r = 0.231$, $p = 0.073$). Again, beat gesturing had no significant correlation with any of the three form-associated speech measures.

TABLE 8 Descriptive statistics of speech measures for gesturers and non-gesturers of beats.

	Gesturers (beats)			Non-gesturers (beats)		
	Mean	*SD*	*Range*	*Mean*	*SD*	*Range*
Meaning-associated speech measures						
Number of word tokens	73.212	40.44	19–180	47.036	32.274	19–147
RTTR	4.272	0.802	2.524–5.859	3.801	0.625	2.772–5.114
Speech rate	81.275	28.124	13.67–138.382	60.083	25.816	17.817–111.111
Form-associated speech measures						
Percentage of subordination	0.209	0.121	0–0.474	0.264	0.19	0–0.714
Mean length of clauses	5.355	0.877	3.143–7.176	5.508	1.137	4.182–8.25
Percentage of error-free clauses	0.306	0.181	0–0.625	0.367	0.276	0–1
Discourse-associated speech measures						
Number of connectives/clause	0.889	0.287	0.286–1.571	0.684	0.303	0.182–1.389

TABLE 9 The contrast of speech measures produced by gesturers and non-gesturers of four types of co-speech gestures.

	Iconics	Metaphorics	Deictics	Beats
	***F* (*ω*) for iconics, deictics, beats; *U* (*r*) for metaphorics**			
Meaning-associated speech measures				
Number of word tokens	19.614*** (0.468)	530*** (0.492)	5.724* (0.268)	7.611** (0.313)
RTTR	20.826*** (0.495)	517*** (0.462)	7.517** (0.311)	6.354* (0.285)
Speech rate	16.453*** (0.449)	527*** (0.485)	7.742** (0.316)	9.268** (0.345)
Form-associated speech measures				
Percentage of subordination	3.087 (0.182)	256 (−0.127)	2.061 (0.13)	1.703 (0.114)
Mean length of clauses	0.275 (0)	387 (0.169)	0.765 (0)	0.351 (0)
Percentage of error-free clauses	0.641 (0)	448.5* (0.308)	0.282 (0)	1.056 (0.032)
Discourse-associated speech measures				
Number of connectives per clause	12.57*** (0.399)	412 (0.226)	9.314** (0.346)	7.345** (0.307)

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

A summary: The associations between co-speech gestures and L2 speech performance

Positive associations with meaning-related L2 speech measures were observed for all the four types of co-speech gestures, and associations involving iconic gesturing and metaphoric gesturing were tighter. As shown in Table 11, gesturers of iconics performed better than non-gesturers of iconics in the three meaning-related measures, and this was also the case for metaphorics; in addition, both iconic and metaphoric rates were correlated positively with these speech measures. Deictic and beat gesturing had weaker associations with meaning-related speech measures. Though gesturers of deictics and gesturers of beats performed better in terms of the three meaning measures, the effect sizes were smaller compared with gesturers of iconics and metaphorics. Also, only one meaning-related speech measure (i.e., the speech rate) was correlated significantly with the deictic rate and two (i.e., the number of word tokens and the speech rate) with the beat rate in a positive way, and their effect sizes were again smaller compared with the iconic and metaphoric rates.

Form-associated L2 speech measures were not significantly associated with any co-speech gestures except metaphorics. Compared with non-gesturers of metaphorics, gesturers of metaphorics produced a larger percentage of error-free clauses, and the metaphoric rate was significantly correlated with this measure of speech form.

For the discourse-associated L2 speech measure (i.e., the number of connectives per clause), the association with co-speech gestures was strongest for iconics, weaker for deictics and beats, and weakest for metaphorics, as shown in Table 11. Participants who used iconics employed more connectives per clause than participants who did not use iconics, and similar differences also existed for gesturers and non-gesturers of deictics and beats; furthermore, the iconic rate, deictic rate, and beat rate were significantly correlated with the

TABLE 10 Correlations (Spearman's rho) between four types of co-speech gesture rate and speech measures.

	Iconic rate	Metaphoric rate	Deictic rate	Beat rate
Number of word tokens	0.452***	0.485***	0.225	0.333**
RTTR	0.423***	0.479***	0.207	0.231
Speech rate	0.407**	0.483***	0.265*	0.303*
Percentage of subordination	−0.236	−0.109	−0.222	−0.132
Mean length of clauses	0.097	0.154	0.230	−0.002
Percentage of error-free clauses	−0.138	0.312*	−0.062	−0.092
Number of connectives per clause	0.465***	0.212	0.354**	0.272*

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

TABLE 11 Effect sizes and *p* values for associations between four types of co-speech gestures and meaning and discourse-related speech measures.

	Iconics	Metaphorics	Deictics	Beats
Number of word tokens				
Difference analysis	$\omega = 0.468, p < 0.001$	$r = 0.492, p < 0.001$	$\omega = 0.268, p = 0.02$	$\omega = 0.313, p = 0.008$
Correlation analysis	$r = 0.452, p < 0.001$	$r = 0.485, p < 0.001$	$r = 0.225, p = 0.082$	$r = 0.333, p = 0.009$
RTTR				
Difference analysis	$\omega = 0.495, p < 0.001$	$r = 0.462, p < 0.001$	$\omega = 0.311, p = 0.008$	$\omega = 0.285, p = 0.014$
Correlation analysis	$r = 0.423, p < 0.001$	$r = 0.479, p < 0.001$	$r = 0.207, p = 0.11$	$r = 0.231, p = 0.073$
Speech rate				
Difference analysis	$\omega = 0.449, p < 0.001$	$r = 0.485, p < 0.001$	$\omega = 0.316, p = 0.007$	$\omega = 0.345, p = 0.003$
Correlation analysis	$r = 0.407, p = 0.001$	$r = 0.483, p < 0.001$	$r = 0.265, p = 0.039$	$r = 0.303, p = 0.018$
Number of connectives per clause				
Difference analysis	$\omega = 0.399, p < 0.001$	$r = 0.226, p = 0.077$	$\omega = 0.346, p = 0.003$	$\omega = 0.307, p = 0.009$
Correlation analysis	$r = 0.465, p < 0.001$	$r = 0.212, p = 0.101$	$r = 0.354, p = 0.005$	$r = 0.272, p = 0.034$

number of connectives per clause. The use of iconics had the largest effect sizes followed by deictics and beats for both the difference analysis and the correlation analysis. Metaphoric gesture use had no significant association with the discourse measure.

Discussion

This study explored the associations between co-speech gestures and speech performance for lower-intermediate L2 English language learners. We found that all the four types of co-speech gestures were more closely connected with meaning and discourse-associated L2 speech measures than with form-associated measures.

Inter-gesture differences

One inter-gesture difference is that the associations between individual types of co-speech gestures and meaning-associated L2 speech measures were stronger for iconics and metaphorics, weaker for beats, and weakest for deictics. This was shown by the differences in effect sizes (see Table 11). For meaning-associated L2 speech measures, the effect sizes of the difference analysis for gesturers and non-gesturers of iconics and metaphorics were comparable, but the effect sizes were smaller for gesturers and non-gesturers of beats and deictics. In addition, in the correlational analysis, iconic and metaphoric rates had larger effect sizes, the beat rate had smaller effect sizes, and the deictic rate had the smallest effect sizes. Also, all three meaning-associated L2 speech measures were significantly correlated with iconic and metaphoric rates, two of them were significantly correlated with the beat rate, and only one with the deictic rate. This inter-gesture difference can be explained by findings from previous studies. First, as representational gestures, iconics and metaphorics have similar functions (Kita et al., 2017), and they can help gesturers conceptualize speech content (Alibali et al., 2000; Kita and Davies, 2009; Chu et al., 2014) and access words (Rauscher et al., 1996; Krauss et al., 2000; Frick-Horbury, 2002), which explains why they were conducive to meaning expression. Second, beats are also helpful in accessing words

(Ravizza, 2003) and improving fluency (Vilà-Giménez and Prieto, 2020), but beats have a weaker association with speech content than representational gestures (McNeill, 1992). Third, deictics have the weakest association with speech performance measures and they tend to be used to compensate for weak speech proficiency (Nicoladis et al., 1999; Mayberry and Nicoladis, 2000; Nicoladis, 2002; Gan and Davison, 2011).

Another inter-gesture difference is that the association of the discourse-associated L2 speech measure, i.e., the number of connectives per clause, was the strongest with iconics, weaker with deictics and beats, and not significant with metaphorics (see Table 11). Whether or not L2 learners used iconics, deictics, and beats significantly influenced the use of connectives, and the effect sizes decreased from iconics, to deictics, and to beats. Correlation analysis displayed the same trend. The strongest association between iconics and connectives might be due to the function of iconics in conceptualizing speech content (Alibali et al., 2000; Kita and Davies, 2009; Chu et al., 2014), and the use of connectives is an important indicator of the quality of speech content on the discourse level. The reason why metaphorics, which are also representational in nature, bore little association with the use of connectives is unclear. It is possible that metaphorics were used so infrequently that no obvious association with connectives could be observed; it is also possible that by nature metaphorics are less likely to be used with connectives than with other speech measures. Both deictics and beats link relatively weakly to speech content (McNeill, 1992; Nicoladis et al., 1999; Mayberry and Nicoladis, 2000). Although previous studies have observed the co-occurrence of beats and connectives (McNeill, 1992; Dimitrova et al., 2016; Shattuck-Hufnagel and Ren, 2018), the association of beats with connective use was weaker than the association between iconics and connectives in this study, indicating that discourse cohesion is related more to the quality of speech content in general.

The third inter-gesture difference is the contrastive associations of individual types of co-speech gestures with the percentage of error-free clauses. The use of metaphoric gestures had a positive association with the percentage of error-free clauses, whereas the other three types of gestures bore negative, though insignificant, relations with it. This higher degree of accuracy for metaphorics was not due to shorter speech production and fewer opportunities to make language mistakes, since gesturers of metaphorics produced more word tokens than non-gesturers and the metaphoric gesture rate was correlated positively with the number of word tokens. Our results indicate that lower-intermediate L2 learners who use more metaphoric gestures also have a higher ability to monitor language mistakes. To our knowledge, no study has mentioned the function of metaphoric gestures in improving speech accuracy. In learning grammar, metaphoric gesturing is an important indicator to show L2 learners' learning process (Kimura and Kazik, 2017) and an effective interactional strategy to communicate with the lecturer about grammar learning (Matsumoto and Dobs, 2017). It is possible that our participants who benefited from gesturing in learning grammar also inherited such gesture use in monitoring language accuracy during speech production. Another possibility is that producing metaphoric gestures by schematizing abstract concepts saved the type of working memory resources that could be used for other tasks related to metaphoric thinking like monitoring grammar. Further studies are needed to investigate the mechanisms of this relationship.

L2 speech dimensions associated with co-speech gesturing

Our results indicate that meaning and discourse-associated aspects of L2 speech measures are associated with co-speech gestures, whereas form-associated ones have weak associations with gestures. This is generally consistent with previous findings with regard to the functions of co-speech gestures, i.e., conceptualizing information and retrieving words (Rauscher et al., 1996; Alibali et al., 2000). Such functions can explain the associations between co-speech gesturing and speech length, lexical richness, speaking fluency, and discourse cohesion of L2 speech. Co-speech gestures' function of lowering cognitive load (Goldin-Meadow et al., 2001; Wagner et al., 2004; Ping and Goldin-Meadow, 2010; Cook et al., 2012; Chu et al., 2014) is also supported by our findings, but it seems that the working memory benefits brought by gesturing only contributed to improving the meaning expression for L2 learners, but not monitoring language form. Our participants were lower-intermediate level L2 learners for whom the cartoon-retelling task was challenging, and when they undertook the speech task, they were struggling to finish the task. In such cases, the cognitive resources created by using co-speech gestures might have been allocated primarily to conveying meaning. It is still unclear whether the particular measures of L2 speech associated with co-speech gesturing were influenced by language proficiency. It is possible that for more proficient language learners who find it unchallenging to retrieve words and express ideas, using co-speech gestures will benefit the speech performance measures related to form. Another possible explanation is that speakers who chose to gesture might have a greater desire to communicate ideas and thus spent more cognitive resources on meaning expression. More studies on the associations between gesturing, willingness to communicate, and speech performance are needed.

Our study demonstrates positive associations between connectives and gestures in a quantitative way. Using connectives is an important way to show semantic

relations between textual constitutes (Halliday and Hasan, 1976). Their positive associations with co-speech gestures, namely iconics, beats, and deictics, show that co-speech gestures are helpful in not only meaning expression within clauses, but also meaning construction in the discourse.

The function of co-speech gesturing seems to resemble that of strategic pre-task planning in L2 speech production. Strategic planning makes learners prioritize meaning over form, such as producing speech with higher fluency and more diversified words (Yuan and Ellis, 2003; Sangarun, 2005; Li and Fu, 2018). Our participants were not allowed to prepare before the task, and co-speech gesturing was generally associated with measures related to speech meaning, such as the number of word tokens, RTTR that represents lexical richness, and the speech rate. This indicates that the cognitive resources released by using co-speech gestures might be similar to those provided by strategic planning. This possibility makes co-speech gestures a promising strategy when L2 learners face a challenging speech task with no preparation time. However, it is also found that, for advanced L2 learners, more representational and deictic gestures were produced when they were not allowed to do pre-task planning but more iconic ones were generated when they were allowed to do so (Lin, 2020), suggesting a more complex interaction between gesture, task planning, and language proficiency. To have more solid conclusions, future studies need to adopt experimental designs conductive to explore causal relationships between gesture and L2 speech performance, and take language proficiency and other task related factors into consideration.

The associations between gesture and speech for L2 learners

Our study is a necessary complement to the current literature in that most relevant research focused on native and highly proficient bilingual speakers (Nicoladis et al., 2009; Smithson et al., 2011; Laurent and Nicoladis, 2015; Laurent et al., 2015). The associations between gesture use and speech performance reported in such studies were not consistent, which may have been caused by the ceiling effects resulted from participants' high language proficiency (Nicoladis et al., 2007). Our study explored the associations between co-speech gesturing and speech performance for lower-intermediate L2 learners, and we found some positive associations. With a lower language proficiency level, our participants were very likely to have faced a shortage of working memory resources and used gestures to facilitate speech production. It has been shown that the cognitive benefits of co-speech gestures are only obvious when participants face a high cognitive load (Marstaller and Burianová, 2013; Chu et al., 2014; Lewis et al., 2015). Co-speech gesturing produced by participants with lower language proficiency level in our study was not likely to be a reflection of speaking style (Nagpal et al., 2011), but more possibly a strategy to cope with the shortage of cognitive resources required by the speech task. Thus, the associations found in this study support the widely recognized close connection between gesture and speech. Our study can serve as a stepping stone to further research on the causal effect of co-speech gesturing on L2 speech performance. Such explorations have both pedagogical significance and practical value for L2 teachers and speakers.

Conclusion

Based on the speech and gesture data elicited from a cartoon-retelling task completed by lower-intermediate L2 learners, this study found that all the four types of co-speech gestures were positively associated with meaning-related L2 speech measures, with the associations involving iconics and metaphorics stronger and that involving deictics and beats weaker; iconics, deictics and beats were also associated with L2 discourse cohesion in a positive way; and co-speech gestures had little association with form-associated L2 speech measures, except that metaphoric gestures were positively associated with the percentage of error-free clauses. The results show that all the four types of co-speech gestures tend to have a positive association with L2 speech meaning construction both within the clause and on the discourse level.

This study has several limitations. First, the interactions between speech, gestures, and other individual factors, such as personality, willingness to communicate, and working memory, were not taken into consideration. Second, participants were not required to gesture or not to; rather, they used or did not use gestures spontaneously. While this design eliminated possible influencing factors, it also made us lose the opportunity to explore the causal relationship between gesturing and speech. Third, since only L2 learners of one proficiency level were recruited as participants, whether the findings hold for speakers of other proficiency levels is unclear. Fourth, whether the findings apply to L2 learners with different language backgrounds requires further investigation. Lastly, our speech data were monologs instead of dialogues, which made it hard to explore interactional features of gesture and speech, such as pragmatic aspects of gestures (see Kendon, 2017 for more information). Future studies are needed to further explore the relationship between gesture and speech.

Data availability statement

Researchers can contact the corresponding authors for access to de-identified data.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

SM wrote the first draft of the manuscript and GJ revised the draft. Both authors contributed to the conception and design of the study and data collection and analysis.

Funding

This research was supported by the Fundamental Research Funds for the Central Universities in UIBE (Grant No. CXTD11-02). The fund was used to cover part of the salaries paid to research assistants and part of the processing fee.

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

References

Agostinho, S., Tindall-Ford, S., Ginns, P., Howard, S. J., Leahy, W., and Paas, F. (2015). Giving learning a helping hand: finger tracing of temperature graphs on an iPad. *Educ. Psychol. Rev.* 27, 427–443. doi: 10.1007/s10648-015-9315-5

Alamillo, A. R., Colletta, J. M., and Guidetti, M. (2013). Gesture and language in narratives and explanations: the effects of age and communicative activity on late multimodal discourse development. *J. Child Lang.* 40, 511–538. doi: 10.1017/S0305000912000062

Alibali, M. W., Kita, S., and Young, A. J. (2000). Gesture and the process of speech production: we think, therefore we gesture. *Lang. Cogn. Process.* 15, 593–613. doi: 10.1080/016909600750040571

Aziz, J., and Nicoladis, E. (2019). "My french is rusty": proficiency and bilingual gesture use in a majority English community. *Biling.-Lang. Cogn.* 22, 826–835. doi: 10.1017/S1366728918000639

Burns, P., McCormack, T., Jaroslawska, A. J., O'Connor, P. A., and Caruso, E. M. (2019). Time points: a gestural study of the development of space–time mappings. *Cogn. Sci.* 43. doi: 10.1111/cogs.12801 (accessed April 01, 2022).

Cartmill, E. A., and Goldin-Meadow, S. (2016). "Gesture," in *APA Handbook of Nonverbal Communication*, eds D. Matsumoto, H. C. Hwang, and M. G. Frank (Washington, DC: American Psychological Association), 307–333. doi: 10.1037/14669-012

Chu, M., Meyer, A., Foulkes, L., and Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *J. Exp. Psychol. Gen.* 143, 694–709. doi: 10.1037/a0033861

Chui, K. (2011). Conceptual metaphors in gesture. *Cogn. Linguist.* 22, 437–458. doi: 10.1515/cogl.2011.017

Church, R. B., Kelly, S., and Holcombe, D. (2014). Temporal synchrony between speech, action and gesture during language production. *Lang. Cogn. Neurosci.* 29, 345–354. doi: 10.1080/01690965.2013.857783

Cienki, A., and Müller, C., eds (2008). *Metaphor and Gesture*. Amsterdam: John Benjamins. doi: 10.1075/gs.3

Colletta, J. M., Guidetti, M., Capirci, O., Cristilli, C., Demir, O. E., Kunene-Nicolas, R. N., et al. (2015). Effects of age and language on co-speech gesture production: an investigation of french, american, and italian children's narratives. *J. Child Lang.* 42, 122–145. doi: 10.1017/S0305000913000585

Cook, S. W., Yip, T. K., and Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Lang. Cogn. Process.* 27, 594–610. doi: 10.1080/01690965.2011.567074

Delgado, B., Gómez, J. C., and Sarriá, E. (2009). "Private pointing and private speech: developing parallelisms," in *Private Speech, Executive Functioning, and the Development of Verbal Self-Regulation*, eds A. Winsler, C. Fernyhough, and I. Montero (New York, NY: Cambridge University Press), 153–162. doi: 10.1017/CBO9780511581533.013

Dimitrova, D., Chu, M., Wang, L., Özyürek, A., and Hagoort, P. (2016). Beat that word: how listeners integrate beat gesture and focus in multimodal speech discourse. *J. Cogn. Neurosci.* 28, 1255–1269. doi: 10.1162/jocn_a_00963

Ellis, R., and Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: Oxford University Press.

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Appl. Linguist.* 21, 354–375. doi: 10.1093/applin/21.3.354

Frick-Horbury, D. (2002). The use of hand gestures as self-generated cues for recall of verbally associated targets. *Am. J. Psychol.* 115, 1–20. doi: 10.2307/1423671

Gan, Z., and Davison, C. (2011). Gestural behavior in group oral assessment: a case study of higher- and lower-scoring students. *Int. J. Appl. Linguist.* 21, 94–120. doi: 10.1111/j.1473-4192.2010.00264.x

Gillespie, M., James, A. N., Federmeier, K. D., and Watson, D. G. (2014). Verbal working memory predicts co-speech gesture: evidence from individual differences. *Cognition* 132, 174–180. doi: 10.1016/j.cognition.2014.03.012

Ginns, P., Hu, F. T., Byrne, E., and Bobis, J. (2016). Learning by tracing worked examples. *Appl. Cogn. Psychol.* 30, 160–169. doi: 10.1002/acp.3171

Ginns, P., and King, V. (2021). Pointing and tracing enhance computer-based learning. *Educ. Technol. Res. Dev.* 69, 1387–1403. doi: 10.1007/s11423-021-09997-0

Goldin-Meadow, S. (2003). *Hearing Gesture: How our Hands Help us Think*. Cambridge, MA: Harvard University Press. doi: 10.1037/e413812005-377

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., and Wagner, S. (2001). Explaining math: gesturing lightens the load. *Psychol. Sci.* 12, 516–522. doi: 10.1111/1467-9280.00395

Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hu, F. T., Ginns, P., and Bobis, J. (2014). Does tracing worked examples enhance geometry learning? *Aust. J. Educ. Dev. Psychol.* 14, 45–49. Available at https://files.eric.ed.gov/fulltext/EJ1041668.pdf; (accessed April 01, 2022).

Hu, F. T., Ginns, P., and Bobis, J. (2015). Getting the point: tracing worked examples enhances learning. *Learn. Instr.* 35, 85–93. doi: 10.1016/j.learninstruc.2014.10.002

Iverson, J. M., and Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychol. Sci.* 16, 367–371. doi: 10.1111/j.0956-7976.2005.01542.x

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511807572

Kendon, A. (2017). Pragmatic functions of gestures: some observations on the history of their study and their nature. *Gesture* 16, 157–175. doi: 10.1075/gest.16.2.01ken

Kimura, D., and Kazik, N. (2017). Learning in-progress: on the role of gesture in microgenetic development of L2 grammar. *Gesture* 16, 126–150. doi: 10.1075/gest.16.1.05kim

Kita, S., Alibali, M. W., and Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychol. Rev.* 124, 245–266. doi: 10.1037/rev0000059

Kita, S., and Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Lang. Cogn. Process.* 24, 761–775. doi: 10.1080/01690960802327971

Kita, S., and Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal? Evidence for an interface representation of spatial thinking and speaking. *J. Mem. Lang.* 48, 16–32. doi: 10.1016/S0749-596X(02)00505-3

Korbach, A., Ginns, P., Brünken, R., and Park, B. (2020). Should learners use their hands for learning? Results from an eye-tracking study. *J. Comput. Assist. Learn.* 36, 102–113. doi: 10.1111/jcal.12396

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.

Krauss, R. M., Chen, Y., and Gottesman, R. (2000). "Lexical gestures and lexical access: a process model," in *Language and Gesture*, ed. D. McNeill (New York, NY: Cambridge University Press), 261–283. doi: 10.1017/CBO9780511620850.017

Lakoff, G., and Johnson, M. (1980). *Metaphors we Live by*. Chicago, IL: The University of Chicago Press.

Laurent, A., and Nicoladis, E. (2015). Gesture restriction affects french-english bilinguals' speech only in french. *Biling. Lang. Cogn.* 18, 340–349. doi: 10.1017/S1366728914000042

Laurent, A., Nicoladis, E., and Marentette, P. (2015). The development of storytelling in two languages with words and gestures. *Int. J. Biling.* 19, 56–74. doi: 10.1177/1367006913495618

Lausberg, H., and Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Res. Methods, Instruments, Comput.* 41, 841–849. doi: 10.3758/BRM.41.3.841

Levy, E. T., and McNeill, D. (2013). Narrative development as symbol formation: gestures, imagery and the emergence of cohesion. *Cult. Psychol.* 19, 548–569. doi: 10.1177/1354067X13500328

Lewis, C., Lovatt, P., and Kirk, E. (2015). Many hands make light work: the facilitative role of gesture in verbal improvisation. *Think. Ski. Creat.* 17, 149–157. doi: 10.1016/j.tsc.2015.06.001

Li, S., and Fu, M. (2018). Strategic and unpressured within-task planning and their associations with working memory. *Lang. Teach. Res.* 22, 230–253. doi: 10.1177/1362168816684367

Lin, Y. L. (2020). A helping hand for thinking and speaking: effects of gesturing and task planning on second language narrative discourse. *System* 91, 120–130. doi: 10.1016/j.system.2020.102243

Lin, Y. L. (2019). Speech-accompanying gestures in L1 and L2 conversational interaction by speakers of different proficiency levels. *IRAL*. 60, 123–142. doi: 10.1515/iral-2017-0043

Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology* 3, 71–89. doi: 10.1515/lp-2012-0006

Lucero, C., Zaharchuk, H., and Casasanto, D. (2014). "Beat gestures facilitate speech production," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36, (Quebec City, QC, Canada), 898–903. Available online at: https://escholarship.org/uc/item/8pw1x0sx (accessed April 01, 2022).

Ma, S., Jin, G., and Barlow, M. (2021). Representational gestures correlated with meaning associated aspects of L2 speech performance. *Gesture* 20. (in press) doi: 10.1075/gest.19017.ma

Marstaller, L., and Burianová, H. (2013). Individual differences in the gesture effect on working memory. *Psychon. Bull. Rev.* 20, 496–500. doi: 10.3758/s13423-012-0365-0

Matsumoto, Y., and Dobs, A. M. (2017). Pedagogical gestures as interactional resources for teaching and learning tense and aspect in the ESL grammar classroom. *Lang. Learn.* 67, 7–42. doi: 10.1111/lang.12181

Mayberry, R. I., and Nicoladis, E. (2000). Gesture reflects language development: evidence from bilingual children. *Curr. Dir. Psychol. Sci.* 9, 192–196. doi: 10.1111/1467-8721.00092

McNeill, D. (1992). *Hand and Mind: What the Hands Reveal About Thought*. Chicago, IL: University of Chicago Press.

Morett, L. M. (2014). When hands speak louder than words: the role of gesture in the communication, encoding, and recall of words in a novel second language. *Mod. Lang. J.* 98, 834–853. doi: 10.1111/modl.12125

Mota, M. B. (2005). Working memory capacity and fluency, accuracy, complexity, and lexical density in l2 speech production. *Fragmentos*, 69–104. Available online at: https://www.researchgate.net/profile/Mailce-Mota/publication/242149105_Working_memory_capacity_and_fluency_accuracy_complexity_and_lexical_density_in_L2_speech_production/links/55b77a9f08aec0e5f4381e98/Working-memory-capacity-and-fluency-accuracy-complexity-and-lexical-density-in-L2-speech-production.pdf (accessed April 01, 2022).

Nagpal, J., Nicoladis, E., and Marentette, P. (2011). Predicting individual differences in L2 speakers' gestures. *Int. J. Biling.* 15, 205–214. doi: 10.1177/1367006910381195

Nematizadeh, S., and Wood, D. (2019). Willingness to communicate and second language speech fluency: an investigation of affective and cognitive dynamics. *Can. Mod. Lang. Rev.* 75, 197–215. doi: 10.3138/cmlr.2017-0146

Nicoladis, E. (2002). Some gestures develop in conjunction with spoken language development and others don't: evidence from bilingual preschoolers. *J. Nonverbal Behav.* 26, 241–266. doi: 10.1023/A:1022112201348

Nicoladis, E., Marentette, P., and Navarro, S. (2016). Gesture frequency linked primarily to story length in 4–10-year old children's stories. *J. Psycholinguist. Res.* 45, 189–204. doi: 10.1007/s10936-014-9342-2

Nicoladis, E., Mayberry, R. I., and Genesee, F. (1999). Gesture and early bilingual development. *Dev. Psychol.* 35, 514–526. doi: 10.1037/0012-1649.35.2.514

Nicoladis, E., Pika, S., and Marentette, P. (2009). Do french-english bilingual children gesture more than monolingual children? *J. Psycholinguist. Res.* 38, 573–585. doi: 10.1007/s10936-009-9121-7

Nicoladis, E., Pika, S., Yin, H., and Marentette, P. (2007). Gesture use in story recall by chinese-english bilinguals. *Appl. Psycholinguist.* 28, 721–735. doi: 10.1017/S0142716407070385

Oya, T., Manalo, E., and Greenwood, J. (2004). The influence of personality and anxiety on the oral performance of japanese speakers of english. *Appl. Cogn. Psychol.* 18, 841–855. doi: 10.1002/acp.1063

Ping, R., and Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cogn. Sci.* 34, 602–619. doi: 10.1111/j.1551-6709.2010.01102.x

Rauscher, F. H., Krauss, R. M., and Chen, Y. (1996). Gesture, speech, and lexical access: the role of lexical movements in speech production. *Psychol. Sci.* 7, 226–231. doi: 10.1111/j.1467-9280.1996.tb00364.x

Ravizza, S. (2003). Movement and lexical access: do noniconic gestures aid in retrieval? *Psychon. Bull. Rev.* 10, 610–615. doi: 10.3758/BF03196522

Sangarun, J. (2005). "The effects of focusing on meaning and form in strategic planning," in *Planning and Task Performance in a Second Language*, ed. R. Ellis (Philadelphia, PA: John Benjamins), 111–141. doi: 10.1075/lllt.11.08san

Shattuck-Hufnagel, S., and Ren, A. (2018). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Front. Psychol.* 9:1514. doi: 10.3389/fpsyg.2018.01514

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press. doi: 10.1177/003368829802900209

Skehan, P. (2009). Modelling second language performance: integrating complexity, accuracy, fluency, and lexis. *Appl. Linguist.* 30, 510–532. doi: 10.1093/applin/amp047

Skehan, P., ed. (2014). *Processing Perspectives on Task Performance*. Amsterdam: John Benjamins. doi: 10.1075/tblt.5

Smithson, L., Nicoladis, E., and Marentette, P. (2011). Bilingual children's gesture use. *Gesture* 11, 330–347. doi: 10.1075/gest.11.3.04smi

Sun, P. P. (2022). The influence of cognitive, affective, and sociocultural individual differences on L2 Chinese speech performance: a multiple indicators multiple causes approach. *J. Psycholinguist. Res.* doi: 10.1007/s10936-022-09847-x (accessed June 15, 2022).

Vilà-Giménez, I., Dowling, N., Demir-Lira, E., Prieto, P., and Goldin-Meadow, S. (2021). The predictive value of non-referential beat gestures: early use in parent–child interactions predicts narrative abilities at 5 years of age. *Child Dev.* 92, 2335–2355. doi: 10.1111/cdev.13583

Vilà-Giménez, I., and Prieto, P. (2020). Encouraging kids to beat: children's beat gesture production boosts their narrative performance. *Dev. Sci.* 23, 1–14. doi: 10.1111/desc.12967

Wagner, S. M., Nusbaum, H., and Goldin-Meadow, S. (2004). Probing the mental representation of gesture: is handwaving spatial? *J. Mem. Lang.* 50, 395–407. doi: 10.1016/j.jml.2004.01.002

Wang, B., Ginns, P., and Mockler, N. (2022). Sequencing tracing with imagination. *Educ. Psychol. Rev.* 34, 421–449. doi: 10.1007/s10648-021-09625-6

Wang, S. (2018). The application of china's standards of English language ability to college English teaching. *Foreign Lang. Educ.* 39, 1–4. doi: 10.16362/j.cnki.cn61-1023/h.2018.04.001

Weissheimer, J., and Mota, M. B. (2009). Individual differences in working memory capacity and the development of L2 speech production. *Issues Appl. Linguist.* 17, 93–112. doi: 10.5070/L4172005115

Wesp, R., Hesse, J., Keutmann, D., and Wheaton, K. (2001). Gestures maintain spatial imagery. *Am. J. Psychol.* 114, 591–600. doi: 10.2307/1423612

Yuan, F., and Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Appl. Linguist.* 24, 1–27. doi: 10.1093/applin/24.1.1

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 30 August 2022
DOI 10.3389/fpsyg.2022.968946

OPEN ACCESS

EDITED BY
Peijian Paul Sun,
Zhejiang University, China

REVIEWED BY
Majid Elahi Shirvan,
University of Bojnord, Iran
Su-hie Ting,
Universiti Malaysia Sarawak, Malaysia

*CORRESPONDENCE
Scott Aubrey
scaubrey@cuhk.edu.hk

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

RECEIVED 14 June 2022
ACCEPTED 03 August 2022
PUBLISHED 30 August 2022

CITATION
Aubrey S (2022) The relationship between anxiety, enjoyment, and breakdown fluency during second language speaking tasks: An idiodynamic investigation. *Front. Psychol.* 13:968946. doi: 10.3389/fpsyg.2022.968946

The relationship between anxiety, enjoyment, and breakdown fluency during second language speaking tasks: An idiodynamic investigation

Scott Aubrey*

Department of Curriculum and Instruction, Faculty of Education, The Chinese University of Hong Kong, Hong Kong, Hong Kong SAR, China

Research has found that levels of enjoyment and anxiety fluctuate on a moment-to-moment timescale during second language (L2) spoken task performances as learners attempt to cope with various communication challenges. For L2 speakers, surges in these emotions can limit or expand cognitive resources, affecting speech processing capability. However, at an intra-individual level, there is very little empirical evidence on how emotions and fluency are related during L2 spoken task performances. The present study uses the idiodynamic approach to examine the relationship between enjoyment, anxiety, and breakdown fluency (i.e., average length of pauses) during monolog tasks performed by university students who use English as an L2. After watching a video recording of their task performances, participants rated their anxiety and emotion levels on a per-second basis. Immediately after, they were interviewed about their attributions for fluctuations in their ratings. After segmenting task performances into 26 7-s segments of speech, per-person correlations revealed that (1) the (negative) relationship between anxiety and enjoyment varied from strong to very weak, and (2) the (positive) relationship between anxiety and breakdown fluency was much stronger than the (negative) relationship between enjoyment and breakdown fluency. Triangulation of anxiety and enjoyment ratings, stimulated recall interviews, and performance data led to the identification of four categories of factors that influenced the emotion-fluency relationship: task design factors (task structure, task topic), task implementation factors (pre-task planning, task time limits), cognitive-linguistic factors (momentary breakdowns in conceptualization and formulation speech processes), and achievement outcome factors (self-evaluations of appropriate and inappropriate language used).

KEYWORDS

anxiety, enjoyment, fluency, task-based language teaching, idiodynamic method, second language learning, spoken task performances

Introduction

With the so-called "affective turn" in second language acquisition (SLA) (Pavlenko, 2013), there has been a growing understanding that emotions play an important role in second language (L2) learning (e.g., Swain, 2013; Prior, 2019; Lambert and Zhang, 2019; MacIntyre and Wang, 2021). Although the exact mechanism is unclear, scholars agree that emotions tend to motivate and influence cognition (Izard, 2009). Specifically, negative emotions are thought to restrict learners' attention and cognitive resources, whereas positive emotions play a facilitative function by broadening learners' perspectives (Fredrickson, 2004). Given that emotions can fluctuate rapidly during language production (Boudreau et al., 2018), the cognitive processes that impact speech fluency during communication may also undergo temporary changes.

The role that emotion and cognition play in speech production can be interpreted through Levelt's (1989, 1999) model of first language speech production, which divides the speech production process into three sequential stages: *conceptualization*, when the speaker decides the relevant content to use; *formulation*, when the speaker encodes the content with relevant lexis and grammar in preparation for expression; and *articulation*, where a phonological plan is mapped onto the formulated message. In the case of L2 learners, disfluent speech is related to their inability to process L2 speech efficiently in real-time communication (Kormos, 2006; Skehan, 2014). Thus, momentary fluctuations in emotions may cause momentary disturbances in cognitive processing at the conceptualization or formulation stage, enabling more or less fluent speech at times throughout a performance. However, researching the dynamic role emotion plays in speech production during tasks requires innovative methods to capture the within-person variability of emotion at a micro-level. Such methods are beyond the prevalent "post-task questionnaire" approach used in task-based language teaching (TBLT) research.

To address this issue, the current study uses the idiodynamic approach (MacIntyre and Ducker, 2022) to examine the relationship between anxiety, enjoyment and breakdown fluency (average length of pausing) during monolog L2 speaking tasks. This relationship is explored by dividing participants' monolog task performances into short, segmented units of speech, establishing per-person correlations between the two emotion variables and breakdown fluency, and exploring reasons behind momentary or prolonged changes in enjoyment and anxiety during task performances. This research is potentially important as it establishes an agenda for investigating emotions and language production in TBLT research from a complex dynamics systems theory (CDST) perspective.

Literature review

Emotions in second language acquisition

Emotion can be described as consisting of "neural circuits (that are at least partially dedicated), response systems, and a feeling state/process that motivates and organizes cognition and action" (Izard, 2010, p. 367). SLA researchers have gone so far as to suggest that emotion and cognition are "inseparable" (Swain, 2013). At the very least, emotions are thought to direct cognitive resources, affecting attention, memory, and behavior (Schumann, 1994; Pekrun, 2006). A spectrum of emotions has been identified as influencing language learning and communication. These include *positive emotions* such as enjoyment, love, pride, or gratitude, and *negative* emotions such as anxiety, boredom, guilt, and anger (for a review, see Shao et al., 2020). The characteristics of emotions can be understood through the multidimensional model of affect (Linnenbrink, 2007), which describes emotions as having dimensions of valence (e.g., *love*: negative; *guilt*: positive) and activation (e.g., *anxiety*: high activation; *boredom*: low activation). An implication of this model is that both negative and positive emotions can be highly activating (e.g., *anxiety, enjoyment*), energizing learners in terms of motivation, effort, and performance. From another perspective, the control-value theory framework (Pekrun, 2006) posits that emotions can be conceptualized in terms of achievement. That is, they can be either connected to achievement activities (e.g., frustration from a difficult task), with resultant *anticipatory emotions* (e.g., hopelessness) or connected to achievement outcomes (e.g., anxiety of failure at task), with resultant *retrospective emotions* (e.g., shame). Such classifications are important in SLA as they provide frameworks for hypothesizing which emotions facilitate or inhibit learning behavior as well as how emotions are generated during learning tasks.

Of the multitude of emotions that language learners can experience, by far the most researched emotion in SLA is anxiety, defined as "the worry and negative emotional reaction aroused when learning or using a second language" (MacIntyre, 1999, p. 27). High levels of anxiety can be debilitating for language performance and hinder language acquisition (Horwitz, 2010; Gregersen and MacIntyre, 2014) as it can lead to disruptions in cognitive ability and cause one to withdrawal from involvement to seek self-protection (Arnold and Brown, 1999). However, as a high activation emotion, anxiety has also been shown to motivate learners, causing them to increase effort in the face of increased cognitive demands (Marcos-Llinàs and Garau, 2009).

Recent interest in positive psychology in SLA (MacIntyre and Gregersen, 2012; MacIntyre and Mercer, 2014) has resulted in an uptick in research investigating the role of positive emotions in language learning, particularly enjoyment.

Enjoyment can be broadly defined as an experience, event, or action that is perceived as offering joy, happiness, and pleasure (Brantmeier, 2005). Enjoyment is also used synonymously with *flow*, or "a deep, spontaneous involvement with the task at hand" (Csikszentmihalyi, 2014, p. 181). In a similar way, Dewaele and MacIntyre (2016) define enjoyment as "a complex emotion, capturing interacting dimensions of the challenge and perceived ability that reflects the human drive for success in the face of difficult tasks" (p. 216). Scholars, such as Csikszentmihalyi and Dewaele, thus emphasize the productive aspect of enjoyment, highlighting that it is a positive, activity-focused emotion.

The relationship between enjoyment and anxiety has been the target of recent interest in SLA (e.g., Dewaele and MacIntyre, 2014; Boudreau et al., 2018; Dewaele and Pavelescu, 2021). Rather than comprising opposite ends of the same dimension, current thinking suggests enjoyment and anxiety are independent emotions (Dewaele and MacIntyre, 2014). As enjoyment and anxiety are highly activating, they both serve to energize learners' behaviors in complementary ways (Linnenbrink, 2007). Supporting this view, Elahi Shirvan and Talebzadeh (2018a) found that, in the same L2 learning environment, learners experienced both anxiety, as a result of pressure from time limits or competitive goals, and enjoyment, caused by positive feedback or a sense of achievement. Both emotions "pushed [them] forward to take actions in the process of learning" (p. 127). Thus, an important emerging viewpoint is that enjoyment and anxiety do not always change together in a "see-saw fashion" (Dewaele and Pavelescu, 2021) but converge and diverge over time (Dewaele and MacIntyre, 2014). This nuanced perspective suggests that anxiety and enjoyment can be co-occurring emotions, and successful language learners must "embrace 'joy' as well as 'pain'" (Prior, 2019, p. 522).

Anxiety and enjoyment in spoken L2 tasks

One way to deepen our understanding of how anxiety and enjoyment emerge in language learning activities, their relationship to each other, and their impact on cognition, is through examining learners' performances during L2 tasks. In SLA, the study of emotion in L2 speaking has been researched from the perspective of both TBLT and L2 psychology.

TBLT research has a pedagogic focus and views tasks as distinct units of learning that create meaningful communication situations where the L2 can be learned *through* language use (Ellis et al., 2020). Recent research in this vein seeks to understand how task design and implementation conditions can affect learners' emotional responses in relation to task performance. A consistent finding is that designing tasks with familiar content can engender greater levels of interest, enthusiasm, confidence, and enjoyment (Aubrey, 2017a,b, 2022; Lambert et al., 2017, 2022; Qiu and Lo, 2017; Lambert and Zhang, 2019; Aubrey et al., 2020; Nakamura et al., 2021). Such tasks have an embedded dimension of learner control (e.g., students choose their own photo in a photo-description task) and so provide learners with a sense of meaning and autonomy—a key pre-condition for the emergence of positive emotions (Csikszentmihalyi et al., 2005; Aubrey, 2017b, 2021). However, the effect of learner control on both enjoyment and anxiety can be complex. For example, Nakamura et al. (2021) found that giving learners such control—in their case, choice over task content—led to increases in anxiety *and* enjoyment. This suggests that when learners are free to use their own ideas and language during a task, they may enter into a vulnerable, anxious state, but this state is necessary for the kind of exploration and play that generates feelings of enjoyment. Certain task implementation factors, particularly, pre-task planning, can also affect learners' task emotions (e.g., Kim, 2013; Bui and Teng, 2018). As planning can reduce the cognitive burden of conceptualization and formulation (Skehan, 2018), planning content and language to be used may lead to overall reduced anxiety during performances. However, as Bui and Teng (2018) argue, some learners prefer not to plan because preparation constrains learners' choices and limits their ability to speak extemporaneously on a topic. In a similar way, repeating a task may alleviate anxiety as learners need to devote progressively less attention to lexical retrieval processes with each attempt (Bygate and Samuda, 2005), but too many repetitions of the same task can cause boredom/apathy as they become "tired of doing the same thing" (Kim, 2013, p. 17).

Despite the aforementioned insights, methodologically, TBLT research has been mostly limited to using retrospective post-task questionnaires to measure emotional experiences. Thus, task-based research has largely ignored the substantial within-person variation in emotion levels that occur *during* a task. Filling this gap, researchers in L2 psychology have developed new methods which aim to describe "the complex intraindividual emotional reactions and changes in self-perception, not group averages, that occur during brief episodes... of L2 communication" (MacIntyre and Ducker, 2022, p. 1). These studies employ the idiodynamic method, which involves recording an episode of task communication, having participants repeatedly rate affective states on a per-second basis using computer software while watching their recorded performance, and conducting a stimulated recall with participants to elicit reasons for changes in ratings. While idiodynamic studies often measure conative variables, such as willingness to communicate (WTC) (e.g., MacIntyre and Legatto, 2011; Ducker, 2021), others have measured dynamic changes in emotions, such as anxiety (Gregersen et al., 2014; Elahi Shirvan and Talebzadeh, 2018b; MacIntyre, 2019; Macintyre and Gregersen, 2022), enjoyment (Talebzadeh et al., 2020) or both anxiety and enjoyment (Boudreau et al., 2018). These studies take a CDST perspective (Larsen-Freeman and Cameron, 2008), demonstrating that emotions fluctuate

during communication as learners attempt to cope with ongoing communication demands.

Most idiodynamic studies have employed dyadic tasks (conversations, interviews). A common finding is that changes in emotions tend to be related to interpersonal issues. For example, Elahi Shirvan and Talebzadeh (2018b) found that learners' anxiety levels faced constant change, which was influenced by ongoing self-comparisons with their interlocutor, as well as verbal and non-verbal feedback. Similarly, Macintyre and Gregersen (2022), who examined conversations between cross-cultural dyads, revealed that anxiety stemmed from difficulties in comprehending accents from unfamiliar interlocutors, while decreases in anxiety were attributed to finding common ground or changing to more familiar topics mid-conversation. Finally, Talebzadeh et al. (2020) investigated enjoyment during teacher-student conversations. Results revealed that automatic mimicry (i.e., mirroring an interlocutor's behavior) was the main mechanism through which enjoyment was transferred from between student and teacher (e.g., facial expressions, posture, movement, and vocalization). An implication of these studies is that emotions often emerge because of a desire to maintain interpersonal harmony.

Gregersen et al. (2014) and Boudreau et al. (2018) both report on investigations that are most relevant to the current study. Gregersen et al. (2014) triangulated idiodynamic ratings of anxiety levels and physiological responses (heartrate) for learners during presentation tasks in a Spanish as a foreign language class. They found that anxiety levels were highly volatile, with momentary fluctuations attributed to the presentation topic, learners' vocabulary knowledge, the audience's reaction, and learners' ongoing evaluation of the task experience. A notable finding was that when learners attempted to memorize presentations word-for-word, they were vulnerable to spikes in anxiety caused by problems recalling specific planned vocabulary. This sometimes led to continually rising anxiety levels, exacerbated by learners' conscious awareness of their anxious state. Underscoring the significance of pre-task planning strategies, Gregersen et al. (2014) notes that learners who planned general ideas but spoke extemporaneously on their topic (i.e., they did not rely on a memorized script) could maintain lower levels of anxiety. Boudreau et al. (2018) conducted an intriguing examination of the rapidly changing relationship between enjoyment and anxiety during a variety of tasks performed by L2 learners. They found that per-person correlations between anxiety and enjoyment (i.e., correlations on individual learner data) varied from highly negative to almost zero, indicating considerable variation between learners. Inspecting intra-individual fluctuation patterns revealed that, on a moment-to-moment timescale, enjoyment and anxiety interacted continuously in complex ways. This dynamic relationship partly depended on task type, with the photo description task engendering both positive and negative emotions as learners' past experiences depicted in the photos influenced emotions felt during communication. Similar to Gregersen et al. (2014), momentary surges in anxiety were frequently associated with cognitive-linguistic difficulties (e.g., problems retrieving a word from memory).

In sum, idiodynamic studies have focused mostly on anxiety fluctuation patterns in dyadic conversations. They have demonstrated that emotions fluctuate rapidly with corresponding changes in learners' cognitive processing ability. They also show that anxiety and enjoyment have a complex relationship that is influenced by a wide range of factors, including task factors, pre-task preparation, characteristics of interlocutors, and characteristics of the learners themselves.

L2 speech fluency

Fluency in L2 task performances can be measured in several ways: *speed fluency* (or speech rate), which is generally measured by the number of syllables of pruned discourse per second (see Ellis and Barkhuizen, 2005; Ch. 7); *repair fluency*, which is operationalized as the frequency of overt repairs or reformulations (see Gotz, 2013); and *breakdown fluency*, as measured by pause frequency (see Saito et al., 2018). Such variables comprise cognitive fluency (Segalowitz, 2010) and reflect the level of efficiency of speech processing mechanisms (i.e., conceptualization, formulation, and articulation) (Levelt, 1989, 1999).

Of interest to this study is breakdown fluency. Higher breakdown fluency can signal a problem related to conceptualization of ideas (Lambert et al., 2021) or a breakdown in the linguistic encoding process (Kormos, 2006; Gotz, 2013). Moreover, breakdown fluency can also vary as a function of the tasks performed (Tavakoli and Skehan, 2005). For example, structured tasks, or tasks that have predictable discourse patterns (e.g., problem-solution tasks, see Leeming et al., 2020; Lambert et al., 2021; Aubrey, 2022; Aubrey and Philpott, 2022; Aubrey et al., 2022) can aid learners in producing more fluent speech (Ahmadian and Tavakoli, 2011; Ahmadian, 2015). These structured tasks might decrease breakdown fluency as they provide learners with multiple "starting points" from which they can gain control over content, which frees attentional resources for conceptualization and formulation processes (Skehan, 2014).

Breakdown fluency might also be closely related to emotion. For example, pausing has been associated with highly stressful moments as a result of momentary "jams" in cognitive processing (MacIntyre and Legatto, 2011; Gregersen et al., 2014; MacIntyre and Serroul, 2015; Boudreau et al., 2018; Macintyre and Gregersen, 2022). Examining fluency more directly, Wood (2016) measured dynamic WTC and fluency in Japanese learners of English and found that, in general, when WTC is high, fluency is high, and when WTC is

low, fluency is low. However, the influence of one variable on the other occurred in both directions. That is, fluency sometimes influenced WTC while WTC sometimes influenced fluency. This two-way interaction was similarly observed in Nematizadeh and Wood's (2019) case study in which L2 learners of Farsi completed a picture description task. They found that high WTC sometimes coincided with frequent pausing when learners had several ideas but hesitated in choosing appropriate lexis and grammar. These studies are important as they shed light on how problems at different stages in speech processing (e.g., idea conceptualization, encoding ideas with language) can lead to breakdown in fluency. It is notable, however, that claims about relationships between emotions and fluency have so far been either speculative (e.g., inferring emotional states via WTC ratings) or qualitative (e.g., inferring emotional states via interviews only). Seeking to advance this line of research, the current study attempts to connect emotion ratings to language production, first quantitatively, using inferential statistics (correlations), and then qualitatively, by exploring reasons for emotional change. Specifically, this research aims to answer the following questions.

1. Based on pausing behavior and idiodynamic ratings, what are the per-person relationships between anxiety, enjoyment, and breakdown fluency during problem-solving task performances?
2. Based on stimulated recall interviews, what factors influence the relationships between anxiety, enjoyment, and breakdown fluency at moments of emotional change during problem-solving task performances?

Materials and methods

Participants

Participants included four learners of English who were attending a university in Hong Kong. Based on information from a background questionnaire, all participants spoke Cantonese as their first language (L1) and had scored "4" on the English language subject level of their Hong Kong Diploma of Secondary School Exam (HKDSE), which is benchmarked to the IELTS score range of 6.31–6.51 (Hong Kong Examinations and Assessment Authority, Hong Kong Examinations and Assessment Authority) and equivalent to the Common European Framework of Reference (CEFR) B2/C1 level. Participants were born and raised in Hong Kong and reported having no experience living in an overseas English-speaking country. Despite these common characteristics, there were differences in age, gender, major at university, and use of English, which are summarized in **Table 1**.

Instruments

Participants performed an oral problem-solution task (for the task instructions, see **Appendix 1**). This task closely followed versions found in recent research (Leeming et al., 2020; Lambert et al., 2021; Aubrey and Philpott, 2022; Aubrey et al., 2022). Learners were required to individually give a 3-min monolog that explained a problem, compared possible solutions, and recommended a solution with reasons. Before the performance, participants were given 10 min to plan for the task by completing a planning worksheet. The worksheet contained the task instructions, the task problem description, and questions to guide learners to understand the situation and problem and to conceptualize their responses and evaluations of their responses (for the planning worksheet, see **Appendix 2**). The task required participants to consider a problem related to lack of interaction between local and international students on campus that provided the data which was used for this study. In conducting speaking tasks with learners under laboratory conditions similar to the current study, MacIntyre (2019) observed that "an awareness of the camera and researchers" can impact the quality of data collected (p. 6). Thus, to optimize reliability of data collection instruments, a practice task (same task type, different topic) was done with each participant beforehand to orient them to the structure of the desired task performance, familiarize them with the research environment, and train them in the idiodynamic software procedures (see procedures below).

Procedures

Data were collected during separate sessions with participants. During the task planning stage, the participant made notes on the planning worksheet. The worksheet was collected by the researcher after the 10 min of planning time. A video camera recorded the 3-min speech performance. A timer set for 3 min was placed in front of the participant during the task.

Immediately following the task, the participant watched his/her task performance video and rated their anxiety on a per-second basis using idiodynamic software (for a detailed description of the idiodynamic method, see MacIntyre and Ducker, 2022). The software played the video and simultaneously collected participants' ratings by recording clicks from a computer mouse. The software displayed the ratings on-screen with colored bars showing increasing or decreasing ratings. The software features an auto-zero function that returned the rating to zero if the mouse was not clicked. The video was then watched a second time to rate enjoyment.

While it is possible that there may be inconsistencies in how different participants used the rating software (Boudreau et al., 2018), learners in this study were provided with feedback

TABLE 1 Summary of participant information.

Pseudonym	Gender	Age	Major	Use of English with family or friends	Use of English in the classroom
Isabel	F	19	Hospitality and real estate	Not often	Not often
Diana	F	19	Chinese language education	Never	Never
Brenda	F	22	Chinese language education	Not often	Not often
Travis	M	22	Chinese language education	Sometimes	Sometimes

on their use of the software during a practice rating session to mitigate such inconsistencies during data collection. During this practice session, learners were also provided with a written definition of anxiety and enjoyment so that they could more intuitively understand what was required from them.

After learners completed their ratings, data were then copied to a Microsoft Excel sheet. A graph of the participant's anxiety and enjoyment ratings were displayed on a computer and shown to the participant to begin the interview (see **Figures 1–4**). The researcher and the participant looked at the graph together and discussed the trends for each variable across the 3-min period. Examples of questions that were asked by the researcher include: Why did you rate anxiety/enjoyment low at this stage of the task? Can you explain why your ratings remained stable but then increased over this time interval? Why did your ratings for anxiety/enjoyment in the final minutes of the task suddenly decrease? When necessary, the researcher played key segments of the video again for the participant to assist in their recall of specific segments of speech. The participant was also asked to self-evaluate their task performance more generally. A video camera recorded the interviews.

Analysis

The first aim of the analysis was to produce per-person correlations between enjoyment, anxiety, and breakdown fluency. As an initial step, the 3-min task performances were transcribed by a research assistant (RA) and verified by the researcher. To obtain a measure of breakdown fluency, the task performance sound files were analyzed with PRAAT 6.053 (Boersma and Weenink, 2019) to identify silent pauses of 0.25 s or more (De Jong and Bosker, 2013). The length and position of each of these pauses were added to the transcripts. After initial discussion and practice, interrater reliability was established between the RA who independently coded the transcripts. Pearson product moment correlation coefficients revealed high interrater reliability on the measure of pauses (r = 0.921). The transcript was then divided into 25 7-s segments and one 5-s segment (180 s total, or 3 min). This length of segment was chosen because it represented the largest interval of time that any participant produced a run of fluent speech without pausing. This number of data points also satisfied minimum thresholds considered acceptable for inferential statistics (Bonett and Wright, 2000). The mean length of pausing per second for each segment was calculated. Examples of two consecutive segments, with length of pauses (seconds) indicated in parentheses, are provided below:

Example 1: 7-second segment 1 (mean length of pausing per second = 0.69 s):

> uh (1.65) vibes (0.65) because (0.56) uh (0.35) we (0.38) we all have (0.74) different cultures (0.51) and

Example 2: 7-s segment 2 (mean length of pausing per second = 0.34 s):

> (1.10) we gather and meet each other and many different country people in ah (0.82). the universities (0.45).

Following a precedent for examining idiodynamic ratings in different segments of speech (see MacIntyre and Serroul, 2015; Macintyre and Gregersen, 2022), mean anxiety and enjoyment ratings were calculated for each 7-s speech segment. This was deemed necessary for performing the correlations to answer the first research question. Using the resulting 26 data points, Pearson correlations for each participant were determined between anxiety, enjoyment, and fluency. This approach provides an estimate of the strength of the relationships between task emotions and fluency.

The secondary aim of the analysis was to identify learners' reasons for patterns of enjoyment and anxiety in the tasks. All interviews were transcribed and examined for specific rationales given for changes in emotions.

Results

To answer the first research question, measures of anxiety, enjoyment, and breakdown fluency for each individual performance were subjected to correlation analysis. The correlations are shown in **Table 2**. As can be seen, there were significant positive correlations (strong to moderate) between anxiety and breakdown fluency for three out of the four task performances (Isabel, Brenda, Diana) and one non-significant, very weak correlation (Travis). In contrast, only one task performance produced a significant, moderate correlation between enjoyment and breakdown fluency (Isabel),

TABLE 2 Correlations between per-7-s ratings of enjoyment, anxiety and breakdown fluency for each participant.

	Correlation with breakdown fluency		Correlation between anxiety and enjoyment (*r*)
Participant	**Anxiety (*r*)**	**Enjoyment (*r*)**	
Isabel	0.65**	−0.43*	−0.71**
Diana	0.44*	0.05	−0.27
Brenda	0.51*	−0.38	−0.37
Travis	0.14	−0.13	−0.16

** Is significant at the $p < 0.01$ level; * is significant at the $p < 0.05$ level.

while others were either non-significant and weak (Brenda), very weak (Travis), and near zero (Diana). Isabel's task experience produced the only significant and strongly negative correlation between anxiety and enjoyment; others produced non-significant and weak correlations (Brenda, Diana, Travis). Thus, we can see overall positive (anxiety and fluency) and negative relationships (anxiety and enjoyment), but the strength of these correlations varied widely among participants.

To answer research question 2, each individual task experience will now be examined more closely. It should be noted that, during the interviews, participants often repeated reasons for emotional changes; thus, the following interview summary does not tabulate these reasons, but rather provides illustrative examples in full context for each reason given by each participant.

Isabel's task performance

Isabel described herself as "not good at speaking fluent English" and claimed that her "ability in grammar and tenses is weak." However, she felt confident about her task performance, which she attributed to her familiarity with the task topic and opportunity to plan for the task. Among all participants, Isabel's idiodynamic ratings were second lowest for anxiety ($M = -0.1$; $SD = 0.85$), second highest for enjoyment ($M = 0.13$; $SD = 0.63$), and her breakdown fluency was average ($M = 0.29$; $SD = 0.16$). Her significant (strong to moderate) correlations indicate there are some regular patterns in the way her anxiety, enjoyment and breakdown fluency fluctuate together throughout the task. These patterns are reflected in Figure 1.

Isabel exhibited more frequent breakdowns, higher anxiety, and lower enjoyment to start the task. The period from 49 to 77 s was particularly unenjoyable and anxiety-provoking, with a higher-than-average breakdown fluency. In the corresponding speech segment below, frequent pausing coincided with the use of hesitation devices and false starts:

> . . . accept (0.35) some ah (1.67) cultural (0.89) habits or chara- (0.25) characteristics of different (0.85) people (1.29) For example, cannot- uh (1.68) we are not- (1.16) we will- (1.29) we cannot uh (0.57) accept (0.59) some Indian people having their hands to eat (0.59) because (0.79) we-(0.58) many of us would think that's dirty (0.46) and maybe we (1.01) don't.

In explaining this period, Isabel described cognitive problems associated with the struggle to organize and encode ideas with the L2, causing a highly anxious state: "many things come up in my brain at that moment, so I need to organize them again and I feel anxious." A transition point occurred at the 77–91-s period of the task, in which enjoyment spiked, anxiety dropped to its lowest point, and breakdowns decreased. Isabel attributed this to the problem-solution structure of the task: "at this point, I changed [from describing problems to giving solutions]- I transition successfully. . . I start giving some solutions more easily." As she transitioned from the "problem" stage to the "solution" stage, she could draw on new ideas that she was more confident using. This transition seemed to launch her into a period of low anxiety, high enjoyment, and low breakdown fluency that lasted for the remaining task performance. Of particular interest is the period 126–140 s, in which she experienced her highest enjoyment, lowest anxiety, and decreasing breakdown fluency. This portion of the task performance is given below:

> . . . the better solution is that we provide more (0.28) interactive activities so that (0.35) uh (0.59) we can- for example we hold some com (0.32) competition we can gather some different countries (0.31) people to get into a team, (0.30) therefore they will have common language, topic and (0.30) they can have some.

Isabel emphasized that her emotional change in this segment was linked to a surge in confidence from using an idea that she had selected during planning as her recommended solution for the task problem: "I suggest the better solution here. . . I can imagine how it works when I planned it. . . it seems more organized." This example offers some evidence to support Isabel's general view that planning improved her self-confidence and optimism in expressing herself. Utilizing a pre-planned idea likely lessened cognitive load (i.e., conceptualization) during this portion of the task performances so that Isabel could

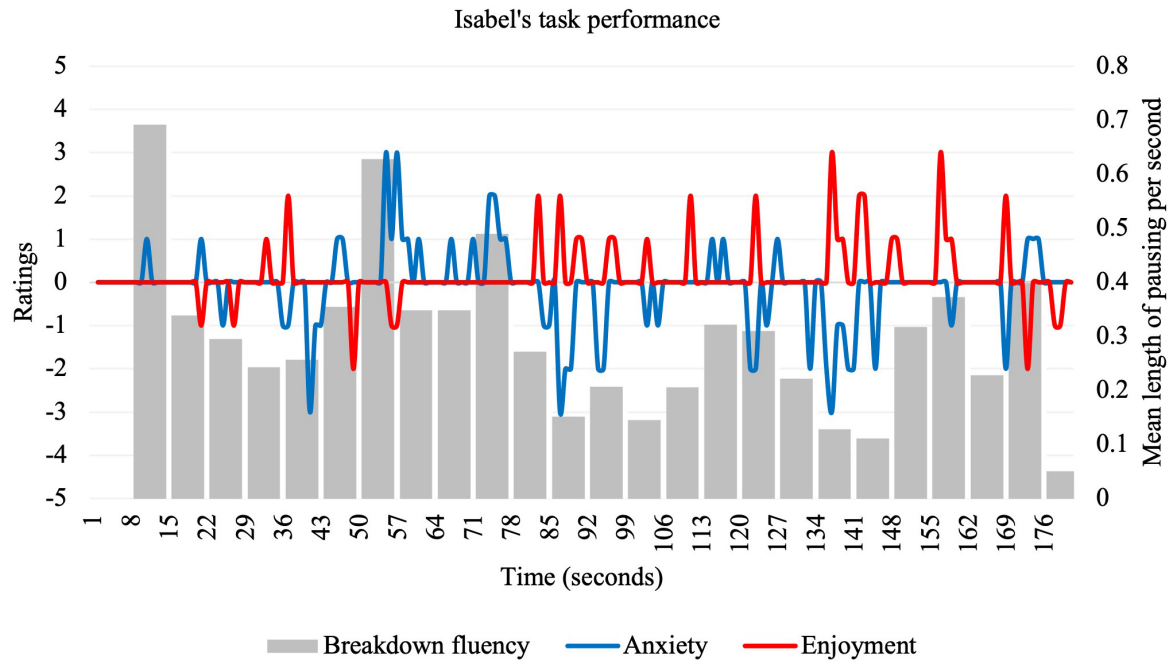


FIGURE 1
Idiodynamic ratings and breakdown fluency per segment for Isabel's task performance.

direct cognitive resources toward producing fluent speech (i.e., formulation and articulation).

Diana's task performance

Diana described herself as competent at using English for "simple communication" but said that she often needs to use Google translate. She reported never using English at home. When she uses English with friends, her communication is usually confined to written chat messages using the WhatsApp mobile application. Compared to other participants, Diana's idiodynamic ratings were average for both anxiety (M = 0.11; SD = 1.75) and enjoyment (M = 0.03; SD = 1.00), and exhibited average breakdown fluency (M = 0.29; SD = 0.19), but variance for each was higher. The more rapid and extreme fluctuations for ratings and breakdown fluency can be seen in Figure 2. For Diana, anxiety tended to sharply rise and fall with similar changes in pausing during the task.

When the interviewer drew Diana's attention to the several peaks in anxiety, she immediately attributed these to an inability to retrieve lexis from memory: "Yeah um mostly it is I cannot find the right word to say." As can be seen in Figure 2, these peaks coincide with increases in mean length of pauses, suggesting a problem during the formulation stage of speech production:

A noticeable change in anxiety during the task was the 56–77-s period. She started with a peak anxiety of 4 and ended with the lowest anxiety of the task, −5. This period also coincided with a decrease in average pausing and an increase in enjoyment. Diana pointed out that this was the time when she transitioned from describing the problem to giving a solution, which can be seen in corresponding speech segment below:

> . . . and actually local students are pretty ah (1.12) stubborn (0.36) because (0.86) we (0.25) kind of we can get kind of intimidated by other culture (0.36) So there are some solutions to this problem (0.25). Namely (0.25) firstly (0.40) the university can offer more scholarships so that um (0.36) more university students can (0.35) go on exchange tours and then (0.27) increase their different exposure to other cultures.

A characteristic of Diana's speech was that breakdown fluency and anxiety were often unrelated to enjoyment. For example, her highest peak in enjoyment was during the 98–106-s period despite also experiencing an increase in pausing. This corresponds to the following speech segment:

> . . . and then ah (1.71) the whole university can be more inclusive to (0.88) different cultures.

Explaining why she felt enjoyment, she described that she planned to talk about cultural inclusivity as a key point, and this was an opportunity to implement her plan: "I was enjoying the speech when I say about cultural inclusive. . . [I

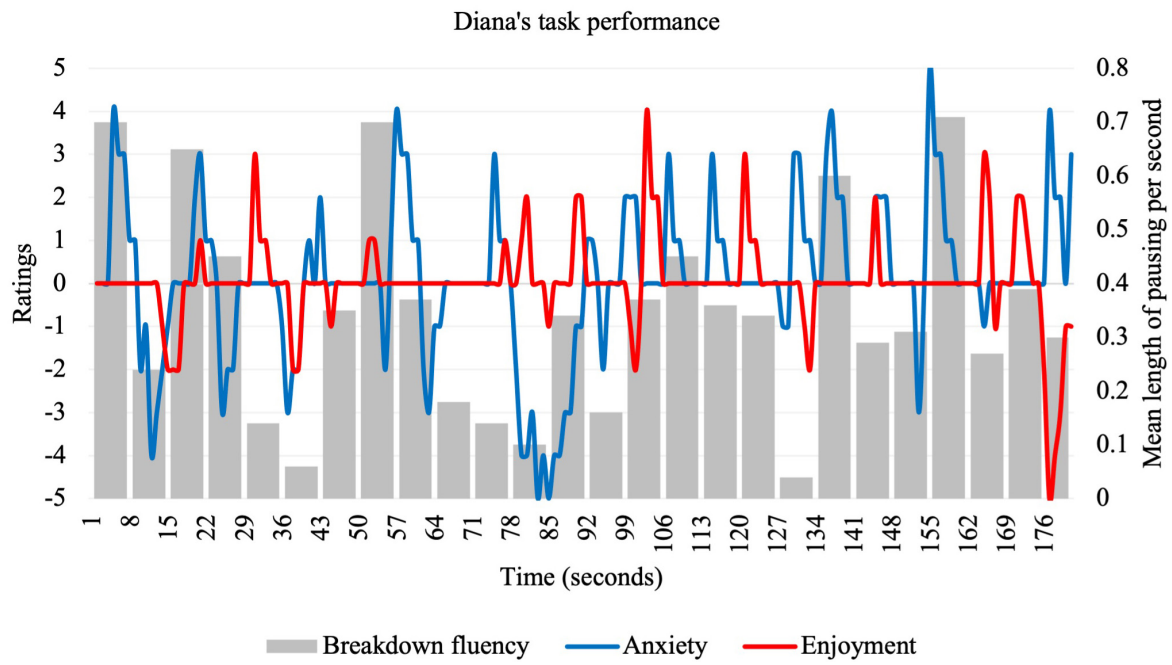


FIGURE 2
Idiodynamic ratings and breakdown fluency per segment for Diana's task performance.

tried] to put inclusive part into my speech here." However, when asked about why she also experienced an immediate spike in anxiety, she described her mixed feelings about planning:

> [Planning] helps reducing of anxiety for me because I have a rough- rough idea of what I should talk about. . . but also it can restrict like what I say- it can restrict what I want to say because. . . I cannot say anything else that I want to say.

Thus, it seems that, for Diana, planning led to both self-confidence in using ideas but also a loss of freedom to choose what ideas to express in the moment. Toward the end of the task, Diana commented on the following segment of speech where enjoyment dipped, anxiety spiked but breakdown fluency decreased to its lowest point (126–133 s):

> . . . their experience to share their culture with the local students in a very (0.28) deep way.

Despite occurring during a fluent run of speech, Diana's negative emotions were aroused by her use of "deep way"—a phrase that she judged to be inappropriate:

> I think I was talking about guest speakers coming to give talks about their own cultures in a deep way and. . . I said "deep way" and I was very very disappointed, so I was less enjoyable.

Even though Diana used the phrase in a fluent segment of speech, she negatively evaluated her language use which aroused an immediate feeling of disappointment.

Brenda's task performance

Brenda described herself as proficient at "basic communication" in English but emphasized that she makes grammar mistakes easily. In general, she doesn't use English with family or friends and uses English only when required in classes.

Compared to other participants, Brenda reported the highest anxiety ratings (M = 0.18; SD = 0.74), the lowest enjoyment ratings (M = −0.19; SD = 0.79), and exhibited the highest breakdown fluency (M = 0.31; SD = 0.16). Similar to Isabel, Brenda's breakdown fluency significantly correlated with anxiety (positive, weak) but non-significantly correlated with enjoyment (negative, very weak). Fluctuation patters are shown in Figure 3.

Brenda tended to increase her anxiety and breakdown frequency as the task progressed. To prompt Brenda to explain why she ended performance more anxious than she started, she was played the final speech segment (168–180 s) again:

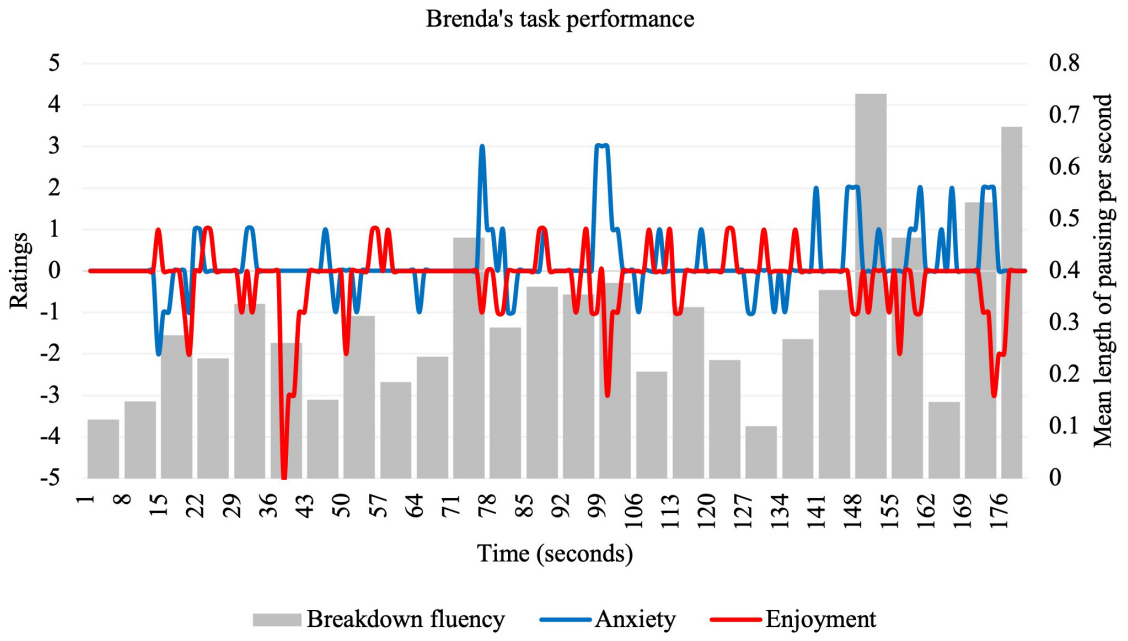


FIGURE 3
Idiodynamic ratings and breakdown fluency per segment for Brenda's task performance.

> . . . it won't be (0.25) too (2.20) too crowded (1.27) and we can leave (2.2) some space for (1.85) students (0.70) to know each other.

Brenda attributed this high anxiety/low enjoyment/high pausing period to cognitive problems (i.e., conceptualizing the idea and struggling to formulate words), which was exacerbated by an awareness of time constraints:

> I'm not very fluent [at this time] . . . when time is running out, my head cannot function. . . I have a lot of ideas in my mind, but I cannot express it clearly, so I became very nervous. I didn't enjoy it very much.

An exception to Brenda's relatively successful start to the task was the 35–42-s segment, in which enjoyment dipped to its lowest point at −5. This period corresponds to the following segment of speech:

> . . . comp (1.31) the wish can come true (0.51) and one of the causes may be that.

Brenda connected her decease in enjoyment to her use of the phrase "wish can come true":

> It's the sentence "the wish can come true". . . I think it's a silly sentence. . .. I really hate "the wish can come true" part.

Thus, similar to Diana, Brenda's negative evaluation of how she used language led to an immediate emotional response that seemingly did not impact her fluency.

Travis's task performance

In contrast to other participants, Travis reported that he often speaks to friends and family in English outside of class, and in the interview, expressed a reasonable level of comfort when speaking English. He evaluated his task performance favorably, describing it as "organized" and commending himself on his "time management." Reflecting his confidence, compared to other participants, Travis exhibited the lowest ratings for anxiety (M = −0.22; SD = 0.78) and the highest ratings for enjoyment (M = 0.21; SD = 0.59), with relatively low variance for each. Breakdown fluency during the task (M = 0.30; SD = 0.16) was comparable to other participants. Figure 4 visually displays fluctuations for the three key variables.

His relatively enjoyable task experience was corroborated by the interview. Travis repeatedly attributed his positive emotions to the personal nature of the topic. In one instance, he explained that he is "currently living with a foreign roommate"—a situation which provided him with ideas he could draw on to use in the task, which generated a spike in enjoyment, decrease in anxiety, and a decrease in breakdown fluency (63–77-s):

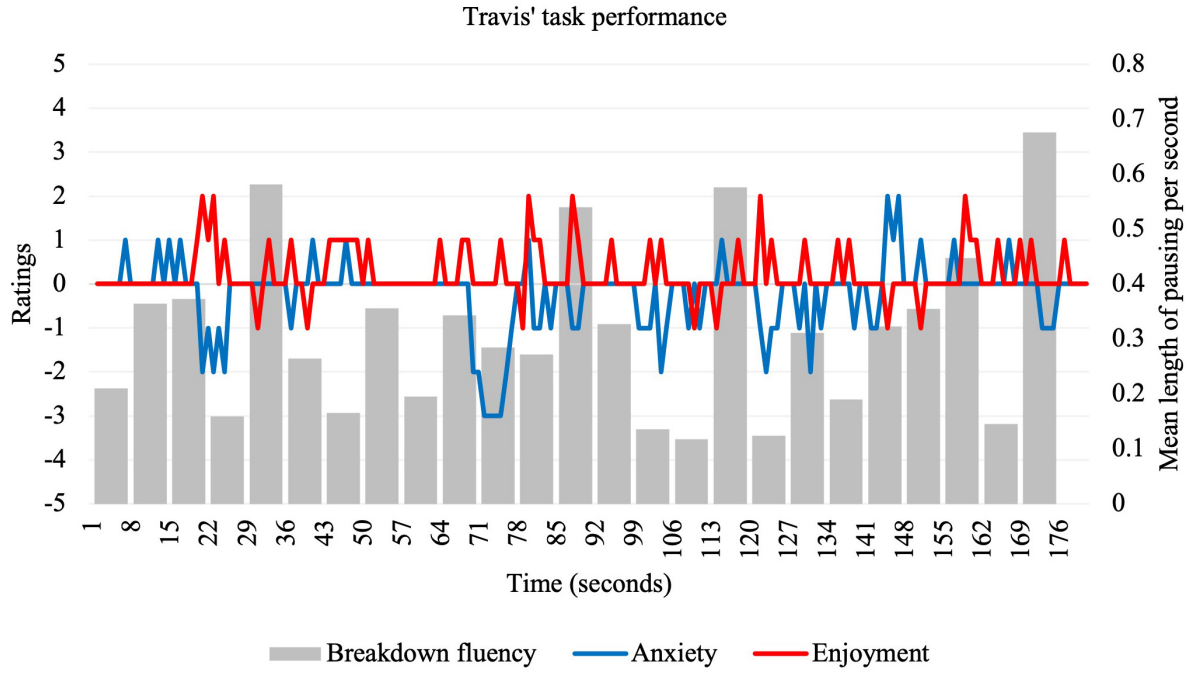


FIGURE 4
Idiodynamic ratings and breakdown fluency per segment for Travis's task performance.

> . . . is that (0.49) there's a policy that can apply in the student residence (0.38) which is to (1.53) um (0.80) ask (0.33) the international students to be allocated with (0.86) one local students.

He explained that the personal nature of this speech segment led to a positive emotional response:

> I am totally talking about my personal experience, and I am confident in delivering that- it's because that's what we are trying to do in the residence.

Travis also described his appropriate word choices as influencing feelings of enjoyment. He referred to the following speech segment (14–28 s) which coincided with high enjoyment and low anxiety:

> . . . in Hong Kong (1.33) actually so many so called, like, (1.27) multi-cultural university, (0.51) the international students and the local students just couldn't (0.60) mingle together.

Travis noted that he "used the word 'mingle.'" This was significant because it was a word he would not normally use, yet he found it matched his idea well. He rated this moment highly enjoyable as it engendered a sense of satisfaction and pride.

Travis also provided insight into instances where he felt both enjoyment and anxiety simultaneously. Referring to the period 42–49 s, he commented:

> I think I am quite satisfied with my delivering of the idea but at the same time I was thinking 'oh I couldn't talk about the cultural part'. . . I planned about saying this. . ..so it gave me the mix of enjoyment and anxiety. . . I wanted to talk more but I don't have the time.

Travis thus connected positive emotion to confidently and fluently expressing the idea, but also felt negative emotions associated with regret for not having time to express another planned idea.

The sole instance he felt highly anxious was during the following 142–149-s speech segment:

> . . . just kind of mandatory for them to (1.26) to- (1.00) to try to understand more about.

His explanation for anxiety echoed other participants' reasons related to an inability to retrieve lexis: "I think during that part I was hesitant because I couldn't find the right word."

A summary of reasons provided by each participant for changes in emotional levels during the task, as well as the corresponding directions of emotion and breakdown fluency change, is given in Table 3.

TABLE 3 Perceived factors leading to change in emotions and breakdown fluency during the task.

Reason for change in emotions	Change in anxiety	Change in enjoyment	Change in breakdown fluency
Struggle with formulating language	Isabel + Diana + Travis +	Isabel − Diana − Travis −	Isabel + Diana + Travis +
Struggle with conceptualizing/ organizing ideas	Diana + Brenda +	Diana − Brenda −	Diana + Brenda +
Self-evaluating language used	Travis − Diana + Brenda +	Travis + Diana − Brenda −	Travis − Diana − Brenda −
Personalization of task topic	Travis −	Travis +	Travis −
Task time limits	Brenda +	Brenda −	Brenda +
Task structure	Isabel − Diana −	Isabel + Diana +	Isabel − Diana −
Use of ideas/language from task planning	Isabel − Diana ± Travis +	Isabel + Diana + Travis +	Isabel − Diana + Travis −

Discussion

The first research question asked about the intra-individual relationships between emotions and breakdown fluency. Three out of four learners (Isabel, Diana, Brenda) exhibited significant positive correlations between anxiety and breakdown fluency (see **Table 2**). This not only suggests that, for most learners, anxiety was debilitating for language performance (Horwitz, 2010; Gregersen and MacIntyre, 2014), but that momentary increases in anxiety tended to lead to momentary increases in length of pauses (and vice versa) during speaking tasks. Compared to anxiety, the relationship between enjoyment and breakdown fluency was weak, as only one learner (Isabel) exhibited a significant negative relationship (see **Table 2**). Isabel's feelings of enjoyment might have fostered a more open, less inhibited state (Fredrickson, 2004) that coincided with less hesitant speech. However, for most learners, a diverse set of factors—not directly related to breakdown fluency—may have been behind changes in enjoyment during tasks. Similarly, Isabel was the only learner who exhibited a significant negative correlation between anxiety and enjoyment (see **Table 2**). This is consistent with the notion that anxiety and enjoyment are independent emotions that do not always change together in a "see-saw fashion" (Dewaele and Pavelescu, 2021), even at a per-second timescale during short communication tasks (Boudreau et al., 2018). The substantial variation in the strength of correlations between learners suggests that, even when learners are exposed to the same task, they each have unique emotional task experiences.

The second research question asked about the factors that influenced the relationship between enjoyment and anxiety. Learners' reflections were elicited from graphs of anxiety and enjoyment ratings (see **Figures 1–4**) and centered around significant emotional experiences during the tasks. Reflecting characteristics of a dynamic system (Larsen-Freeman and Cameron, 2008), these factors (see **Table 3**) were interconnected, rather than distinct, and exerted influence on different timescales. These influences affected learners in different ways, resulting in rapidly fluctuating anxiety and enjoyment, with some unexpected effects on fluency.

Reminiscent of task-based studies that have investigated learners' affective responses (e.g., Aubrey et al., 2020; Dao and Sato, 2021; Aubrey, 2022), *task-level* factors relate to task design (topic, task structure) and task implementation (time limits, task planning). These factors interacted with individual learner-internal characteristics, such as interests and experiences, to produce emotional change. Travis, for example, drew on his personal life experiences (e.g., living with a foreign roommate) for ideas used in his performance (e.g., suggestions for improving intercultural interaction), providing content to periodically sustain his fluency while also increasing his enjoyment. This reflects findings from previous research that suggests the alignment of task topic and personal experience engenders positive emotions (Lambert et al., 2021, 2022; Dewaele and Pavelescu, 2021; MacIntyre and Wang, 2021). Additionally, the problem-solution structure represented an interesting task design influence. Both Isabel and Diana reported the transition from describing the task problem to providing task solutions as a reason for decreasing anxiety and breakdown fluency. This is consistent with previous assertions that tasks with predictable discourse patterns aid learners in producing fluent speech (Ahmadian and Tavakoli, 2011; Ahmadian, 2015). It is interesting to note, however, that for Isabel, the transition from describing the problem to providing solutions seemed to trigger a positive effect that lasted for the duration of the task, but for Diana, the effect was only momentary.

Task implementation factors comprised two influences on emotions and fluency: time limitations and pre-task planning. For Brenda, the approaching time limit induced growing anxiety until it became the dominant emotion toward the end of the task. This led to disruptions in cognitive processing and a steady rise in average length of pauses. Thus, rather than causing a momentary spike, anxiety was continually affected (Gregersen et al., 2014; Boudreau et al., 2018). In terms of pre-task planning, three learners (Travis, Isabel, Diana) attributed planning to an increase in enjoyment, and, for two learners, a decrease in breakdown fluency (Travis, Isabel). This might be expected as preparation for a task is likely to increase self-confidence and optimism, creating the conditions for enjoyment. Pre-task planning has also previously been shown to aid in the conceptualization process, leading to greater fluency during the

subsequent performance (Skehan and Foster, 1997; Lambert et al., 2021; Aubrey and Philpott, 2022; Aubrey et al., 2022). However, Travis and Diana reported that planning led to surges in anxiety as they felt they had sacrificed chances to speak more extemporaneously on the task topic. Such mixed effects of planning reflect Bui and Teng's (2018) finding that some learners may prefer not to plan for a task, especially those who are of higher proficiency and familiar with the task topic (as was the case with Travis). This finding also supports Gregersen et al.'s (2014) view that certain pre-task planning strategies, in which learners feel they need to adhere strictly to what they have prepared, render learners more vulnerable to surges in anxiety during performances.

A second category could be described as *cognitive-linguistic* factors (momentary breakdowns in conceptualization and formulation) which relate to lower-level issues that have immediate consequences on speech production processes (Levelt, 1989, 1999). Most participants (Isabel, Diana, Travis) reported how the struggle to remember words led to momentary surges in anxiety, lowered enjoyment, and increased length in pausing, which is indicative of problems in speech processing. This result aligns with previous idiodynamic studies that suggest problems retrieving specific words co-occur with spikes in anxiety (MacIntyre and Legatto, 2011; Gregersen et al., 2014; MacIntyre and Serroul, 2015; Macintyre and Gregersen, 2022) and increased pausing behavior (Wood, 2016; Nematizadeh and Wood, 2019). Such findings provide the clearest evidence for the interconnectedness of cognitive, emotional, and linguistic subsystems, and reinforce the high anxiety/low enjoyment/high breakdown fluency relationship.

A final category could be described as *achievement outcome* factors, which emerged as retrospective emotional responses to language used. According to control-value theory of achievement emotions (Pekrun, 2006), retrospective emotions are outcome-focused emotions experienced when learners evaluate an outcome positively (success) or negatively (failure), which are associated with activating or deactivating effects, respectively. In this study, learners experienced surges in emotions when evaluating language used that they had either planned or placed considerable importance on. For example, when Travis used a word (*mingle*) that he deemed appropriate to explain the problem (*local and international students do not mingle*), he experienced a momentary surge in enjoyment, lowered anxiety, and decrease in breakdown fluency (i.e., the corresponding "activation" response). Such retrospective emotional responses are consistent with flow experiences in which individuals feel a sense of accomplishment that reaffirms they are performing an activity optimally (Nakamura and Csikszentmihalyi, 2002; Aubrey, 2017a,b). In contrast, Diana and Brenda experienced negative retrospective emotional responses when they perceived to have failed in producing an appropriate phrase that was important for expressing their ideas, leading to heightened anxiety and a decrease in enjoyment; interestingly, though, breakdown fluency decreased. This might be explained by the fact that the emotions triggered were low activation (e.g., shame experienced after using inappropriate language) and could be more easily overcome in this context (i.e., in a monolog task with no interlocutor). As retrospective emotions are rarely studied in the context of task performances, more research needs to investigate additional outcome emotions, which would help researchers understand how these moments of reflection energize task behavior.

Conclusion, limitations, and pedagogical implications

This study has explored the per-person relationships between enjoyment, anxiety, and breakdown fluency, and the factors that learners perceive to influence these relationships at pivotal moments of change during spoken L2 tasks. Significant methodological features of this study include (1) the dynamic measurement of breakdown fluency (pausing behavior) during short segments of speech, and (2) the combined measurement of both enjoyment and anxiety. The findings from this study suggest that anxiety and enjoyment are influenced by task design and implementation factors (which remain relatively fixed during the task) as well as cognitive-linguistic and achievement outcome factors (which arise as task communication unfolds). However, each learner—mediated by their own experiences—responded to these influences in unique ways. Overall, it was observed that surges in anxiety frequently coincided with breakdowns in fluency. In contrast, enjoyment—often manifesting as confidence when using planned ideas, a feeling of achievement from using appropriate language, and positive emotions related to past experiences—had a much weaker relationship with fluency.

Importantly, this research has pedagogical implications for L2 learning and teaching. If teachers want to optimize learners' fluency during tasks, they should seek ways to reduce anxiety and improve enjoyment during task performances. This research suggests three ways this can be achieved. First, providing learners with a clear task structure can help scaffold their performance. In the case of a monolog task, this might be done by describing to learners the specific stages they need to progress through in their talk (e.g., situation, problem, solution); alternatively, for more advanced learners, teachers might provide them with time to plan their own outline for the task. Second, connecting the task topic to learners' personal experiences can improve learners' emotional investment in the task. The teacher can do this by choosing topics that are in line with learners' interests, or learners might be encouraged to make their own connections to task topics as a preparation activity. Finally, teachers can cultivate

positive emotions by highlighting achievement outcomes. After the task is completed, teachers can conduct a post-task reflection activity (e.g., reporting on language they felt they used successfully); during the task, other learners (or the teacher) might be encouraged to use non-verbal gestures (e.g., a nod, thumbs up) to signal to the speaker successful use of language or content.

This study is not without its limitations. First, as this research contains a small sample size and relies on learners' imperfect ability to express complex thought-processes, results cannot be generalized. Although there is clear value in depicting individual participants' experiences, there is certainly room for future idiodynamic studies to expand the number of participants to the extent that they might provide generalized findings (for a rare large-scale idiodynamic study, see Ducker, 2021). Second, learners' anxiety and enjoyment ratings reflect their understanding of these terms, how they interpret their own feelings, and their competency in using the computer software. While learners were given clear definitions of the target emotions and each practiced using the rating software before the task, participants' ratings should still be interpreted with some caution. Despite these shortcomings, it is hoped that this study provides a more comprehensive picture of emotion and fluency in task performances.

Based on the findings of this study, some suggestions for future research can be made. Pre-task planning is a well-studied implementation variable in the TBLT domain, and considered an important, easy-to-implement condition that reduces learners' speech processing demands, resulting in more fluent task performances (Willis and Willis, 2012; Pang and Skehan, 2014; Lambert et al., 2021). However, the current study revealed that pre-task planning sometimes provokes surges in anxiety related to feelings of constraint or regret. Future research should employ the idiodynamic method to further explore the impact of different types of pre-task planning on emotions and fluency in task performances. This would add a more nuanced understanding to an established implementation practice. Research should also strive to explore a wider range of emotions in relation to learners' language production. Selection could be based on the multidimensional model of affect (i.e., dimensions of valence and activation) (Linnenbrink, 2007) or control-value theory (achievement outcome/activity emotions) (Pekrun, 2006). These frameworks may be helpful in predicting which emotions are worthwhile investigating in the context of task-based speaking performances.

Data availability statement

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

Ethics statement

The studies involving human participants were reviewed and approved by the Survey and Behavioral Research Ethics Faculty Sub-committee (Faculty of Education), The Chinese University of Hong Kong. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SA contributed to conception, design of the study, collected the data, performed the analysis, and wrote the manuscript.

Funding

This work was supported by the Chinese University of Hong Kong, Direct Grant 2020/21 (reference no. 4058089).

Acknowledgments

The author would like to thank the Chinese University of Hong Kong, Direct Grant, for providing the funding to support this project. The author would also like to thank the research assistants, Bella, Roanna, and Miang, for assisting in data collection and analysis.

Conflict of interest

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

Publisher's note

References

Ahmadian, M. J. (2015). "Working Memory, Online Planning and L2 Self-Repair Behaviour," in *Working Memory in Second Language Acquisition and Processing*, eds Z. Wen, M. B. Mota, and A. McNeill (Bristol: Multilingual Matters), 16–174.

Ahmadian, M. J., and Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Lang. Teach. Res.* 15, 23–49. doi: 10.1177/1362168810383329

Arnold, J., and Brown, H. (1999). "A map of the terrain," in *Affect in Language Learning*, ed. J. Arnold (Cambridge, CA: Cambridge University Press), 1–24.

Aubrey, S. (2017a). Inter-cultural contact and flow in a task-based Japanese EFL classroom. *Lang. Teach. Res.* 21, 717–734. doi: 10.1177/1362168816683563

Aubrey, S. (2017b). Measuring flow in the EFL classroom: learners' perceptions of inter- and intra-cultural task-based interactions. *TESOL Q.* 51, 661–692. doi: 10.1002/tesq.387

Aubrey, S. (2022). Dynamic engagement in second language computer-mediated collaborative writing tasks. Does communication mode matter? *Stud Second Lang. Learn Teach.* 12, 59–86. doi: 10.14746/ssllt.2022.12.1.4

Aubrey, S. (2021). Enhancing long-term learner engagement through project-based learning. *ELT J.* 1–11. doi: 10.1093/elt/ccab032

Aubrey, S., and Philpott, A. (2022). Use of the L1 and L2 in strategic planning and rehearsal for task performances in an online classroom. *Lang Teach Res.* 1–26. doi: 10.1177/13621688221077421

Aubrey, S., King, J., and Almukhaild, H. (2020). Language learner engagement during speaking tasks: a longitudinal study. *RELC J.* 1–15. doi: 10.1177/0033688220945418

Aubrey, S., Lambert, C., and Leeming, P. (2022). The impact of first as opposed to second language pre-task planning on the content of problem-solving task performance. *Lang. Teach. Res.* 26, 867–892. doi: 10.1177/1362168820917844

Boersma, P., and Weenink, D. (2019). *Praat: Doing phonetics by computer* (Version 6.0.56) *[Computer program]*. Available online at: http://www.praat.org (accessed May 15, 2022).

Bonett, D. G., and Wright, T. A. (2000). Sample size requirements for estimating pearson, kendall and spearman correlations. *Psychometrika* 65, 23–28. doi: 10.1007/BF02294183

Boudreau, C., MacIntyre, P., and Dewaele, J.-M. (2018). Enjoyment and anxiety in second language communication: an idiodynamic approach. *Stud Second Lang Learn Teach.* 8, 149–170. doi: 10.14746/ssllt.2018.8.1.7

Brantmeier, C. (2005). Nonlinguistic variables in advanced second language reading: learners' self-assessment and enjoyment. *Foreign Lang. Ann.* 38, 494–504. doi: 10.1111/j.1944-9720.2005.tb02516.x

Bui, G., and Teng, F. T. (2018). Exploring learners' self-reported behavioral patterns in two task-readiness conditions: a qualitative study. *Chinese J. Appl. Linguist.* 41, 129–148. doi: 10.1515/cjal-2018-0008

Bygate, M., and Samuda, V. (2005). "Integrative planning through the use of task repetition," in *Planning and task Performance in a Second Language*, ed. R. Ellis (Amsterdam: John Benjamins), 37–74. doi: 10.1075/lllt.11.05byg

Csikszentmihalyi, M. (2014). *Applications of Flow in Human Development and Education*. Berlin: Springer. doi: 10.1007/978-94-017-9094-9

Csikszentmihalyi, M., Abuhamdeh, S., and Nakamura, J. (2005). "Flow," in *Handbook of Competence and Motivation*, eds A. J. Elliot and C. S. Dweck (New York, NY: Guilford Press), 598–608.

Dao, P., and Sato, M. (2021). Exploring fluctuations in the relationship between learners' positive emotional engagement and their interactional behaviours. *Lang. Teach. Res.* 25, 972–994. doi: 10.1177/13621688211044238

De Jong, N. H., and Bosker, H. R. (2013). "Choosing a threshold for silent pauses to measure second language fluency," in *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)*, ed. R. Eklund (Stockholm), 17–20.

Dewaele, J. M., and MacIntyre, P. (2016). "Foreign language enjoyment and foreign language classroom anxiety. The right and left feet of FL learning?," in *Positive Psychology in SLA*, eds P. D. MacIntyre, T. Gregersen, and S. Mercer (Bristol: Multilingual Matters), 215–236. doi: 10.21832/9781783095360-010

Dewaele, J. M., and MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Stud. Second Lang Learn Teach.* 4, 237–274. doi: 10.14746/ssllt.2014.4.2.5

Dewaele, J.-M., and Pavelescu, L. M. (2021). The relationship between incommensurable emotions and willingness to communicate in English as a foreign language: a multiple case study. *Innov Lang. Learn. Teach.* 15, 66–80. doi: 10.1080/17501229.2019.1675667

Ducker, N. (2021). Protecting and enhancing WTC with idiodynamic peer-peer strategy sharing. *System* 103:102634. doi: 10.1016/j.system.2021.102634

Elahi Shirvan, M., and Talebzadeh, N. (2018a). Foreign language anxiety and enjoyment in an imagined community. *Eurasian J. Appl. Linguist.* 4, 109–133. doi: 10.32601/ejal.464043

Elahi Shirvan, M., and Talebzadeh, N. (2018b). Exploring the fluctuations of foreign language enjoyment in conversation: an idiodynamic perspective. *J. Intercult. Commun. Res.* 47, 21–37. doi: 10.1080/17475759.2017.1400458

Ellis, R., and Barkhuizen, G. (2005). Analysing learner language. Oxford: Oxford University Press.

Ellis, R., Skehan, P., Li, S., Shintani, N., and Lambert, C. (2020). *Task-based Language Teaching: Theory and Applications*. Cambridge, CA: Cambridge University Press. doi: 10.1017/9781108643689

Fredrickson, B. L. (2004). "The broaden-and-build theory of positive emotions," in *The Science of well-being*, eds F. A. Huppert, N. Baylis, and B. Keverne (Oxford: Oxford University Press), 217–238.

Gotz, S. (2013). *Fluency in Native and Nonnative English Speech*. Amsterdam: John Benjamins. doi: 10.1075/scl.53

Gregersen, T., and MacIntyre, P. D. (2014). *Capitalizing on Language Learners' Individuality: From Premise to Practice*. Bristol: Multilingual Matters. doi: 10.21832/9781783091218

Gregersen, T., MacIntyre, P. D., and Meza, M. D. (2014). The motion of emotion: idiodynamic case studies of learners' foreign language anxiety. *Mod. Lang. J.* 98, 574–588. doi: 10.1111/modl.12084

Hong Kong Examinations and Assessment Authority (2015). *IELTS*. Available online at: http://www.hkeaa.edu.hk/en/recognition/benchmarking/hkdse/ielts/ (accessed May 15, 2022).

Horwitz, E. K. (2010). Foreign and second language anxiety. *Lang. Teach.* 43, 154–167. doi: 10.1017/S026144480999036X

Izard, C. E. (2009). Emotion theory and research: highlights, unanswered questions, and emerging issues. *Ann. Rev. Psychol.* 60, 1–25. doi: 10.1146/annurev.psych.60.110707.163539

Izard, C. E. (2010). The many meanings/aspects of emotion: definitions, functions, activation, and regulation. *Emot. Rev.* 2, 363–370. doi: 10.1177/1754073910374661

Kim, Y. (2013). "Promoting attention to form through task repetition in a Korean EFL context," in *Second Language Interaction in Diverse Educational Settings*, eds K. McDonough and A. Mackey (Amsterdam: John Benjamins), 3–24. doi: 10.1075/lllt.34.04ch1

Kormos, J. (2006). *Speech Production and L2 Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Lambert, C., and Zhang, G. (2019). Engagement in the use of English and Chinese as foreign languages: the role of learner-generated content in instructional task design. *Mod. Lang. J.* 103, 391–411. doi: 10.1111/modl.12560

Lambert, C., Aubrey, S., and Leeming, P. (2021). Task preparation and second language speech production. *TESOL Q.* 55, 331–365. doi: 10.1002/tesq.598

Lambert, C., Aubrey, S., Sung, B., and Tan, S. (2022). "Learner-generated content, conative and affective variation, and memory for input-based task content during computer-mediated communication," in *Individual Differences in Task-based Language Learning and Teaching*, ed. S. Li (Amsterdam: John Benjamins).

Lambert, C., Philp, J., and Nakamura, S. (2017). Learner-generated content and engagement in L2 task performance. *Lang. Teach. Res.* 21, 665–680. doi: 10.1177/1362168816683559

Larsen-Freeman, D., and Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.

Leeming, P., Aubrey, S., and Lambert, C. (2020). Collaborative pre-task planning processes and second-language task performance. *RELC J.* 1–17. doi: 10.1177/1362168819890825

Levelt, W. (1989). *Speaking: From intention to Articulation*. Cambridge, CA: MIT Press.

Levelt, W. (1999). "Producing spoken language: a blueprint of the speaker," in *Neurocognition of Language*, eds C. Brown and P. Hagoort (Oxford: Oxford University Press), 83–122. doi: 10.1093/acprof:oso/9780198507932.003.0004

Linnenbrink, E. A. (2007). "The role of affect in student learning: a multi-dimensional approach to considering the interaction of affect, motivation, and engagement," in *Emotion in Education*, eds P. A. Schutz and R. Pekrun (Amsterdam: Elsevier), 107–124. doi: 10.1016/B978-012372545-5/50008-3

MacIntyre, P. D. (1999). "Language anxiety: a review of the research for language teachers," in *Affect in Foreign Language and Second Language Learning: A Practical Guide to Creating a Low-anxiety Classroom Atmosphere*, ed. D. J. Young (New York, NY: McGraw-Hill), 24–45.

MacIntyre, P. D. (2019). Anxiety, uncertainty management and communication accommodation in women's brief dyadic conversations with a stranger: an idiodynamic approach. *Sage Open* 9:215824401986148. doi: 10.1177/2158244019861482

MacIntyre, P. D., and Ducker, N. (2022). The idiodynamic method: a practical guide for researchers. *Res. Methods Appl. Linguist.* 1, 1–10. doi: 10.1016/j.rmal.2022.100007

Macintyre, P. D., and Gregersen, T. (2022). The idiodynamic method: willingness to communicate and anxiety processes interacting in real time. *IRAL* 60, 67–84. doi: 10.1515/iral-2021-0024

MacIntyre, P. D., and Legatto, J. J. (2011). A dynamic system approach to willingness to communicate: Developing an idiodynamic method to capture rapidly changing affect. *Appl. Linguist.* 32, 149–171. doi: 10.1093/applin/amq037

MacIntyre, P. D., and Mercer, S. (2014). Introducing positive psychology to SLA. *Stud. Second Lang. Teach. Learn.* 4, 153–172. doi: 10.14746/ssllt.2014.4.2.2

MacIntyre, P. D., and Serroul, A. (2015). "Motivation on a per-second timescale: examining approach-avoidance motivation during l2 task performance," in *Motivational Dynamics in Language Learning*, eds Z. Dornyei, P. D. MacIntyre, and A. Henry (Bristol: Multilingual Matters), 109–138. doi: 10.21832/9781783092574-013

MacIntyre, P. D., and Wang, L. (2021). Willingness to communicate in the L2 about meaningful photos: Application of the pyramid model of WTC. *Lang. Teach. Res.* 25, 878–898. doi: 10.1177/13621688211004645

MacIntyre, P., and Gregersen, T. (2012). Emotions that facilitate language learning: the positive-broadening power of the imagination. *Stud Second Lang. Learn Teach.* 2, 193–213. doi: 10.14746/ssllt.2012.2.2.4

Marcos-Llinàs, M., and Garau, M. J. (2009). Effects of language anxiety on three proficiency-level courses of Spanish as a foreign language. *Foreign Lang. Ann.* 42, 94–111. doi: 10.1111/j.1944-9720.2009.01010.x

Nakamura, J., and Csikszentmihalyi, M. (2002). "The concept of flow," in *Handbook of Positive Psychology*, eds C. R. Snyder and S. J. Lopez (Oxford: Oxford University Press), 89–105.

Nakamura, S., Phung, L., and Reinders, H. (2021). The effect of learner choice on L2 task engagement. *Stud. Second Lang. Acquisit.* 43, 428–441. doi: 10.1017/S027226312000042X

Nematizadeh, S., and Wood, D. (2019). Willingness to communicate and second language speech fluency: An investigation of affective and cognitive dynamics. *Canad. Mod. Lang. Rev.* 75, 197–215. doi: 10.3138/cmlr.2017-0146

Pang, F., and Skehan, P. (2014). "Self-reported planning behaviour and L2 performance in narrative retelling," in *Processing Perspectives on Task Performance*, ed. P. Skehan (Amsterdam: John Benjamins), 95–128. doi: 10.1075/tblt.5.04pan

Pavlenko, A. (2013). "The affective turn in SLA: From "affective factors" to "language desire" and "commodification of affect," in *The Affective Dimension in Second Language Acquisition*, eds D. Gabry ìs and J. Bielska (Bristol: Multilingual Matters), 3–28.

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* 18, 315–341. doi: 10.1007/s10648-006-9029-9

Prior, M. T. (2019). Elephants in the room: an "affective turn," or just feeling our way? *Mod. Lang. J.* 103, 516–527. doi: 10.1111/modl.12573

Qiu, X., and Lo, Y. (2017). Content familiarity, task repetition and Chinese EFL learners' engagement in second language use. *Lang. Teach. Res.* 21, 681–698. doi: 10.1177/1362168816684368

Saito, K., Ilkan, M., Magne, V., Tran, M. N., and Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Appl. Psycholinguist.* 39, 593–617. doi: 10.1017/S0142716417000571

Schumann, J. (1994). Where is cognition? Emotion and cognition in second language acquisition. *Stud. Second Lang. Acquisit.* 16, 231–242. doi: 10.1017/S0272263100012894

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. Milton Park: Routledge. doi: 10.4324/9780203851357

Shao K., Nicholson L. J., Kutuk G. and Lei F. (2020). Emotions and instructed language learning: proposing a second language emotions and positive psychology model. *Front. Psychol.* 11, 1–13. doi: 10.3389/fpsyg.2020.02142

Skehan, P. (2014). "The context for researching a processing perspective on task performance," in *Processing Perspectives on Task Performance*, ed. P. Skehan (Amsterdam: John Benjamins), 1–26. doi: 10.1075/tblt.5.01ske

Skehan, P. (2018). *Second Language Task-based Performance: Theory, Research, and Assessment*. Milton Park: Routledge. doi: 10.4324/9781315629766

Skehan, P., and Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Lang. Teach. Res.* 1, 185–211. doi: 10.1177/136216889700100302

Swain, M. (2013). The inseparability of cognition and emotion in second language learning. *Lang. Teach.* 46, 195–207. doi: 10.1017/S0261444811000486

Talebzadeh, N., Elahi Shirvan, M., and Khajavy, G. H. (2020). Dynamics and mechanisms of foreign language enjoyment contagion. *Innov. Lang. Learn. Teach.* 14, 399–420. doi: 10.1080/17501229.2019.1614184

Tavakoli, P., and Skehan, P. (2005). "Strategic planning, task structure, and performance testing," in *Planning and Task Performance in a Second Language*, ed. R. Ellis (Amsterdam: John Benjamins), 239–273. doi: 10.1075/lllt.11.15tav

Willis, D., and Willis, J. (2012). *Doing Task-based Teaching*. Oxford: Oxford University Press.

Wood, D. (2016). Willingness to communicate and second language speech fluency: an idiodynamic investigation. *System* 60, 11–28. doi: 10.1016/j.system.2016.05.003

Appendix

APPENDIX 1 Task instructions.

A problem related to education in Hong Kong is described below. For this task you must give a 3-min speech in English in response to the problem. In the speech, you should do the following four things in order:
(1) explain the problem,
(2) compare possible solutions to the problem,
(3) recommend one of these solutions, and
(4) give reasons for your recommendation.
Before giving your speech, you will have 10 min to prepare using a worksheet. After the preparation, the worksheet and problem description will be taken away.
Problem:
Having a multicultural university campus does not in itself ensure interaction between international and local students. In fact, at many so-called "multicultural universities" in Hong Kong, international students and local students do not mix and remain isolated in their own social groups. This is a problem because without international-local student interaction on campus, universities are not benefiting from an exchange of ideas, cultures and languages, which is very important to remain globally competitive.

APPENDIX 2 Planning worksheet.

Pre-Task worksheet To help you prepare for your 3-min speech, you have 10 min to complete this worksheet. Please write in English.
Part 1. Answer the questions below.
What is the cause of the problem? _____
Who or what is responsible for the problem? _____
How would you describe the problem for someone who has not read about it?
_____
Part 2. First, complete table by deciding on four possible solutions to the problem. Then, think about the advantages and disadvantages of each solution and note them in the column.
Possible solutions Advantages Disadvantages
1

2

3

Part 3. Which of these solutions do you think is the best?
Recommended solution: _____
Please decide on 3 reasons for your choice.
_____
_____
_____

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 08 September 2022
DOI 10.3389/fpsyg.2022.940269

OPEN ACCESS

EDITED BY
Xun Yan,
University of Illinois
at Urbana-Champaign, United States

REVIEWED BY
Phillip Hamrick,
Kent State University, United States
Peijian Paul Sun,
Zhejiang University, China

*CORRESPONDENCE
Mona Roxana Botezatu
botezatum@health.missouri.edu

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

RECEIVED 10 May 2022
ACCEPTED 17 August 2022
PUBLISHED 08 September 2022

CITATION
Botezatu MR, Weissheimer J, Ribeiro M, Guo T, Finger I and Mota NB (2022) Graph structure analysis of speech production among second language learners of Spanish and Chinese.
Front. Psychol. 13:940269.
doi: 10.3389/fpsyg.2022.940269

Graph structure analysis of speech production among second language learners of Spanish and Chinese

Mona Roxana Botezatu[1]*, Janaina Weissheimer[2], Marina Ribeiro[3], Taomei Guo[4], Ingrid Finger[5] and Natalia Bezerra Mota[3,6]

[1]Department of Speech, Language and Hearing Sciences, University of Missouri, Columbia, MO, United States, [2]Brain Institute, Federal University of Rio Grande do Norte, Natal, Brazil, [3]Research Department at Motrix Lab, Motrix, Rio de Janeiro, Brazil, [4]State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China, [5]Department of Modern Languages, Federal University of Rio Grande do Sul, Porto Alegre, Brazil, [6]Department of Psychiatry and Legal Medicine, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil

Language experience shapes the gradual maturation of speech production in both native (L1) and second (L2) languages. Structural aspects like the connectedness of spontaneous narratives reveal this maturation progress in L1 acquisition and, as it does not rely on semantics, it could also reveal structural pattern changes during L2 acquisition. The current study tested whether L2 lexical retrieval associated with vocabulary knowledge could impact the global connectedness of narratives during the initial stages of L2 acquisition. Specifically, the study evaluated the relationship between graph structure (long-range recurrence or connectedness) and L2 learners' oral production in the L2 and L1. Seventy-nine college-aged students who were native speakers of English and had received classroom instruction in either L2-Spanish or L2-Chinese participated in this study. Three tasks were used: semantic fluency, phonemic fluency and picture description. Measures were operationalized as the number of words per minute in the case of the semantic and phonemic fluency tasks. Graph analysis was carried out for the picture description task using the computational tool *SpeechGraphs* to calculate connectedness. Results revealed significant positive correlations between connectedness in the picture description task and measures of speech production (number of correct responses per minute) in the phonemic and semantic fluency tasks. These correlations were only significant for the participants' L2- Spanish and Chinese. Results indicate that producing low connectedness narratives in L2 may be a marker of the initial stages of L2 oral development. These findings are consistent with the pattern reported in the early stages of L1 literacy. Future studies should further explore the interactions between graph structure and second language production proficiency, including more advanced stages of L2 learning and considering the role of cognitive abilities in this process.

KEYWORDS

bilingual language production, second language proficiency, graph structure analysis, Spanish, Chinese, English

Introduction

Much of what is known about speech production comes from the study of single word and sentence production. The production of units of language above a single sentence (i.e., discourse) has received less attention in the literature, even though it represents one of the most complex forms of communication. Speakers engage in the production of spontaneous monologic speech for pragmatic purposes, such as describing a scene or event, giving instructions, telling a story, or arguing for a point of view. A critical part of everyday conversation, monologic speech is a complex task that presents distinct demands on speech planning and production, involving multiple stages of processing. The most pervasive model of speech production (Levelt's, 1999) "blueprint of the speaker," and theories of discourse production (e.g., Eggins and Martin, 1997; Halliday, 2004; Sherratt, 2007) agree that the stages of speech production include the selection of a topic/message, the retrieval of relevant information which is then shaped into a logical structure, the selection of the lexical items and grammatical features that map onto the message content, the specification of the phrase structure of each utterance, along with the retrieval of phonological representations of lexical items and the motor execution of the phonetic plan. Interactive effects among these processing stages have been documented, particularly in the speech production literature (for a review, see Goldrick, 2006), suggesting that the distinct stages of processing may influence one another. The current study aims to add to this literature by evaluating how lexico-semantic processes may influence structural aspects of discourse, such as the connectedness of speech produced in continuous sequence. Connected speech may be thought of as the "the rapid, smooth, accurate, lucid and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" (Lennon, 2000, p. 26). For the purpose of the current study, we define connected speech as the continuous sequence of spoken words that occurs in monologic discourse.

Production of connected speech in the second language

The production of connected speech is highly automatized in the native language (L1), yet remains open to the influence of age and education (Le Dorze and Bedard, 1998). In the second language (L2), the production of connected speech is not fully automatized (Kormos, 2006) as a consequence of limited L2 proficiency, best reflected in measures of lexical complexity (Lu, 2012; Kang, 2013; Révész et al., 2016), grammatical complexity (Hahne, 2001; Iwashita, 2006; Rossi et al., 2006; Gan, 2012; De Clercq and Housen, 2017) and phonological encoding (Wong et al., 2021). Critically, the efficiency of L2 lexical access, operationalized as L2 vocabulary knowledge (Hilton, 2008; Koizumi and In'nami, 2013; Uchihara and Saito, 2019) and retrieval speed (De Jong et al., 2013), plays a key role in determining the quality of L2 speech (Kormos, 2006; Liu, 2020). The efficient retrieval of L2 lexical items is dependent not only upon L2 vocabulary knowledge (Hilton, 2008), but also on the ability to resolve high levels of competition from the more dominant L1 (e.g., Meuter and Allport, 1999; Misra et al., 2012), which is co-activated and competes for selection (e.g., Costa et al., 2000; Hoshino and Kroll, 2008; Colomé and Miozzo, 2010). Additionally, speakers have to deal with a limited amount of cognitive resources to provide the system with the necessary energy to operate. This makes L2 speech production even more demanding in the case of beginner L2 speakers, since they have to allocate a great amount of cognitive resources to mobilize lexical, syntactic and phonemic searches while trying to meet the demands of real-time communication (Green, 1986; Green and Abutalebi, 2013). In this sense, we can think of L2 proficiency as a bottleneck that speakers need to reach in order to further be able to employ discourse strategies as a next step in communication.

Emergent findings from research on second language acquisition have revealed a positive relationship between L2 lexical access and various measures of L2 speech quality, such as fluency, accuracy and complexity (Liu, 2020). Yet it is unclear whether this relationship is specific to the weaker, non-dominant L2, or whether it is also encountered in the dominant L1. The current study tested the hypothesis that unlike the native language, where connected speech production is highly automatic, connected speech production in the weaker L2 is highly dependent upon L2 vocabulary knowledge, regardless of the structural distance between speakers' two languages (e.g., structurally similar languages: English and Spanish; structurally dissimilar languages: English and Chinese). To test this prediction, the current study employed graph structure analysis to investigate the relationship between discourse connectedness and classic measures of lexical diversity (i.e., semantic and phonemic fluency) in speakers' L1 and L2 in two groups of college-aged L2 learners: native speakers of English who received classroom instruction in either L2-Spanish or L2-Chinese.

Measures of speech connectedness

To measure discourse connectedness, we have employed graph structure analysis, a method originally created to characterize formal thought disorders in clinical populations (Mota et al., 2012, 2014), but also used with monolingual (Mota et al., 2016, 2018) and bilingual children and adults (Leandro, 2020; Lemke et al., 2021). Formal thought disorders are a set of symptoms identified based on the way a narrative is produced. In this sense, evaluating the spontaneous word trajectory in narrative production mirrors the mental processes involved in the planning and production of discourse. Inspired by the

10.3389/fpsyg.2022.940269

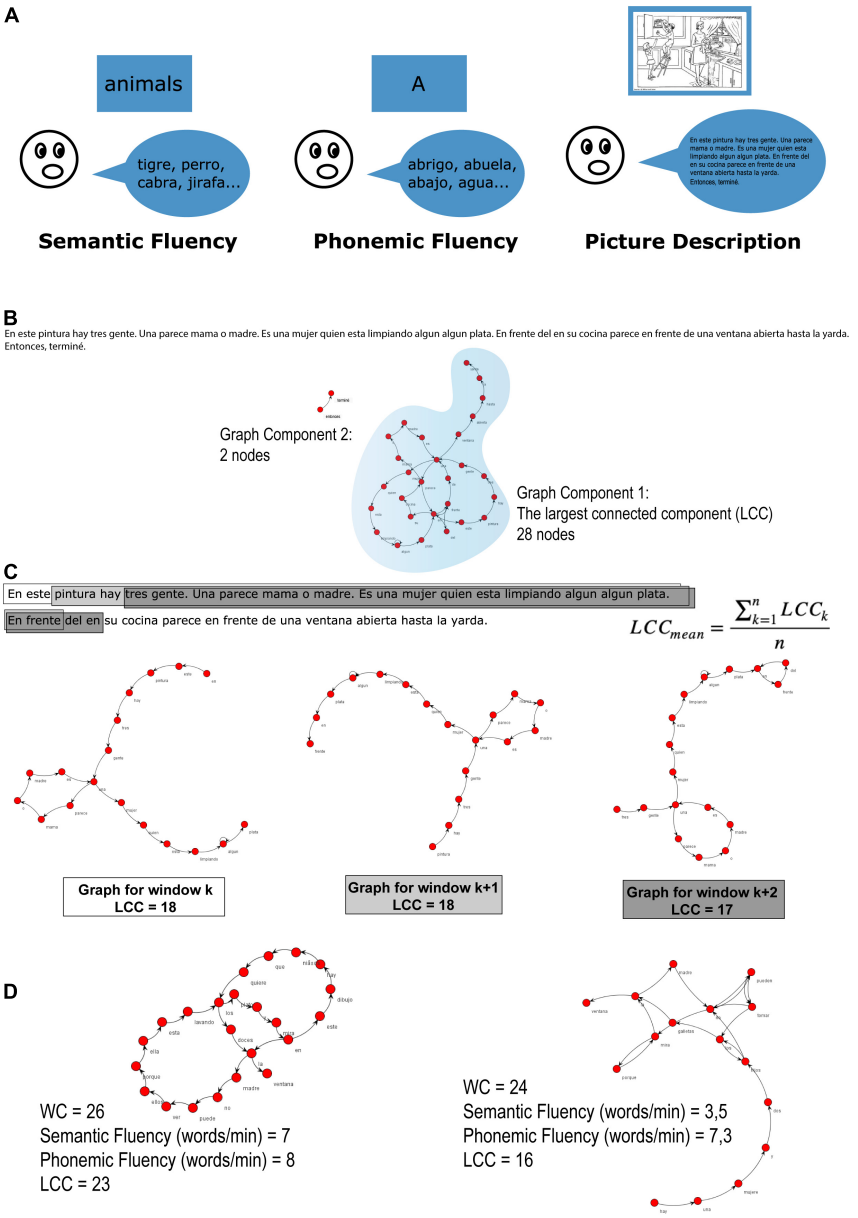


FIGURE 1
Verbal fluency tasks and graph analysis procedures. **(A)** Semantic fluency, phonemic fluency and picture description were operationalized as the number of words per minute. **(B)** An illustrative example of a graph from a text considering interruptions (here, when there is an interruption from the oral narrative, the following text after the interruption is transcribed in another line). If there are no repeated words, there will be two different components. The LCC counts the number of nodes inside the largest connected component (LCC, indicated by the blue shade). **(C)** To control for verbosity, narratives were analyzed using a moving window of a fixed word length (20 words) with a step of two words. LCC is averaged over the text windows. An example of a text divided into windows of 20 words, jumping two words to the following window. After computing all the 20-word graphs, the average of all the LCCs from all the windows was calculated (as shown in the equation).
(D) Representative examples of graphs of two bilingual subjects [English (L1) and Spanish (L2)], with different performances in fluency.

description of formal thought disorders, word graph analysis involves the study of word trajectory by means of representing each word as a node and the spontaneous sequence as directed edges (see **Figure 1**; Mota et al., 2012, 2014). Representing the narrative as a graph makes it possible to calculate topological aspects (e.g., connectedness) that characterize the word trajectory structure based on the recurrence pattern (Mota et al., 2014). The production of discourse involves a certain degree of word association and repetition. Word graph analysis distinguishes between more or less direct word associations by calculating short and long-range recurrences. Short-range recurrences refer to the repetitions of the same

word association (edges that link the same pair of nodes), while long-range recurrences represent the number of nodes inside a connected component (or a set of nodes with at least some connection between them) (Mota et al., 2018). Long-range recurrences provide a measure of global connectedness. Applying this method to characterize thought disorders, we found that the higher the connectedness, the lower the cognitive decline associated with mental illness, demonstrating that word graph connectedness may predict a diagnosis of schizophrenia (Mota et al., 2014, 2017; Palaniyappan et al., 2019; Morgan et al., 2021; Spencer et al., 2021), as well as the cognitive decline associated with dementia (Bertola et al., 2014; Malcorra et al., 2021). Moreover, studying the typical development of discourse patterns, we found that connectedness develops in association with general intelligence (IQ), theory of mind and verbal memory performance, predicting reading acquisition months in advance (Mota et al., 2016, 2020).

In the current study, long-range recurrences, measured by the number of nodes (or different words) inside the largest connected component (LCC), were used as a marker of speech connectedness during a spontaneous speaking task. Although the term "connectedness" is more commonly used in the field of mathematics, where it has emerged, we believe that the closest equivalent in psycholinguistics would be "textual cohesion." It is assumed that the adjacency between lexical items in a discursive fragment, represented and measured here using graph theory, may be an alternative way to obtain a quantitative measure of text unity; that is, of the relationship between the elements that make up its unity and determine its comprehension. As far as we know, in psycholinguistics, there have been few attempts to find linguistic markers of speech connectivity, one being the measure of syntactic complexity in terms of T-Units (Lemke et al., 2021).

Previous work employing these quantitative measures of speech connectedness has revealed that the production of long-range recurrences changes across lifespan and is associated with L2 proficiency. Mota et al. (2018) described the dynamics of short and long-range recurrence during typical development and their association with formal education, which reveals an interesting pattern of speech connectedness across lifespan. The authors showed that short-range recurrences (e.g., the repetitions of the same word associations) decreased during children's emerging literacy, but increased with advancing age. Conversely, the ability to produce long-range recurrences in a well-connected narrative increased over school years, and maturation is reached only during high school (Mota et al., 2018), but decreased in older adults in typical aging, as well as in dementia (Malcorra et al., 2021).

Speech connectedness in bilinguals

In the realm of bilingualism, Lemke et al. (2021) investigated the effects of bilingualism and biliteracy on connectedness and syntactic complexity in the written production of 11-year-old Portuguese-English bilingual children. The authors reported a correlation between graph attributes (i.e., connectedness) and the levels of syntactic complexity in both languages, demonstrating that, as children advance in the development of more complex writing strategies in Portuguese, they progress in their written production in English to the same extent. However, the study conducted by Lemke et al. (2021) did not include oral production tasks, only written ones. The current study addresses this methodological gap by investigating oral production through the analysis of graph attributes. Leandro (2020) was the first to extend this line of work to oral production in adult bilinguals and to show an association between measures of L2 oral proficiency and graph attributes in the case of Portuguese-English adult bilinguals. In his study, graph analysis (i.e., long-range connectedness and short-range repetitions) successfully predicted fluency in the continuum between pre-intermediate and near-native levels of L2 speech proficiency. In general, the more fluent speakers were, in terms of number of words per minute, the more connected their speech was found to be and the fewer short-range repetitions the participants produced. However, the author did not evaluate this relationship in the speakers' L1, Portuguese. Therefore, the present study fills this gap by looking at the interaction between verbal fluency and speech connectedness in both bilinguals' first and second languages. Additionally, the studies evaluating speech connectedness in bilinguals have solely focused on speakers of structurally similar languages, such as English and Portuguese. The present study extends the analysis to bilingual speakers of structurally similar languages (i.e., English and Spanish) and structurally dissimilar languages (i.e., English and Chinese) to provide a better representation of possible language pairings in emergent bilinguals and to evaluate whether the relationship between connected speech production and lexical retrieval changes as a function of the structural distance between bilinguals' two languages.

The current study

The current study tested the hypothesis that connected speech production in the weaker L2 is highly dependent upon L2 lexical retrieval (vocabulary knowledge), regardless of the structural distance between learners' two languages. Critically, we predicted that the same association would not be found in the L1 because lexical retrieval is highly automatic in the L1 in adulthood. Alternatively, if lexical retrieval is equally challenging in speakers' L1 and L2, then we should see an association between measures of lexical retrieval and discourse connectedness in both languages.

Materials and methods

Participants

A total of seventy-nine college-aged students who were native speakers of English and reported past or current enrollment in L2-Spanish (n = 54, mean age = 20.35, SD ± 2.47, 15 male, average age of initial L2-Spanish exposure = 11.06, SD ± 4.22) or L2-Chinese (n = 25, mean age = 21.68, SD ± 3.01, 9 male, average age of initial L2-Chinese exposure = 17.4, SD ± 3.04) courses were recruited from the University of Missouri and Beijing Normal University and completed the study for payment. Participants reported normal hearing, normal or corrected-to-normal vision and no history of neurological, language or learning deficits.

Materials

Speech production was assessed using a picture description task and lexical retrieval was measured using two distinct verbal fluency tasks (i.e., semantic fluency and phonemic fluency), which are described in detail below (see **Figure 1A**). All participants completed the tasks in English and in the foreign language in which they had received instruction, Spanish or Chinese, respectively. In addition to the discourse and lexical retrieval measures, participants also completed a language history questionnaire (Marian et al., 2007). All materials are presented below. Additional details on the materials and procedures can be found in Botezatu et al. (2022).

Picture description

The Cookie Theft scene from the Boston Diagnostic Aphasia Examination (Goodglass and Kaplan, 1983) was used in the picture description task. Participants were given 5 min to produce a narrative describing the picture and were instructed to speak for the entire time. Each trial began with a 1,000 ms blank screen, which was followed by a picture that cued participants to speak for 5 min and ended with a 1,000 ms blank screen. The resulting oral language samples were then transcribed offline by independent raters and scored in terms of average words-per-minute (96% interrater reliability), following the rules for counting words proposed by Nicholas and Brookshire (1993), providing a measure of discourse fluency.

Semantic fluency

A minute-long semantic category fluency task assessed retrieval of lexical items. Data samples were transcribed offline by independent raters (98% interrater reliability) and scored separately in participants' L1 and L2 in terms of the average number of correct responses (excluding simple and inflected repetitions) produced across four named semantic categories (i.e., *animals*, *clothing, fruit, furniture*).

Phonemic fluency

A minute-long letter fluency task was also used to measure retrieval of lexical items. Data samples were transcribed offline by independent raters (97% interrater reliability) and scored separately in participants' L1 and L2 in terms of the average number of correct responses (excluding repetitions and proper names) produced by participants across three named letters (*F*, *A*, or *S* in English; *P*, *M*, or *R* in Spanish; not assessed in Chinese due to no agreed-upon equivalent measure, but imputed using the Multivariate Imputation by Chained Equations R package (Van Buuren and Groothuis-Oudshoorn, 2010).

Language history background

Language experience was measured using the Language Experience and Proficiency Questionnaire (Marian et al., 2007). Participants self-rated their L1 and L2 proficiency, learning experience, frequency and context of exposure and use on a scale from 0 (*no proficiency, never*) to 10 (*native-like proficiency, always*). The questionnaire was administered at the end of the testing session, after participants completed all other tasks.

Data collection procedure

During one in-person testing session, participants completed the picture description, semantic fluency and phonemic fluency tasks in both the L1 and the L2, as well as a language history questionnaire administered in the L1. Participants were tested in the L1 first and L2 second to avoid L1-inhibition following performance in the weaker L2 (Misra et al., 2012). The experimental tasks were presented electronically using the E-Prime 2.0 software (Psychology Software Tools Incorporated, 2012). The Language History Questionnaire was administered electronically using Qualtrics (2019, Qualtrics, Provo, UT).

Data analysis

Proficiency analyses

To characterize the language proficiency and dominance of the participant sample, we compared L2 and L1 proficiency scores for both L2-Spanish and L2-Chinese groups using Wilcoxon Ranksum Tests. The Wilcoxon Ranksum Test is a non-parametric statistical analysis aiming to check the null hypothesis that two independent samples are equal.

Graph analyses

The oral narrative transcriptions from the picture description task, which included all the words spoken spontaneously by participants, were coded as a word-trajectory

graph using the *SpeechGraphs* software.[1] The software represents each word as a node and the sequence of words as directed edges (see Figure 1). This computational tool is used to map the spontaneous relationship between different words in a narrative. The method represents a narrative as a graph, allowing for topological characterization. It provides a number of useful measures (i.e., graph attributes), from elementary measures such as the total number of nodes and edges, to connectedness measures, such as the LCC. In the word graph trajectory, the LCC is defined as the largest set of nodes directly or indirectly linked by some path (see Figure 1). The number of nodes (i.e., different words) found in the LCC provides a measure of global connectedness that may be used to evaluate the lexical diversity of a narrative.

As there was no maximum limit for oral reports, to control for word count differences (i.e., verbosity), we analyzed graphs of 20 words, using a step of two words (corresponding to an overlap of 90% between consecutive graphs) to plot the next graph (see Figure 1C). We used a sliding window technique, in which we chose an initial set of 20 words, plotted a graph, moved two words to the next window and plotted the next graph with the following set of 20 words, and so on consecutively, until the complete set of 20 words in the text was graphed. This allowed us to screen the entire text in 20-word consecutive graphs. We then calculated the LCC of all 20 word-graphs and averaged all LCCs from the same reports. Representative examples of graphs of two bilingual subjects [English (L1) and Spanish (L2)], with different performances in fluency were represented in Figure 1D.

Statistical analyses

The analysis revealed that the data were not normally distributed (Shapiro-Wilk test). Therefore, Spearman non-parametric analyses were conducted to assess the association between the graph scores (LCC) originated in the Cookie Theft analysis and both semantic and phonemic verbal fluency measures. We corrected the significance level by using the Bonferroni test for 4 comparisons ($\alpha = 0.0125$). All the analyses were performed in Python 3.9.7 (Van Rossum and Drake, 1995).

1 The SpeechGraphs software was created by Mota et al. (2014), originally to be used with psychiatric populations, and is freely available at the following website (https://neuro.ufrn.br/softwares/speechgraphs). The software, which uses plain text as input and generates graphs and mathematical attributes as output, can be used in different platforms, such as Linux, Windows or OSX.

Results and discussion

Language proficiency

Participants varied on measures of L1 and L2 production (i.e., discourse fluency; semantic and phonemic fluency) and self-reported proficiency ratings on a 10-point scale (see Table 1). Self-reported proficiency ratings revealed that both L2-Spanish and L2-Chinese learners had a relatively low level of L2 proficiency, with a mean score of 3.9 (SD ± 2.62) in the case of L2-Spanish learners, and a mean score of 3.8 (SD ± 2.46) in the case of L2-Chinese learners. The difference between participants' L1 and L2 fluency scores was evaluated as an additional measure of proficiency. In L2-Spanish learners, Wilcoxon Ranksum Tests revealed a mean difference of W = –8.35, p = 7E-17 in the case of L2 Spanish–L1 English phonemic fluency; and of W = –8.93, p = 4E-19 in the case of L2 Spanish–L1 English semantic fluency. In the case of L2-Chinese learners, a mean difference of W = –5.85, p = 5E-09 in the case of L2 Chinese–L1 English phonemic fluency; and of W = –5.84, p = 5E-09 in the case of L2 Chinese–L1 English semantic fluency were found. Taken together, differences in self-reported proficiency ratings and fluency means between the two languages have led us to characterize the present sample as two groups of beginner L2 learners who maintained dominance of their native language, English.

Speech connectedness

Multiple regression analysis results in Figure 2 indicate that semantic and phonemic fluency predict speech connectedness only in the case of the L2. Although both semantic and phonemic fluency in L2-Spanish and L2-Chinese significantly contributed to explain speech connectedness in the picture description task (R^2 = 0.222, $p < 0.001$ for Spanish and R^2 = 0.293, $p < 0.005$ for Chinese), in the case of L1-English we see a different pattern, with phonemic fluency and semantic

TABLE 1 Mean (SD) psycholinguistic data.

Measure	Spanish learners		Chinese learners	
	English proficiency	Spanish proficiency	English proficiency	Chinese proficiency
Discourse fluency	157.43 (63.00)	57.33 (27.00)	121.17 (28.37)	77.19 (35.88)
Semantic fluency	18.95 (3.79)	6.21 (2.47)	18.45 (3.89)	7.39 (3.64)
Phonemic fluency	15.81 (3.66)	7.89 (2.48)	16.96 (3.63)	8.57 (3.05)
Average proficiency self-rating (/10)	10 (0)	3.9 (2.62)	10 (0)	3.8 (2.46)

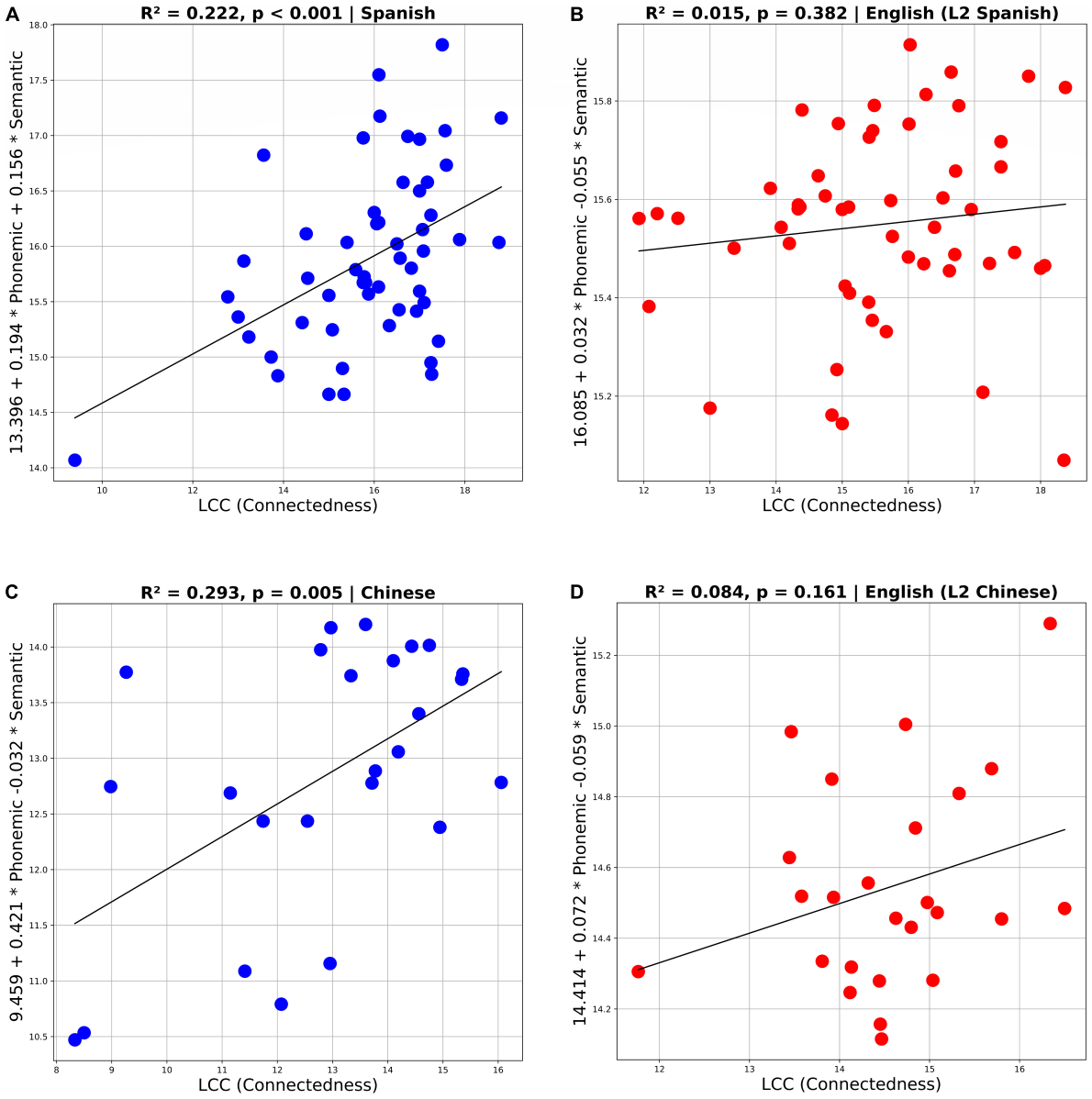


FIGURE 2
Multiple regression scatterplots showing the contribution of phonemic and semantic fluency to explain connectedness (LCC) in L2 Spanish and Chinese **(A,C)** and in L1 English **(B,D)**.

fluency not contributing to the prediction model (R^2 = 0.015, p = 0.382 for Spanish/English and R^2 = 0.084, p = 0.161 for Chinese/English). In other words, phonemic and semantic fluency explained 22% of connectedness variance in the spontaneous narratives in Spanish, and 29% of connectedness variance in the spontaneous narratives in Chinese. These results confirm our hypothesis that the speech production of beginner L2 learners is highly dependent on L2 lexical and phonemic retrieval and that connectedness is better explained by fluency in L2 than in L1, regardless of the structural distance between learners' two languages.

The regression analysis results also showed that phonemic fluency was more closely related to L2 connectedness than semantic fluency, especially for Chinese (Coefficient for phonemic fluency = 0.194 and coefficient for semantic fluency = 0.156, in Spanish; and Coefficient for phonemic fluency = 0.421 and coefficient for semantic fluency = 0.032 in Chinese). That led us to run Spearman correlations to evaluate more closely the relationship between phonemic fluency and connectedness in L2 and L1. Again, results revealed positive correlations between long-range recurrences (LCC), measured in the picture description task, and phonemic fluency

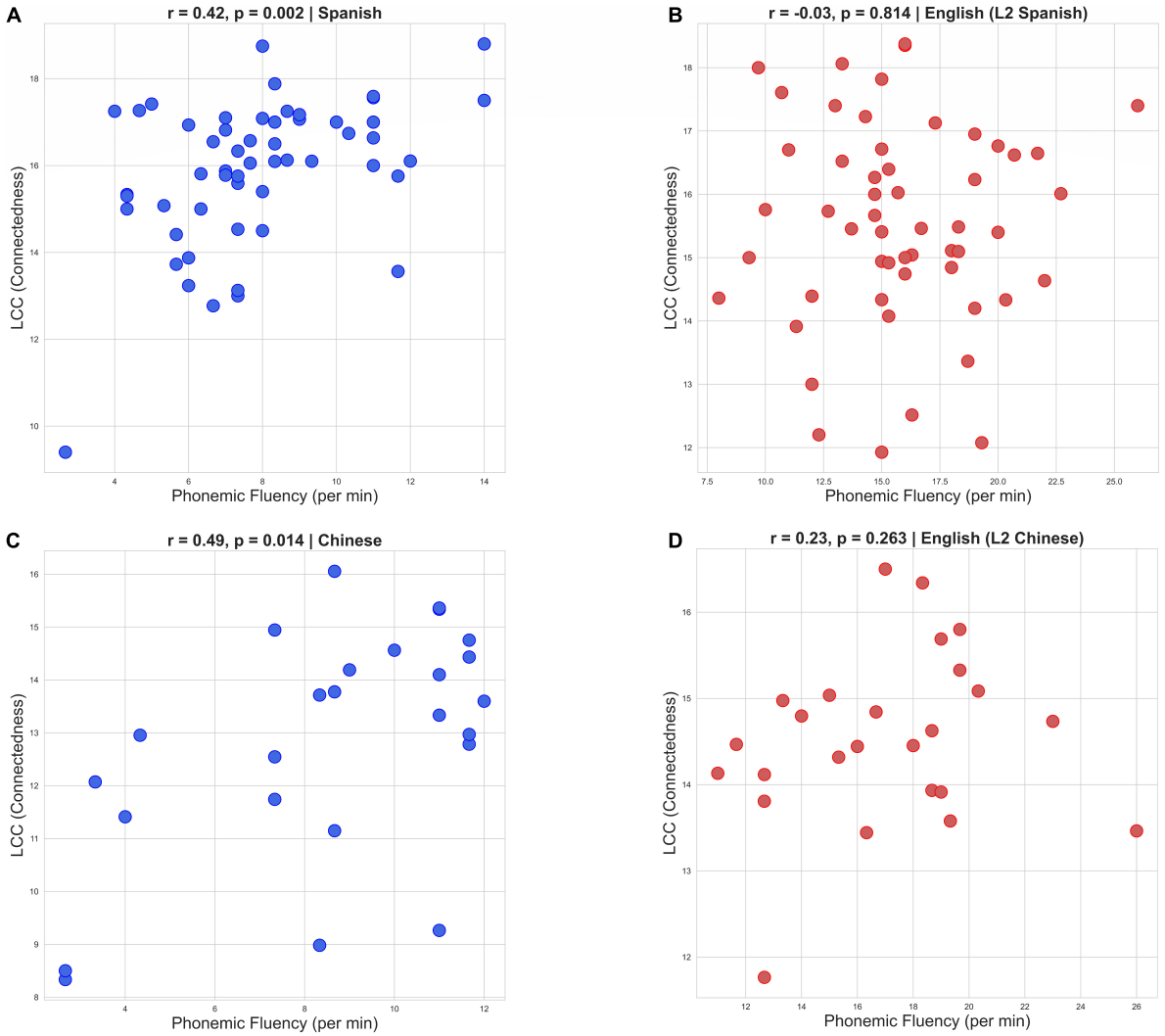


FIGURE 3
Correlation scatterplots of the Largest Connected Component (LCC) measure in English and L1 Phonemic Fluency **(B,D)** and the LCC measure in the L2 and L2 phonemic fluency (Spanish and Chinese; **A,C**).

(R = 0.42, p = 0.02 for Spanish and R = 0.49, p = 0.014 for Chinese). Once more, these correlations were only significant for the participants' L2–Spanish and Chinese (see Figure 3), reinforcing the claim we have put forward here that the speech production of beginner L2 learners is highly dependent on L2 lexical retrieval. The fact that L2 connectedness is better explained by phonetic fluency (rather than semantic fluency), regardless of learners' L2, seems to indicate that L2 learners in the current study relied more on phonetic cues to access lexical structures in order to meet the demands of the picture description task. This finding is consistent with previous reports of a progression from reliance on word form in beginner L2 learners to reliance on word meaning in more advanced L2 learners (e.g., Talamas et al., 1999). Additionally, the results of the current study demonstrate that the relationship between connected speech production in the L2 and L2 lexical retrieval in emergent bilinguals does not change as a function of the structural distance between bilinguals' two languages.

Taken together, our findings indicate that producing a lower number of long-range recurrences may be a marker of individual differences in the initial stages of L2 oral development, when the ability to produce a well-connected narrative tends to be dependent on a lexical repertoire, which is still under development, in order to incrementally aid connectedness in speech. These findings are consistent with the pattern reported in the early stages of L1 literacy, where the increase in longer recurrences has also been associated with the development of literacy (Mota et al., 2016, 2018). In other words, connectedness in an adult's L1 speech seems to

be well-structured and, therefore, less likely to be explained by variability in the individuals' L1 mental lexicon. The picture is different in the case of the developing L2, in which variability in individuals' ability to produce a narrative is linked to vocabulary size (e.g., Hilton, 2008; Koizumi and In'nami, 2013; Uchihara and Saito, 2019) and the speed at which learners can access their lexical repertoire (e.g., De Jong et al., 2013), therefore closely dependent on L2 proficiency (e.g., Kormos, 2006).

The developmental perspective adopted here reveals different strategies to produce a well-connected narrative in a new language. As we can see here, in the initial stages of second language acquisition, phonemic cues seem to play an important role in a naturalistic task such as narrating a scene as a monologue. At more advanced stages we could find different results, as we have presented evidence that the L1 narrative production is not associated with vocabulary retrieval. Also, differences in the bilingual experience or learning context may also reveal other strategies to be differently recruited.

Limitations and future directions

There are a number of limitations that we would like to acknowledge. First, the design of the current study, which tested participants in the L1 first and L2 second to avoid L1 inhibition following L2 retrieval, likely led to practice effects in the L2. These practice effects may have resulted in increased speech connectedness in the second language, but we cannot test this empirically based on the available data. Future studies should test the influence of practice on speech connectedness in the weaker second language. Second, we had access to a small sample of participants (particularly for the Chinese group), so the results should be replicated with larger samples. Third, we did not have access to participants with higher levels of L2 proficiency, which could reveal differences in the association between narrative production mechanisms and lexical retrieval. More studies with larger and more diverse samples in terms of proficiency levels are needed to advance our current understanding of the association between vocabulary acquisition and naturalistic use of a second language in the production of narratives. Future studies should further explore the interactions between graph structure and second language production proficiency, including more advanced stages of L2 learning and considering the role of cognitive abilities in this process. Associations between cognitive abilities (IQ, memory and theory of mind), academic achievement and speech connectedness have been documented in the past (Mota et al., 2016), revealing that children with higher cognitive and academic scores produced more long-range connections and fewer repetitions. Future research should test these associations in the L2.

Conclusion

Given that individual difference factors can reveal disparities in L2 speech production among learners, such factors have attracted researchers' growing interest. Here, we addressed individual differences in L2 speech production by employing graph structure analysis to evaluate the relationship between L2 lexical retrieval and the global connectedness of narratives during the initial stages of L2 acquisition and whether results can be replicated in the dominant L1. The current study contributes to the literature on second language acquisition by demonstrating that in the initial stages of L2 oral development, the connectedness of L2 speech is explained by variability in L2 lexical access. The study also demonstrates that a non-semantic graph strategy may be used to measure dynamics of narrative production in naturalistic settings, promoting the use of computational approaches to track L2 development, allowing for individualized feedback and helping to adjust speech trajectory over time. In addition, speech graphs may offer an alternative to refine the evaluation of L2 speech performance, with teachers and examiners being able to provide a faster and visually informative representation and assessment of learners' L2 speech production.

Data availability statement

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

Ethics statement

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

Author contributions

MB contributed to the data collection, data analysis, and manuscript writing. JW, MR, IF, and NM contributed to the data analysis and manuscript writing. TG contributed to the data collection. All authors contributed to the article and approved the submitted version.

Funding

This research was supported by a Catalyst Award and a Richard Wallace Faculty Incentive Grant from the University of

Missouri and an Advancing Academic-Research Careers Award from the American Speech-Language-Hearing Association to MB, by grant 31871097 from the National Natural Science Foundation of China to TG, as well as by grant 306659/2019-0 to JW and grant 312123/2019-1 to IF from the Brazilian National Council for Scientific and Technological Development (CNPq) and grant 88887.584264/2020-00 from the Coordination of Superior Level Staff Improvement (CAPES) to IF.

Acknowledgments

We thank Kathleen Acord, Madison Backes, Ashley Bramer, Jennifer Calvin, Sierra Cheung, Sierra Clemetson, Sarah D'Amico, Ryley Ewy, Laura Fry, Madison Hinmon, Jaclyn Johnson, Zeping Liu, Hanna Lowther, Sarah Marx, Carlos Martinez Villar, Allie Mitan, Xi Ren, Istvan Romhany, Morgan Trachsel, Jason Wong, No-Ya Yu, and Qiming Yuan for help with data collection and coding.

Conflict of interest

MR and NM are employed by Motrix, an EduTech startup. Also, NM has been a consultant for Boehringer Ingelheim.

The remaining 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

References

Bertola, L., Mota, N. B., Copelli, M., Rivero, T., Diniz, B. S., Romano-Silva, M. A., et al. (2014). Graph analysis of verbal fluency test discriminate between patients with Alzheimer's disease, mild cognitive impairment and normal elderly controls. *Front. Aging Neurosci.* 6:185. doi: 10.3389/fnagi.2014.00185

Botezatu, M. R., Guo, T., Kroll, J. F., Peterson, S., and Garcia, D. (2022). Sources of variation in second and native language speaking proficiency among college-aged second language learners. *Stud. Second Lang. Acquisit.* 44, 305–330. doi: 10.1017/S0272263121000188

Colomé, À, and Miozzo, M. (2010). Which words are activated during bilingual word production? *J. Exp. Psychol.* 36, 96–109. doi: 10.1037/a0017677

Costa, A., Caramazza, A., and Sebastian-Galles, N. (2000). The cognate facilitation effect: implications for models of lexical access. *J. Exp. Psychol.* 26, 1283–1296. doi: 10.1037/0278-7393.26.5.1283

De Clercq, B., and Housen, A. (2017). A cross-linguistic perspective on syntactic complexity in L2 development: syntactic elaboration and diversity. *Modern Lang. J.* 101, 315–334. doi: 10.1111/modl.12396

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., and Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Appl. Psycholinguist.* 34, 893–916. doi: 10.1017/S0142716412000069

Eggins, S., and Martin, J. (1997). "Genres and registers of discourse," in *Discourse as Structure and Process: Discourse Studies: A Multidisciplinary Introduction*, ed. T. V. Dijk (Los Angeles, CA: Edward Arnold), 230–256. doi: 10.4135/9781446221884.n9

Gan, Z. (2012). Complexity measures, task type, and analytic evaluations of speaking proficiency in a school-based assessment context. *Lang. Assess. Q.* 9, 133–151. doi: 10.1080/15434303.2010.516041

Goldrick, M. (2006). Limited interaction in speech production: chronometric, speech error, and neuropsychological evidence. *Lang. Cognit. Processes* 21, 817–855. doi: 10.1080/01690960600824112

Goodglass, H., and Kaplan, E. (1983). *The Assessment of Aphasia and Related Disorders*, 2nd Edn. Germany: Lea & Febiger.

Green, D. W. (1986). Control, activation, and resource: a framework and a model for the control of speech in bilinguals. *Brain Lang.* 27, 210–223. doi: 10.1016/0093-934X(86)90016-7

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

Hahne, A. (2001). What's different in second-language processing? Evidence from event-related brain potentials. *J. Psycholinguist. Res.* 30, 251–266. doi: 10.1023/A:1010490917575

Halliday, M. (2004). *An Introduction to Functional Grammar*. London: Hodder Education.

Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *Lang. Learn. J.* 36, 153–166. doi: 10.1080/09571730802389983

Hoshino, N., and Kroll, J. F. (2008). Cognate effects in picture naming: Does cross-language activation survive a change of script? *Cognition* 106, 501–511.

Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Lang. Assess. Q.* 3, 151–169. doi: 10.1207/s15434311laq0302_4

Kang, O. (2013). Linguistic analysis of speaking features distinguishing general English exams at CEFR levels. *Res. Notes* 52, 40–48.

Koizumi, R., and In'nami, Y. (2013). Vocabulary knowledge and speaking proficiency among second language learners from novice to intermediate levels. *J. Lang. Teach. Res.* 4, 900–913. doi: 10.4304/jltr.4.5.900-913

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. London: Routledge.

Le Dorze, G., and Bedard, C. (1998). Effects of age and education on the lexico-semantic content of connected speech in adults. *J. Commun. Disord.* 31, 53–71. doi: 10.1016/S0021-9924(97)00051-8

Leandro, D. C. (2020). *Pre-Task Planning, Working Memory Capacity and l2 Speech Production: An Exploratory Study Using Graph Analysis*. Ph.D. thesis, Federal University of Rio Grande do Norte, Brazil.

Lemke, C. E., Weissheimer, J., Mota, N. B., de Souza Brentano, L., and Finger, I. (2021). The effects of early biliteracy on thought organization and syntactic complexity in written production by 11-year-old children. *Lang. Teach. Res. Q.* 26, 1–17. doi: 10.32038/ltrq.2021.26.01

Lennon, P. (2000). "The lexical element in spoken second language fluency," in *Perspectives on Fluency*, ed. H. Riggenbach (Ann Arbor, MI: University of Michigan Press), 25–42.

Levelt's, W. J. (1999). "Producing spoken language: a blueprint of the speaker," in *The Neurocognition of Language*, eds C. M. Brown and P. Hagoort (Oxford: Oxford University Press), 83–122. doi: 10.1093/acprof:oso/9780198507932.003.0004

Liu, Y. (2020). Relating lexical access and second language speaking performance. *Languages* 5:13. doi: 10.3390/languages5020013

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Lang. J.* 96, 190–208. doi: 10.1111/j.1540-4781.2011.01232_1.x

Malcorra, B. L. C., Mota, N. B., Weissheimer, J., Schilling, L. P., Wilson, M. A., and Hübner, L. C. (2021). Low speech connectedness in Alzheimer's Disease is associated with poorer semantic memory performance. *J. Alzheimers Dis.* 82, 905–912. doi: 10.3233/JAD-210134

Marian, V., Blumenfeld, H. K., and Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): assessing language profiles in bilinguals and multilinguals. *J. Speech Lang. Hear. Res.* 50, 940–967. doi: 10.1044/1092-4388(2007/067)

Meuter, R. F., and Allport, A. (1999). Bilingual language switching in naming: asymmetrical costs of language selection. *J. Mem. Lang.* 40, 25–40. doi: 10.1006/jmla.1998.2602

Misra, M., Guo, T., Bobb, S. C., and Kroll, J. F. (2012). When bilinguals choose a single word to speak: electrophysiological evidence for inhibition of the native language. *J. Mem. Lang.* 67, 224–237. doi: 10.1016/j.jml.2012.05.001

Morgan, S. E., Diederen, K., Vértes, P. E., Ip, S. H., Wang, B., Thompson, B., et al. (2021). Natural Language Processing markers in first episode psychosis and people at clinical high-risk. *Transl. Psychiatry* 11, 1–9. doi: 10.1038/s41398-021-01722-y

Mota, N. B., Callipo, R., Leite, L., Torres, A. R., Weissheimer, J., Bunge, S. A., et al. (2020). Verbal short-term memory underlies typical development of "thought organization" measured as speech connectedness. *Mind Brain Educ.* 14, 51–60. doi: 10.1111/mbe.12208

Mota, N. B., Copelli, M., and Ribeiro, S. (2017). Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *NPJ Schizophr.* 3:18. doi: 10.1038/s41537-017-0019-3

Mota, N. B., Furtado, R., Maia, P. P., Copelli, M., and Ribeiro, S. (2014). Graph analysis of dream reports is especially informative about psychosis. *Sci. Rep.* 4:3691. doi: 10.1038/srep03691

Mota, N. B., Sigman, M., Cecchi, G., Copelli, M., and Ribeiro, S. (2018). The maturation of speech structure in psychosis is resistant to formal education. *NPJ Schizophr.* 4:25. doi: 10.1038/s41537-018-0067-3

Mota, N. B., Vasconcelos, N. A. P., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., et al. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS One* 7:e34928. doi: 10.1371/journal.pone.0034928

Mota, N. B., Weissheimer, J., Madruga, B., Adamy, N., Bunge, S. A., Copelli, M., et al. (2016). A naturalistic assessment of the organization of children's memories predicts cognitive functioning and reading ability. *Mind Brain Educ.* 10, 184–195. doi: 10.1111/mbe.12122

Nicholas, L. E., and Brookshire, R. H. (1993). A system for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *J. Speech Lang. Hear. Res.* 36, 338–350. doi: 10.1044/jshr.3602.338

Palaniyappan, L., Mota, N. B., Oowise, S., Balain, V., Copelli, M., Ribeiro, S., et al. (2019). Speech structure links the neural and socio-behavioural correlates of psychotic disorders. *Prog. Neuro Psychopharmacol. Biol. Psychiatry* 88, 112–120. doi: 10.1016/j.pnpbp.2018.07.007

Psychology Software Tools Incorporated (2012). *E-Prime*. Sharpsburg, PA: Psychology Software Tools

Révész, A., Ekiert, M., and Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Appl. Linguist.* 37, 828–848.

Rossi, S., Gugler, M. F., Friederici, A. D., and Hahne, A. (2006). The impact of proficiency on syntactic second-language processing of German and Italian: evidence from event-related potentials. *J. Cognit. Neurosci.* 18, 2030–2048. doi: 10.1162/jocn.2006.18.12.2030

Sherratt, S. (2007). Multi-level discourse analysis: a feasible approach. *Aphasiology* 21, 375–393. doi: 10.1080/02687030600911435

Spencer, T. J., Thompson, B., Oliver, D., Diederen, K., Demjaha, A., Weinstein, S., et al. (2021). Lower speech connectedness linked to incidence of psychosis in people at clinical high risk. *Schizophr. Res.* 228, 493–501. doi: 10.1016/j.schres.2020.09.002

Talamas, A., Kroll, J. F., and Dufour, R. (1999). From form to meaning: stages in the acquisition of second-language vocabulary. *Bilingualism* 2, 45–58. doi: 10.1017/S1366728999000140

Uchihara, T., and Saito, K. (2019). Exploring the relationship between productive vocabulary knowledge and second language oral ability. *Lang. Learn. J.* 47, 64–75. doi: 10.1080/09571736.2016.1191527

Van Buuren, S., and Groothuis-Oudshoorn, K. (2010). Mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45, 1–68.

Van Rossum, G., and Drake, F. L. Jr. (1995). *Python Reference Manual*. Amsterdam: Centrum voor Wiskunde en Informatica Amsterdam.

Wong, S. W., Dealey, J., Leung, V. W., and Mok, P. P. (2021). Production of English connected speech processes: an assessment of Cantonese ESL learners' difficulties obtaining native-like speech. *Lang. Learn. J.* 49, 581–596. doi: 10.1080/09571736.2019.1642372

frontiers | Frontiers in Communication

TYPE Brief Research Report
PUBLISHED 05 October 2022
DOI 10.3389/fcomm.2022.1012811

OPEN ACCESS

EDITED BY
Thomas L. Spalding,
University of Alberta, Canada

REVIEWED BY
Ali Sorayyaei Azar,
Management and Science University, Malaysia
Mona Roxana Botezatu,
University of Missouri, United States

*CORRESPONDENCE
Jimin Kahng
jkahng@olemiss.edu

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

RECEIVED 05 August 2022
ACCEPTED 20 September 2022
PUBLISHED 05 October 2022

CITATION
Kahng J (2022) Individual differences in adults' second language fluency development: Motivation and language use.
Front. Commun. 7:1012811.
doi: 10.3389/fcomm.2022.1012811

Individual differences in adults' second language fluency development: Motivation and language use

Jimin Kahng*

Department of Modern Languages, University of Mississippi, Oxford, MS, United States

Utterance fluency (UF) in a second language (L2) has been found to be associated with L2 proficiency. Nevertheless, the longitudinal development of UF has been underexamined and its relationship with individual differences such as motivation and L2 use has rarely been investigated simultaneously. The current study investigated L1-Chinese L2-English speakers' changes in UF (speed, pausing, and repair phenomena) before and after 5-month study-abroad and related UF changes to L2 use and motivation. The results showed that participants improved in mean syllable duration and end-clause silent pausing. Individuals' changes in certain UF measures, such as mean syllable duration and mid-clause pause frequency, were positively correlated with daily use of L2. Motivation measures largely did not exhibit significant correlations with UF changes, whereas ideal L2 self and intended effort/commitment demonstrated significant positive relationships with daily L2 use.

KEYWORDS

second language (L2) acquisition, utterance fluency, L2 use, L2 motivation, individual differences

Introduction

Speaking is a skill under time pressure and delivering one's message in a timely manner constitutes an essential part of having a conversation. Compared to their first language (L1), second language (L2) speakers often have less L2 knowledge, and are also considerably less fluent using the L2 knowledge they possess (Segalowitz, 2010). This highlights the importance of investigating L2 fluency. The current study explores the longitudinal development of adults' L2 fluency and its relationships with individual differences such as motivation and L2 use. In what follows previous studies on L2 utterance fluency and its development, and the role of motivation in L2 acquisition are discussed in turn.

L2 utterance fluency and its development

According to Segalowitz (2010), fluency has three distinct facets—utterance, cognitive, and perceived fluency. Utterance fluency (UF), the focus of this study, refers to the temporal and hesitation phenomena in speech and can be further categorized into speed, breakdown (pausing), and repair aspects (Skehan, 2003).

Previous studies have demonstrated significant differences between L1 and L2 speech in speed, the frequency of pauses and repairs, and pause distribution, where L2 speech has more pauses within a clause or utterance (e.g., Kahng, 2014; De Jong, 2016). UF measures such as articulation rate and mid-clause pause frequency have also exhibited moderate to strong correlations with L2 proficiency (e.g., Ginther et al., 2010; Kahng, 2014).

Although L2 UF has been widely researched, much fewer studies have tracked L2 learners' fluency development longitudinally. In their seminal study, which examined the role of learning context, language contact, and cognition in oral fluency development, Segalowitz and Freed (2004) found that the L1-English L2-Spanish learners improved speech rate and mean length of run without fillers after a semester of study-abroad; however, the amount of language contact could not explain fluency gains. Huensch and Tracy-Ventura (2017) investigated L2 UF development before, during, and after a 9-months residence-abroad and showed that reported gains in mean syllable duration appeared quickly and were maintained after return from study-abroad whereas gains in pause frequency appeared later and were sensitive to attrition after return home. Huensch et al. (2019) further explored the maintenance of L2 fluency 4 years after study-abroad and found that those who had intense L2 exposure after study-abroad maintained fluency gains made during study-abroad 4 years later but there was a lot of individual variation among those who had limited L2 exposure.

One more study worth discussing in line with L2 UF development is Saito et al. (2018). They found significant differences between low- vs. mid/high/native fluency in end-clause pause frequency, differences between low- vs. mid- vs. high/native fluency in mid-clause pause frequency, and differences between all groups for articulation rate. Although the findings stemmed from cross-sectional data, based on their distinctive length of residence (LOR) profile of the three fluency groups (CIs: 0.0–0.8, 3.7–7.1, and 8.8–12.4 years for low-, mid-, and high-fluency groups, respectively), they inferred that L2 fluency development could be observed in different aspects in the order of end-clause pausing, mid-clause pausing, and articulation rate.

Motivation and L2 acquisition

The role of motivation in L2 acquisition has been researched for several decades and the framing of motivation has evolved from a construct that is static, product-oriented into one that is more dynamic, situated, and process-oriented (Ushioda and Dörnyei, 2012). Throughout L2 motivation research, one of the most influential concepts has been integrativeness (Gardner, 1985), which refers to the desire to learn an L2 in order to come closer to the other language community. Integrativeness/integrative motivation has been widely researched through the 1970s and 1980s (e.g., Gardner and MacIntyre, 1993). However, its limitations have been recognized; the concept is not compatible with newly emerged cognitive motivational concepts such as goal theories or self-determination theory and it was often limiting and not applicable to many language learning environments, such as learning a foreign language as a school subject where the language is not spoken (Dörnyei, 2009).

Dörnyei (2005, 2009) proposed the "L2 Motivational Self System (L2MSS)" in order to overcome the limitations of integrativeness/integrative motivation and to broaden the scope of L2 motivation research. The L2MSS consists of the following three components. *Ideal L2 self* refers to the L2-specific aspect of one's ideal self. If our ideal self is one who speaks an L2, the ideal L2 self can motivate us to learn the L2 because we desire to reduce the discrepancy between our actual and ideal selves. *Ought-to L2 self* concerns "the attributes that one believes one ought to possess to meet expectations and to avoid possible negative outcomes" Dörnyei (2009, p. 29). *L2 learning experience* "concerns situated, "executive" motives related to the immediate learning environment and experience" Dörnyei (2009, p. 29). The L2MSS has been empirically supported by various groups of learners in different contexts. For instance, the empirical findings (Dörnyei, 2009) collected from China, Hungary, Iran, Japan, and Saudi Arabia, involving over 6,000 learners in four different learner types (i.e., secondary students, English-major and non-English-major university students, adult learners) supported the L2MSS and the ideal L2 self, in particular, was consistently highly correlated with the criterion measure (i.e., intended effort).

One final point to note in understanding the role of motivation in L2 acquisition is on what motivation has a direct impact. Traditionally the examined relationship was between motivation and L2 achievement. However, "motivation is a concept that explains *why* people behave as they do rather than how successful their behavior will be" (Csizér and Dörnyei, 2005, p. 20) and recently there has been the recognition that beyond L2 achievement we need to investigate what changes in L2 learners' *behavior* motivation can cause. And in a meta-analysis, Al-Hoorie (2018) did find that ideal L2 self exhibited stronger correlations with intended effort ($r = 0.61$) than with L2 achievement ($r = 0.17$). On the other hand, the role of motivation in UF fluency development has not yet been examined.

Current study

Taken together, although L2 UF has been extensively researched, its longitudinal development has been underexamined. In addition, the role of motivation and L2 use on its development has rarely been explored simultaneously. The current study aims to fill the gap in the literature and address the following research questions:

RQ1: Are there changes in L2 utterance fluency (speed, pausing, and repair phenomena) of L1-Chinese L2-English speakers before and after 5-month study-abroad?
RQ2: What are the relationships between motivation, L2 use, and changes in L2 utterance fluency?

Method

Participants

Forty-four Chinese learners of English participated in the project through an informed consent process and received $50 per session for their participation. This study focuses on the data of 31 learners (17 m/14 f) who participated in both sessions, before and after 5-month study-abroad, while taking undergraduate or graduate courses at a university in the US. Their mean age was 28 ($SD_{age} = 6$; $range_{age} = 21–46$) mean length of residence in the US was 2 months ($SD_{LOR} = 1$ month; $range_{LOR} < 6$ months) at the beginning of the study. They started to learn English around the age of 11 ($SD_{AO} = 2.0$). Based on the grammar and vocabulary sections of DIALANG, a diagnostic test developed by Lancaster University, they were mostly intermediate learners (3 A2; 26 Bs; 2 C1s), according to the Common European Framework of Reference (CEFR; Council of Europe., 2001).

Speaking tasks

Materials

Two types of questions were used as prompts (see Supplementary materials)—one on personal preference from a category such as important time or people (e.g., Who is your best friend? Describe this person and say why he/she is your best friend), and the other on personal choice between two options (e.g., Some people prefer to live in a small town. Others prefer to live in a big city. Which place would you prefer to live in? Use details and examples in your decision). For each type, six comparable prompts on daily life were developed to avoid practice effects of using the same prompts before and after study-abroad. In each session, one of six prompts from each type was randomly selected for each participant. Participants answered in total four different prompts across two sessions.

Procedure

In each session, participants answered the two questions described above. For each question, they had 15 s to prepare for their answer and were asked to talk for about a minute. Their speech was recorded using Praat (Boersma and Weenink, 2018), with a Blue Snowball USB microphone (frequency response 40 Hz−18 KHz) at a 44 KHz sampling rate (16-bit resolution; 1 channel).

Utterance fluency measures

All speech samples were transcribed and included information about silent and filled pauses, repetitions, corrections, and clause boundaries (Foster et al., 2000). Silent pauses (>250 ms; De Jong and Bosker, 2013) and filled pauses were identified and their length was measured in milliseconds (ms) using Praat (Boersma and Weenink, 2018). Pauses were further categorized into mid-clause or end-clause pauses based on the identified clause boundaries to examine their differential developmental patterns suggested by Saito et al. (2018). Following Skehan (2003), speed, breakdown, and repair fluency were measured. For speed fluency, mean syllable duration was calculated by dividing speech time excluding pause time by total number of syllables. For breakdown fluency, in addition to mean silent pause duration, the number of silent and filled pauses in the middle and at the end of clauses per 100 syllables were calculated. For repair fluency, the number of repetitions and corrections per 100 syllables were calculated.

Questionnaire on L2 motivation

A questionnaire was designed to measure participants' motivation and attitudes in L2 learning (see Supplementary materials). The questionnaire consisted of 29 Likert-scale items (on a scale of "1: strongly disagree" to "6: strongly agree") encompassing several attitudinal/motivational variables. The selected variables were those which have been shown to play an important role in determining L2 learning behaviors and effort, including integrativeness and the components of the L2MSS (Dörnyei, 2005). The items were adopted or adapted from Schmitt et al. (2004) and Dörnyei (2010). All the variables were comprised of multiple items. Table 1 describes the attitudinal/motivational variables measured in the study (Schmitt et al., 2004, p. 60; Dörnyei, 2005, p. 106) and reports the reliability measures—Cronbach's alpha in Time1 and Time2. The reliability of the motivation questionnaire was satisfactory.

Questionnaire on L2 use

In order to estimate participants' use of English, a questionnaire on L2 use was developed. The questionnaire (see Supplementary materials) included items on the hours of daily L2 listening, speaking, reading, and writing (on a scale of "less than 1 h", "about 1 h", "about 2 h", "about 3 h", or "more than 3 h"), and the percentages of time spent on L2 listening, speaking, reading, and writing, in comparison with the use of corresponding L1 language skills. It also had items on the number of close American friends, and the number of friends to speak in English with. In addition, an item on the frequency of having a long conversation (more than 10 min) in English (on

TABLE 1 The attitudinal/motivational variables measured in the current study.

Variables	Description	Number of items	α Time1	α Time2
Attitudes toward L2 learning	Subjective appraisal of the enjoyment of learning English	3	0.86	0.91
Ideal L2 self	L2-specific facet of one's ideal self	6	0.74	0.83
Ought-to L2 self	The attributes that one believes one ought to possess in order to avoid possible negative outcomes	6	0.74	0.76
Integrativeness	A broad positive disposition toward the L2 speaker community, including an interest in their life and culture	6	0.77	0.82
Language anxiety	Anxiety experienced while using English	2	0.67	0.76
Intended effort/commitment	The perceived importance of mastering a high level of English and the amount of effort the learner is willing to put into learning English	6	0.67	0.73

TABLE 2 Utterance fluency in Time1 and Time2.

	Time1		Time2		*F*	*df*	p^c	η^2
	M	*SD*	*M*	*SD*				
Mean syllable duration (ms)	310	48	290	42	8.183	1	0.035	0.214
Mean silent pause duration[a] (ms)	589	136	579	148	0.246	1	0.713	0.008
Number of[b]								
Mid-clause silent pauses[a]	8.25	4.53	8.32	4.34	0.055	1	0.816	0.002
End-clause silent pauses[a]	8.68	3.02	7.67	3.06	7.870	1	0.035	0.208
Mid-clause filled pauses[a]	3.49	2.83	3.75	2.76	0.597	1	0.668	0.020
End-clause filled pauses	3.30	2.20	2.52	1.79	6.916	1	0.035	0.187
Repetitions[a]	1.89	1.77	1.97	1.60	0.464	1	0.668	0.015
Corrections	1.21	0.96	1.37	0.95	1.429	1	0.482	0.045

[a]Log-transformed; [b]per 100 syllables; [c]p-values corrected using false discovery rate (FDR).

a scale of "never", "one to three times a week", "four to six times a week", "once a day", to "several times a day") was included as the measure was found to be useful in explaining the development of perceived comprehensibility and fluency (Derwing et al., 2008).

Analysis

The recordings of speaking tasks were transcribed, annotated, and measured by two native English-speaking research assistants. Once the recordings were annotated and measured by the first research assistant, their accuracy was checked by a second research assistant and corrections were made, when necessary, by the author.

In order to examine differences in the measures of fluency and motivation between Time1 and Time2, a series of repeated measures ANOVAs was performed and the *p*-values were corrected using false discovery rate (FDR). The variables that violated the assumptions of the repeated measures ANOVA (e.g., mean silent pause duration, number of silent pauses) were log-transformed. All the transformed data improved in terms of normality after the transformation. In examining the relationships between motivation, L2 use, and changes in UF, Pearson correlation was used for the variables that satisfied its assumptions (i.e., variables on motivation and fluency measures) and Spearman's rank order correlation was used when the analysis included ordinal variables (i.e., measures on L2 use).

Results

Changes in utterance fluency

Table 2 shows descriptive statistics and differences of UF measures in Time1 and Time2. The results of repeated-measures ANOVAs show that participants improved in mean syllable duration and the number of silent and filled pauses between clauses, whereas the rest of the measures demonstrated no significant changes before and after study-abroad.

Relationships between motivation, L2 use, and changes in L2 utterance fluency

Participants' responses to the motivation questionnaire in Time1 and Time2 were compared. Table 3 shows that

TABLE 3 Motivation and attitudes in L2 learning in Time1 and Time2 (scale: 1–6).

	Time1		Time2		*F*	*df*	p^a	η^2
	M	*SD*	*M*	*SD*				
Attitudes	4.60	0.73	4.70	0.87	0.58	1	0.542	0.021
Intended effort/commitment	4.93	0.50	4.80	0.69	1.09	1	0.456	0.039
Ideal L2 self	5.21	0.58	5.04	0.71	4.07	1	0.162	0.131
Ought-to L2 self	4.27	0.71	4.25	0.85	0.02	1	0.887	0.001
Integrativeness	4.68	0.58	4.35	0.78	12.60	1	0.006	0.318
Anxiety	3.29	1.10	3.48	1.06	1.35	1	0.456	0.047

[a]p-values corrected using false discovery rate (FDR).

TABLE 4 Participants' L2 use ($N = 30$).

	<1 h	About 1 h	About 2 h	About 3 h	More than 3 h
Daily L2 use on					
Listening	23%	17%	23%	17%	20%
Speaking	47%	37%	10%	3%	3%
Reading	13%	27%	13%	3%	43%
Writing	40%	27%	10%	13%	10%
	Never	**1–3 times a week**	**4–6 times a week**	**Once a day**	**Several times a day**
Frequency of having a long (more than 10 min) conversation in English	13%	53%	23%	3%	7%

their responses on motivation largely did not change between Time1 and Time2 except for integrativeness, which exhibited a significant decrease.

Table 4 presents participants' responses to the questionnaire on L2 use. The results show that the majority of participants spent 3 or more hours on reading, whereas spent no more than 1 h on speaking or writing. It is particularly noteworthy that half of them spent <1 h on speaking. On the frequency of having a long English conversation, only 10% reported to have it at least once a day, whereas 13% reported never to have it, and half of them reported to have it 1–3 times a week.

On the number of close American friends they have, 73% of the participants reported to have none and 13% reported to have one. Regarding the number of friends to speak in English with, 40% reported to have one to three, whereas 17% reported to have none. On the percentages of the time they use English in comparison with L1, they reported to use English 40 (95% CI = 31–50), 26 (95% CI = 17–35), 50 (95% CI = 38–62), 42% (95% CI = 29–55) of the time for listening, speaking, reading, and writing, respectively.

In order to examine the relationships between motivation, L2 use and changes in L2 UF measures, differences in UF measures were calculated by subtracting Time1 measures from Time2 measures. First, Pearson correlations were performed to examine the relationship between attitudinal/motivational variables and changes in UF measures (see Supplementary Table A1). There was a positive correlation between ought-to L2 self and changes in the number of repetitions, $r = 0.42$, $p = 0.022$, and an unexpected negative relationship between intended effort/commitment and the number of corrections, $r = -0.47$, $p = 0.01$.

Next, the relationship between L2 use and changes in UF measures was examined using Spearman correlations (see Supplementary Table A2). The results showed a few significant positive relationships between L2 use in Time2 and UF changes; between the number of friends to speak in English with and changes in mean syllable duration, $r = 0.42$, $p = 0.021$, between the percentage of English speaking and changes in the number of mid-clause silent pause, $r = 0.41$, $p = 0.024$, and between the percentage of English reading and changes in the number of end-clause silent pauses, $r = 0.38$, $p = 0.041$.

Lastly, the relationship between motivation (Time1 and Time2) and L2 use (Time2) was investigated using Spearman correlations (see Supplementary Table A3). Intended effort/commitment exhibited positive correlations in Time1 with daily hours of listening, $r = 0.38$, $p = 0.039$, and in Time2 with daily hours of reading, $r = 0.39$, $p = 0.038$, and those of writing, $r = 0.50$, $p = 0.006$. Ideal L2 self was also positively correlated with daily hours of listening in Time1, $r = 0.42$, $p = 0.022$, and in Time2, $r = 0.40$, $p = 0.03$. The rest of the motivation/attitude variables did not demonstrate significant relationships with any of the L2 use measures.

Discussion and conclusion

The current study investigated changes in L2 UF measures before and after 5-month study-abroad and their relationships with motivation and L2 use. The participants made significant gains in mean syllable duration, the number of end-clause silent and filled pauses. The findings are in line with the significant correlations found in previous studies between articulation rate (inverse of mean syllable duration) and L2 proficiency (e.g., Ginther et al., 2010; Kahng, 2014). The improvement in end-clause pausing is also compatible with Saito et al. (2018), in which development in end-clause pausing was proposed to develop before that in mid-clause pausing.

In terms of the relationships between changes in UF measures, motivation, and L2 use, attitudinal/motivational variables had few significant correlations with changes in UF measures, whereas ideal L2 self and intended effort/commitment exhibited significant positive correlations with L2 use, including daily hours of L2 listening, reading, and writing. The findings highlight the role of ideal L2 self in L2 use and also accord with the recent recognition that the power of motivation needs to be examined in terms of learners' behavior (e.g., Csizér and Dörnyei, 2005).

Changes in UF measures were found to have significant positive correlations with measures of L2 use. For instance, positive correlations were found between the number of friends to speak in English with and changes in mean syllable duration, and between the percentages of daily speaking in English and changes in the number of mid-clause silent pauses.

The current study is one of the first to demonstrate the complex associations between motivation and L2 use, and between various types of L2 use and different aspects of L2 UF development. Some of the novel findings are that, overall, motivation measures were not significantly correlated with UF development; however, they were positively correlated with daily L2 use. Measures of L2 use, in turn, were positively associated with adults' UF development. While this study has provided insights about the relationship between motivation, L2 use, and fluency development, there were some limitations, such as the small sample size and the relatively short-term (5 months) investigation. Future studies can overcome the limitations of the current study by tracking more participants' changes in L2 UF for a longer period, which will further enhance our understanding of L2 UF developmental patterns and its complex relationships with individual differences.

Data availability statement

The datasets presented in this article are not readily available because only the author and her collaborators have access to the dataset. Requests to access the datasets should be directed to JK, jkahng@olemiss.edu.

Ethics statement

The studies involving human participants were reviewed and approved by IRB at Northeastern Illinois University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JK has designed and conducted research and written the manuscript.

Conflict of interest

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

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.1012811/full#supplementary-material

References

Al-Hoorie, A. H. (2018). The L2 motivational self system: a meta-analysis. *Stud. Second Lang. Learn. Teach.* 8, 721–754. doi: 10.14746/ssllt.2018.8.4.2

Boersma, P., and Weenink, D. (2018). *Praat: Doing Phonetics by Computer [Computer Program]. Version 6.0.43*. Available online at: http://www.praat.org/ (accessed September 8, 2018).

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. New York, NY: Cambridge University Press.

Csizér, K., and Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *Mod. Lang. J.* 89, 19–36. doi: 10.1111/j.0026-7902.2005.00263.x

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: the effects of utterance boundaries and word frequency. *Int. Rev. Appl. Linguist. Lang. Teach.* 54, 113–132. doi: 10.1515/iral-2016-9993

De Jong, N. H., and Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. *Paper Presented at the 6th Workshop on Disfluency in Spontaneous Speech, Stockholm*.

Derwing, T. M., Munro, M. J., and Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Appl. Linguist.* 29, 359–380. doi: 10.1093/applin/amm041

Dörnyei, Z. (2005). *The Psychology of the Language Learner: Individual Differences in Second Language Acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Dörnyei, Z. (2009). *The Psychology of Second Language Acquisition*. Oxford: Oxford University Press.

Dörnyei, Z. (2010). *Questionnaires in Second Language Research: Construction, Administration, and Processing, 2nd Edn*. London: Routledge. doi: 10.4324/9780203864739

Foster, P., Tonkyn, A., and Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Appl. Linguist.* 21, 354–375. doi: 10.1093/applin/21.3.354

Gardner, R. C. (1985). *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. London: Edward Arnold.

Gardner, R. C., and MacIntyre, P. D. (1993). A student's contributions to second-language learning. Part II: affective variables. *Lang. Teach.* 26, 1–11.

Ginther, A., Dimova, S., and Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Lang. Test.* 27, 379–399. doi: 10.1177/0265532210364407

Huensch, A., and Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: a multidimensional investigation. *Mod. Lang. J.* 101, 275–293. doi: 10.1111/modl.12395

Huensch, A., Tracy-Ventura, N., Bridges, J., and Cuesta-Medina, J. (2019). Variables affecting the maintenance of L2 proficiency and fluency 4 years post-study abroad. *Study Abroad Res. Second Lang. Acquisit. Int. Educ.* 4, 96–125. doi: 10.1075/sar.17015.hue

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: temporal measures and stimulated recall. *Lang. Learn.* 64, 809–854. doi: 10.1111/lang.12084

Saito, K., Ilkan, M., Magne, V., Tran, M. N., and Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Appl. Psycholinguist.* 39, 593–617. doi: 10.1017/S0142716417000571

Schmitt, N., Dörnyei, Z., Adolphs, S., and Durow, V. (2004). "Knowledge and acquisition of formulaic sequences: a longitudinal study," in *The Acquisition and Use of Formulaic Sequences,* ed N. Schmitt (Amsterdam: John Benjamins), 55–86. doi: 10.1075/lllt.9

Segalowitz, N. (2010). *Cognitive Bases of Second Language Fluency*. New York, NY: Routledge. doi: 10.4324/9780203851357

Segalowitz, N., and Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: learning Spanish in at home and study abroad contexts. *Stud. Second Lang. Acquisit.* 26, 173–200. doi: 10.1017/S0272263104262027

Skehan, P. (2003). Task based instruction. *Lang. Teach.* 36, 1–14. doi: 10.1017/S026144480200188X

Ushioda, E., and Dörnyei, Z. (2012). "Motivation," in *The Routledge Handbook of Second Language Acquisition*, eds S. Gass, and A. Mackey (New York, NY: Routledge), 396–409.

frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 10 November 2022
DOI 10.3389/fcomm.2022.926409

OPEN ACCESS

EDITED BY
Boping Yuan,
University of Cambridge,
United Kingdom

REVIEWED BY
Alessia Cherici,
Indiana University Bloomington,
United States
Fengchao Zhen,
Shanghai Jiao Tong University, China

*CORRESPONDENCE
Jie Gao
gao_jie@fudan.edu.cn

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

RECEIVED 22 April 2022
ACCEPTED 24 October 2022
PUBLISHED 10 November 2022

CITATION
Gao J (2022) An investigation of high-proficiency L2 English speakers' oral test performance: A profiling approach. *Front. Commun.* 7:926409. doi: 10.3389/fcomm.2022.926409

An investigation of high-proficiency L2 English speakers' oral test performance: A profiling approach

Jie Gao*

College of Foreign Languages and Literature, Fudan University, Shanghai, China

Linguistic profiles, which are often established through the measurement of linguistic features, are able to demonstrate characteristics shared by a specific type of text or a group of language learners. This paper examines the contexts and purposes related to profiling research in language studies, meanwhile synthesizing quantitative profiling methods such as cluster analysis, Principal Component Analysis (PCA), and Factor Analysis (FA). A profiling study of high-proficiency L2 English speakers' test performance is also presented, which explains the profiling procedure in L2 speaking assessment. Cluster analysis conducted on speech fluency and vocabulary variables rendered four different speech profiles, which are associated with the speakers' L1 background and L2 English proficiency level. This paper also discusses the interpretation of linguistic profiles, as well as the statistical concerns involved in the profile construction process.

KEYWORDS

linguistic profiling analysis, cluster analysis, L2 speech production, quantitative research methods, language testing

Introduction

Individual-differences (ID) research, which was described by Ciszér and Dornyei (2005) as the study of language learners' stable and systematic deviations from "normal blueprints [...] typically aimed at identifying dimensions of enduring language learner characteristics relevant to the mastery of an L2 that are assumed to apply to everybody and on which people differ by degree" (p. 613). As a method capable of capturing language learners' individual differences in their performance, linguistic profiling has been more frequently applied in studies related to language learning. According to Halteren (2007), the concept of profiling focuses on linguistic features, the statistical calculation of which could assist researchers in looking for information underlying the text.

In this paper, the selection and quantification of linguistic features not only set descriptive parameters for a specific variety of text, but also represent the language practices of a particular learner group. The linguistic features involved, which often undergo procedures such as frequency counts, normalization, and statistical calculations that are inferential, present profiles demonstrating individual patterns with numerical specifications. Furthermore, a comparison among all the linguistic profiles

generated allows researchers to access possible variations across multiple text types and language users. Sufficient methodological support and unbiased interpretation framework are thus pivotal to explaining the linguistic profiles in language studies, which are expected to facilitate researchers' comprehension of language learners' individual characteristics. Through the investigation of literature published on the discussion of linguistic profiles, this paper examines the construction of linguistic profiles from two aspects:

a) What are the research contexts for conducting linguistic profiling research?
b) What are the commonly used statistical methods for constructing linguistic profiles quantitatively?

Answering these two questions will provide practical suggestions for researchers aiming to use profiling as a research method. In addition, this paper features a study that extracts linguistic profiles through L2 speech data collected from an oral English test. Linguistic profiles, which were built through cluster analysis, consist of variables measuring L2 English speakers' speech fluency and their use of vocabulary. This empirical study exemplifies procedural steps of creating profiles through multivariate statistical method, and emphasizes the necessity of reducing dimensions after obtaining the measurement results of multiple linguistic features. The author also provides interpretation for the profiles generated from the study, while discussing possible statistical concerns and pedagogical implications.

The construction of linguistic profiles in language studies

Purposes and themes of linguistic profiling research

Linguistic profiling, which is often embedded as a preliminary phase in the research process, provides supporting evidence for more overarching research questions regarding individual differences in second/foreign language studies. In general, the research purposes for linguistic profiling analysis include: (a) identifying the language proficiency level of L2 learners (Pienemann, 1998; Pienemann and, 2005; Pienemann and Keβler, 2011) and the strategies used by individual learners (Graham et al., 2020), (b) exploring for linguistic profiles of texts through corpus data (Friginal and Weigle, 2014), and (c) offering description for a country where multilingualism is practiced (Esseili, 2017; Banat, 2021) or sociolinguistic profiles of a particular group of second/foreign language learners (Alarcón, 2010). This literature review section further explains the targets and methods of building linguistic profiles, whose interpretation varies across diverse research contexts.

When associated with the learning process of adult L2 speakers, linguistic profiles have been built to describe the linguistic systems of learners at a specific stage (Clahsen, 1985; Brindley, 1998; Bartening, 2000; Ågren et al., 2012). More specifically, researchers created linguistic profiles of learners at different proficiency levels, which were then used as benchmarks in comparison to individual speech samples (Keβler and Liebner, 2011; Grandfeldt and Ågren, 2014). L2 learners would display a variety of morphosyntactic and grammatical patterns at different L2 proficiency levels, which suggest a developmental progression. The linguistic profiles individualizing L2 learning stages have been applied in the research area of Computer Assisted Language Learning (CALL) and Computer Assisted Language Testing (CALT), such as the development of language learning platforms and automatic scoring systems.

From a pedagogical perspective, linguistic profiling has enabled language teachers and instructors to document the progress of L2 speakers. For example, Van Compernolle (2014) recorded L2 learners' development of sociolinguistic knowledge in strategic interaction scenarios, where an individual learner's profile was built through detailed discourse analysis. The profiling process, which is realized by dynamic assessment and a series of pre-designed pedagogical modules, embodies the learner's growing control of verbal negation in French and provides insights into language course design.

Linguistic profiles have also been constructed using a large amount of observational data drawn from corpora. In Russian, for instance, Kuznetsova (2015) recognized gender-related profiles through "verbs that have a prevalence of masculine vs. feminine past tense endings" and then examined "the gender stereotypes that affect the activities denoted by the verbs" (p. 262). Corpus-based profiling results are also grounded in the correlation between semantic and distributional properties, or connections among distribution, form, and meaning (Divjak and Gries, 2006; Gries, 2010). As a result, profiles generated from the correlation between form and meaning have helped researchers predict meaning through the distribution of forms. In addition, corpus-based studies have provided large volumes of descriptive data for linguistic features in a specific category, the frequency of which has led to the establishment of more nuanced profiles. Hoffmann (2012), for instance, reported the cohesive profiles of spoken dialogues and written monologs, which further explained the linguistic difference across genres of texts.

From a sociolinguistic perspective, linguistic profiles of speaker groups have been produced through questionnaire data, where components for profile construction are drawn from both social factors and linguistic features. Research efforts have been dedicated to extracting a detailed profile

of a particular type of speakers from survey responses. For example, Fabricius (2006) examined Danish listeners' attitudes to British Received Pronunciation (RP) through open-ended questions and Likert scales. After listening to pre-recorded speech with RP, listeners responded to questions regarding the speaker's occupation, place of origin, socio-economic background, and personality. All these elements constitute a multi-faceted sociolinguistic profile for the speaker. Wang et al. (2021) designed the Linguistic Multicompetence Questionnaire to weigh in social factors such as migrant status, language maintenance beliefs, and cultural identity, which also assisted speech pathologists with comprehending different representations of multicompetence. Research on language learning and speech pathologies is enriched by the linguistic profiling of multilingual speakers, as multilingualism and speech disorders were treated as different concepts. To examine the linguistic profiles of heritage language speakers more closely, the survey instrument in Prentza and Kaltsa (2020) inquired speakers' exposure to specific languages, their attitudes toward different languages, and language using both at home and school.

Profiling information accessed through sociolinguistic documentation has diversified researchers' understanding of individuality in the age of new media, as both transcribed speech production and text information are considered as convincing evidence of language users' personal characteristics. Along with detecting the possible connections between social factors and individual differences among language users, Author Profiling (AP) on social media has also become a prominent strand of sociolinguistic profiling, which has integrated Natural Language Processing techniques into producing profiles of news media writers. Linguistic features such as text string length and word frequency were used to construct computational models and build text classifiers (Peng et al., 2016; Manna et al., 2019; Kowsari et al., 2020). Multimodal texts including emails, microblogs, movie reviews, and online bulletin boards provided input data for text classifiers, which are capable of predicting text authors' age, gender, and first language background. Author Profiling (AP) exemplifies the contribution of language engineering through text mining techniques and deep learning architecture. With input from corpus data, AP-related research outcome has been referenced for author attribution (Delmondes Neto and Paraboni, 2022; Deutsch and Paraboni, 2022), plagiarism detection (Potthast et al., 2014), and the identification of cyber troll accounts (Lundberg and Laitinen, 2020).

Methodological support, or a clarification of the commonly-used methods, is the prerequisite for the establishment of linguistic profiles. The next section of this paper will synthesize the quantitative procedures for linguistic profile extraction, which depends largely on the core statistics retrieved by the researchers.

Quantitative methods for linguistic profiling

From a feature-based perspective, linguistic profiling could be accomplished through statistical methods that function in either a descriptive or predictive manner. In comparison to descriptive methods, which generate profiles through the measurement of linguistic features, predictive methods involve inferential models that quantify the contribution of each variable to the models' prediction accuracy. This section of paper discusses the application of both descriptive approaches and predictive models for rendering linguistic profiles, both of which function to identify the individuality of language learners, types and genres of texts, or language varieties.

Descriptive approaches of building linguistic profiles

The comparison among language speakers' use of linguistic features is also conducted through non-parametric tests, as the prerequisite of normal distribution is not always fulfilled by linguistic data. In order to profile non-native speakers' nautical communication practices on board ships, John et al. (2017) compared non-native English speakers' speech patterns during bridge team communication (a subgenre of Maritime English) and non-nautical communication. Nonparametric statistics resulted from the Mann–Whitney *U*-tests, the Kruskal–Wallis test, and the calculation of effect sizes were presented in the study, which helped pinpoint a linguistic profile of non-native speakers' use of Maritime English in concrete settings of English for Specific Purposes (ESP).

In addition to non-parametric tests, profiles have been presented through statistical methods such as *Z*-score and chi-square test. As a statistic that measures the distance from the mean, *Z* score has been used to represent intergroup variations of language learners' ability (Potocki et al., 2017). Students' performance on a two-part reading comprehension test were transformed to *Z*-score, representing their decoding and comprehending abilities, respectively. Different combinations of the two-part *Z* score thus yielded multiple learner profiles, which calls for more accommodating pedagogical guidance to address the needs of different learners. The chi-square statistic, which is calculated from a contingency table, also examines the association between categorical variables as a non-parametric statistic. When used in profiling research that quantifies linguistic features through frequency numbers, the rows and columns of the contingency table could be interpreted as numerical representations of a multidimensional linguistic space. Through calculating the frequencies of linguistic variables within each language variety, Delaere et al. (2012) computed profile-based chi-square to measure the distance across multiple translated text types, thus offering explanations for the "standardness" issue in translational studies.

Compared with *Z*-score method, which is conveniently used on a limited number of variables, cluster analysis is a multivariate statistical procedure that places cases with similar numerical patterns into the same group. Labeled as "a statistical procedure that is relatively rarely used" (Ciszér and Dornyei, 2005, p. 613), cluster analysis is currently applied in language learning research that covers a broader scope of topics. Ryslewicz (2008) conducted cluster analysis on the assessment results of L2 learners' aptitude, intelligence, and proficiency level. Cognitive profiles emerged for successful and unsuccessful L2 English learners, which highlighted the contribution of inductive language learning abilities and expert use of first language to ESL students' learning achievement. When analyzing highly-rated English compositions, both Jarvis et al. (2003) and Friginal et al. (2014) implemented cluster analysis on features such as text length, conjuncts, hedges, and nominalization. The quality of students' writing depends on a balanced use of all the feature options, and essays written by native speaker of English have demonstrated a wider variety of styles.

It could be observed from the studies cited above that *Z*-score method and cluster analysis directly present profiles through the measurement results of features and variables. Principal Component Analysis (PCA) and Factor Analysis (FA), however, function to detect possible factors from sets of variables prior to profile construction. Principal Component Analysis (PCA) combines highly-correlated variables into a new component, which is consecutively used in description of the generated profiles. For example, Zheng et al. (2019) conducted Principal Component Analysis (PCA) on college students' responses to questionnaire statements inquiring their attitudes, beliefs, and experiences of learning a third language. The statements were grouped into generalized factors that compose various motivational profiles, which helped disentangle the relationship between motivation and instrumentality. Factor Analysis (FA), also interpreted as Multi-dimensional Analysis (MDA) in corpus linguistics, is also used as an efficient tool for profiling analysis. Friginal and Weigle (2014) computed the rate of occurrence for multiple linguistic features in L2 academic essays, and used Factor Analysis to identify four functional dimensions (e.g., Personal Opinion vs. Interpersonal Evaluation/Assessment). These dimensions have been enumerated as functional profiles that embody text parameters of L2 writing.

Predictive models of building linguistic profiles

Regression models, which are used to recognize statistically significant predictors for a measurable variable, function to further examine the features included for profile construction. Based on the rating results of speech fluency, Saito et al. (2018) categorized L2 speakers as learner profiles of low, medium, and high fluency through cluster analysis. Multiple regression analysis was then conducted to identify acoustic variables that contribute the most to speech fluency ratings. Instead of directly yielding profiling result, regression models provide a more granular view of objective features that constitute linguistic profiles.

In addition, profiling techniques also include building predictive models and developing computational methods for AP, which is often related to Author Attribution (AA). In this research context, digital texts retrieved from the Internet and social network platforms were used to identify their authors. Custódio and Paraboni (2021) reviewed the influence of text representation (e.g., online chats, blogs, reviews), choice of linguistic features (e.g., part of speech n-grams, character n-grams), and a variety of computational methods (e.g., Naïve Bayes, Random Forest, Support Vector Machine) on author identification accuracy. The discussion is continued in Deutsch and Paraboni (2022), where the task of gaining more knowledge of digital text authors is accomplished through building text classifiers.

Explanations have also been provided for the statistic models applied to identify authors of digital texts, particularly in cases where both linguistic features and demographic features were tapped by researchers. Moreno-Sandoval et al. (2021) described the functioning of a Multinominal Logit Model in quantifying the contribution of linguistic features in Twitter posts to understanding the celebrity's demographic background in reality (e.g., gender, fame, and occupation). The application of predictive models, however, cannot be separated from the discussion of dimension reduction, which is an important statistical concern in profiling analysis.

Well-established linguistic profiles are expected to capture the essential characteristics of a group of cases and require the inclusion of sufficient linguistic features. It is not unreasonable to hypothesize that inadequate features will result in the failure of extracting representative linguistic profiles, as all cases might appear to be homogeneous. An overflow of linguistic features, however, may cause problems such as collinearity. Dimension reduction is thus a necessary step to consider before conducting more complicated statistical investigation. In the study of Moreno-Sandoval et al. (2021), both Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) were applied as dimension reduction procedures for interval data and categorical data, which have demonstrated to be efficient in combining highly-correlated features or variables.

The implementation of statistical methods is often embedded in the fundamental steps for profile extraction. It is necessary for researchers to recognize the major dimensions that are essential to build up linguistic profiles. The second step is to pin down features and indices to represent these major dimensions, followed by selecting appropriate techniques to classify data into different categories. These categories, which contain cases sharing similarities in numerical values of all the features, will be rendered as profiles exhibiting individual characteristics.

Researching the individual differences among language learners, however, is sometimes challenged by issues such as a limited number of data cases, learners located in a restricted range of language proficiency levels, or the difficulty in choosing representative linguistic features for profile construction. This paper presents a study conducted by the author, which illustrates the profiling process through a combinational use of dimension reduction technique and multivariate statistical method. Both PCA and cluster analysis were adopted as profiling techniques, which were used to process dataset with possible concerns of collinearity and unequal sample sizes. In this study, speech data collected from an oral English test demonstrated the linguistic profiles of high-proficiency L2 English speakers, which were built upon fluency and vocabulary features. The profiling outcome provides opportunities for exploring interactions among L2 speakers' performances, their L2 proficiency level, and their L1 background. The cognitive activities L2 speakers are experiencing during speech production, which await to be explored in future studies, may offer explanations for the variances among different profiles. In addition to unpacking individual characteristics displayed in high-proficiency L2 English speakers' test responses, the study also holds a discussion of using holistic scales in speaking test, where a balance between rating efficiency and individual differences needs to be delicately maintained.

Linguistic profiles of high proficiency L2 English speakers—A combination of fluency and vocabulary features

Research background

In this study, linguistic profiles were established for L2 English speakers at high proficiency levels, who participated in an oral English test at a university in the United States. This test is generally administered to international graduate students monthly to probe their eligibility to serve as teaching assistants. A six-point holistic scale (35, 40, 45, 50, 55, and 60) was designed to evaluate test takers' performance. Speakers rated 50, 55, and 60 are considered proficient enough to teach undergraduate courses in English independently. Among all the test takers, L1 Mandarin and L1 Hindi speakers constitute the two largest examinee groups, since a high percentage of admitted international graduate students come from China and India.

The performance of each examinee was scored by two raters, and a third rater was consulted when disagreements occurred. While referring to a holistic rubric for test response evaluation, raters need to balance different factors before making their final scoring decision, such as grammar, pronunciation, and syntactic complexity. Before initiating the rating tasks, all raters assigned were required to participate in a monthly training session, which opens a space for discussions over benchmark speech samples and possible difficulties raters have encountered.

TABLE 1 Demographic information of test takers from 2009 to 2015.

	L1 Hindi speakers	L1 Mandarin speakers	Subtotal
Level 50	80	286	366
Level 60	32	11	43
Subtotal	112	297	

During the training sessions, raters repeatedly reported that delivery speed and vocabulary use were two of the most prominent characteristics in differentiating speakers across proficiency levels. Although examinees of high proficiency are often characterized by faster delivery speed and diverse use of vocabulary, speakers who obtained the same high score might still display different patterns of fluency and vocabulary features in their responses. For example, examinees who deliver with a fast speed were scored the same with test takers who speak slower. High speed delivery, however, is accompanied by less diverse or sophisticated vocabulary. In comparison, slower speakers are able to compose their responses with advanced words that appear more frequently in academic contexts. These different combinations of fluency and vocabulary features are not reflected in a holistic rubric, but might cause scoring hesitancy among raters of the test.

This issue becomes more prominent when speakers at a high proficiency level, those who have been rated 50 or above for the oral English test, are involved. Having surpassed the basic "threshold" of linguistic competence, L2 speakers at a high proficiency level may showcase stronger individuality in their use of language. Speakers rated 60 are expected to be differentiated from speakers rated 50 with less effort, and profiles constructed for speakers of these two scores are hypothesized to be clear-cut and more identifiable. This study thus focuses on L2 English speakers rated either 50 or 60 on the test, with an L1 Hindi or L1 Mandarin Chinese background. In total, 409 speech samples were collected from the examinees who participated in the test between the years 2009 and 2015. More detailed information about the speech samples can be found in Table 1 below.

As is shown in Table 1, the sample sizes are uneven for high-proficiency L2 English speakers with an L1 Mandarin or L1 Hindi background. From 2009 to 2015, 3,484 examinees took the oral English test, among whom L1 Mandarin speakers (n =1,166) and L1 Hindi speakers (n = 251) were two prominent examinee groups. As for all the L2 English examinees who scored 50 and above (n = 1,705), 235 out of 251 L1 Hindi speakers were rated as high-proficiency L2 English speakers (Level 50, Level 55, and Level 60), while the number for L1 Mandarin Speakers is 419. It is also noticeable that most of the high-proficiency L1 Mandarin speakers were rated 50 (n = 286), and only a

few L1 Mandarin speakers ($n = 11$) were scored 60. The fact that L1 Mandarin examinees outnumber L1 Hindi examinees resulted in a larger sample size of L1 Mandarin speakers rated 50. The limited number of L1 Mandarin speakers rated 60 may be attributed to the overall English instructional context in China, where more attention is placed on test preparation rather than speaking. Also, the admission policies for some of the departments set rather flexible minimum requirements for the TOEFL speaking score, which might reduce the possible impact of English proficiency level on international students' admission to graduate programs.

The oral English test contains 12 items in total and was designed in four different formats. Examinees' responses were recorded and saved in a data base for research purposes. In this study, the researcher analyzed test takers' response to the first test item, which lasts for 2 min maximum. The speakers needed to read a newspaper headline, and then express their opinions based on a short question. Four testing formats were randomly distributed to speakers during the test. All of the test items are closely related to campus life, which are indicated as the following:

a) Do you think that taking college courses on-line is a good way to study? Why or why not?
b) Do you think a television announcement will have a significant effect on the amount that they recycle? Why or why not?
c) Do you believe that class size affects the quality of education? Why or why not?
d) Do you think it is the university's responsibility to prevent students from illegally downloading music? Why or why not?

Research questions

The current study is designed to investigate the following research questions:

a) Will cluster analysis render linguistic profiles characterized by different combinations of fluency and vocabulary features?
b) Will L2 speakers' English proficiency level and their L1 background have an influence on their profile membership?

Research methods

In this study, utterance fluency and vocabulary are two major dimensions for the construction of linguistic profiles. Five indices were measured in total: Mean Syllables per Run (MSR), Speech Rate (SR), Pause Rate (PR), Measure of Textual Lexical Diversity (MTLD), and percentage of words in the Academic Word List (AWL). Table 2 explains the calculation procedure of the five indices, which represent fluency and vocabulary usage from multiple facets.

The second phase of this study involves a Hierarchical Cluster Analysis (HCA) based on all the five fluency and vocabulary measures. As a statistical method that recognizes homogeneity among data, cluster analysis places cases of similar numerical attribution into the same group. Staples and Biber (2015) provided more detailed explanations for the application of cluster analysis in applied linguistics research:

> Cluster analysis is a multivariate exploratory procedure that is used to group cases (e.g. participants or texts). Cluster analysis is useful in studies where there is extensive variation among the individual cases within predefined categories. For example, many researchers compare students across proficiency level categories, defined by their performance on a test or holistic ratings. But a researcher might later discover that there is an extensive variation among the students within those categories with respect to their use of linguistic features or with respect to attitudinal or motivational variables. (p. 243)

Among all the clustering techniques, Hierarchical Cluster Analysis (HCA) forms the backbone of cluster analysis (Everitt et al., 2011), where the concepts of homogeneity and separation are of great importance. All agglomerative hierarchical methods ultimately reduce data into one single cluster, while divisive techniques help split data into different groups. Agglomerative HCA is capable of producing a series of data partitions, or groups of speech samples demonstrating identical numerical patterns of the five measures.

In this study, the dataset used for cluster analysis include all four groups of speakers: L1 Hindi speakers rated 50, L1 Mandarin speakers rated 50, L1 Hindi speakers rated 60, and L1 Mandarin speakers rated 60. Profiles emerging from these cases could be used as informative evidence to identify individual linguistic features across different types of L2 speech output. Language teachers and educators are also able to collect important information and adjust pedagogical strategies accordingly.

Before cluster analysis is conducted, measurement outcomes for each of the five variables need to undergo normality and correlation check. Clustering results will be heavily influenced when highly-correlated features are simultaneously included, resulting in collinearity and inaccurate profile extraction. Researchers, however, may not be certain whether the variables selected are highly correlated due to the exploratory nature of cluster analysis. These statistical concerns thus demand for researchers' careful consideration before the implementation of cluster analysis.

To curb the influence of possible collinearity, PCA is often conducted to tackle the effect caused by significant correlation. The purpose of conducting PCA also lies in identifying index

TABLE 2 Utterance fluency and vocabulary measures included for cluster analysis.

Dimension	Linguistic feature	Interpretation	Measurement method	Measurement tool
Utterance fluency	Mean Syllables per Run (MSR)	The combination of speed fluency and breakdown fluency	MSR is calculated as syllable number divided by run number, with runs defined as "numbers of syllables produced between two silent pauses" (Ginther et al., 2010)	Fluencing software more information can be found in Park (2016)
	Speech Rate (SR)	Speed fluency	SR is calculated as the number of syllables divided by response time	
	Pause Rate (PR)	Breakdown fluency	PR is defined as the number of filled and unfilled pause divided by response time	
Vocabulary frequency	Percentage of AWL Words	The frequency of words included in the Academic Word List (AWL) in examinees' responses	The percentage of AWL words In each speaker's transcribed response is calculated	AntWordProfiler: Available from http://www.laurenceanthony.net/software
Lexical diversity	Measure of Textual Lexical Diversity (MTLD)	The mean length of word strings that "maintain a given Type-Token Ratio(TTR) value" of 0.72 (McCarthy and Jarvis, 2010, p. 384)	Each word is first evaluated for Type Token Ratio (TTR) sequentially as human readers process the text[1] The second step involves a factor count. If a word has met the cutoff TTR value of 0.72, the factor count would increase by 1[2]. A partial factor count is also provided for the remainder of a lexical item, which is calculated as the range covered between 1.00 and 0.72 The ultimate MTLD value is obtained through dividing the total number of words by the total factor count	Python program adapted from https://pypi.org/project/lexical-diversity/

[1]McCarthy and Jarvis (2010, p. 384) elaborated the TTR calculation procedure with the following sample text: "… of (1.00) the (1.00) people (1.00) by (1.00) the (0.800) people (0.667) for (0.714) the (0.625) people (0.556)… " The computed TTR results are presented in the parentheses.
[2]McCarthy and Jarvis (2010, p. 384) calculated the factor count for the example above as: "… of (1.00) the (1.00) people (1.00) by (1.00) the (0.800) **people (0.667) |||FACTORS = FACTORS +1|||** for (1.00) the (1.00) people (1.00) …" TTR result is reset at 1.00 after factor count increases.

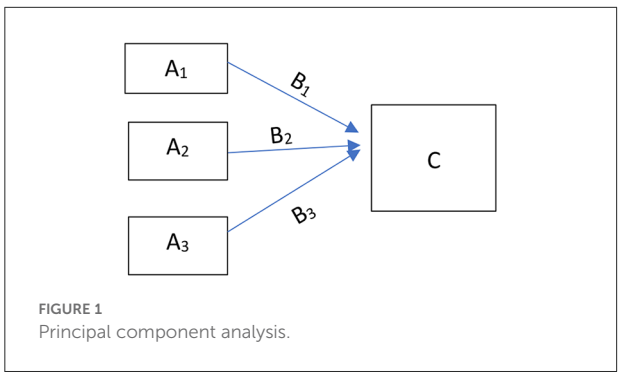


FIGURE 1
Principal component analysis.

variables from a larger set of measures, as researchers are capable of creating a new index variable through linear combination when correlated variables load on the same dimension. Figure 1 is an explanation of the working mechanism of PCA, where variable A_1, A_2, and A_3 are combined into one component C for further analysis. B_1, B_2, and B_3 are coefficient of the linear combination.

The differences between PCA and EFA were further explicated by Phakiti (2018): "While EFA aims at generalizing to the target population, PCA only aims at reproducing the sample being used" (p. 424). Although both Exploratory Factor Analysis (EFA) and Principal Component Analysis (PCA) are dimension reduction techniques that are exploratory in nature, they differ in theoretical assumptions. EFA is grounded in the assumption that all the observed variables could be explained by a latent variable. Under the framework of PCA, however, the variances of observed variables are calculated to derive a new component.

Research results

Linguistic profiling outcome

Descriptive statistics and box plots of Mean Syllables per Run (MSR), Speech Rate (SR), Pause Rate (PR), Measure of Textual Lexical Diversity (MTLD), and the percentage of words on the Academic Word List (AWL) are presented in Table 3 and Figure 2, respectively.

Boxplots in Figure 2 demonstrate that the five variables across the two proficiency levels are approximately normally distributed. In addition, the Kurtosis and Skewness statistics for

TABLE 3 Descriptive statistics of fluency and vocabulary measures.

	Variable	*N*	Mean	SD	Min	Max
Speakers rated 50	MSR	366	7.63	1.85	3.47	17.62
	SR	366	188.72	27.24	107.4	282.00
	PR	366	0.51	0.11	0.22	0.86
	MTLD	366	46.53	12.32	24.1	110.4
	AWL	366	4.12	2.18	0.00	14.30
Speakers rated 60	MSR	43	10.5	2.69	6.47	17.70
	SR	43	222.49	32.07	124.2	276.6
	PR	43	0.42	0.09	0.25	0.70
	MTLD	43	52.15	13.47	26.61	91.19
	AWL	43	4.79	4.79	1.5	10.8

all the variables are within the range between −0.61 and 2.56, indicating that the assumption of normality has been fulfilled.

The correlational results between variables are presented in Table 4. For all of the speakers in this study, Mean Syllables per Run (MSR) is strongly correlated with Speech Rate (SR; $r = 0.75$) and Pause Rate (PR; $r = -0.72$). These results are not unexpected, as MSR is a composite variable that integrates both speed fluency and breakdown fluency. The inclusion of MSR in this study is due to its strong effect in differentiating high proficiency L2 English speakers' performances. The two vocabulary measures, Measure of Textual Lexical Diversity (MTLD) and the percentage of words in the Academic Word List (AWL), are correlated with each other to a lesser extent. It should be pointed out that the correlation examination in the study applies to this particular dataset only, and bears limited inferential capacity. The measurement results for each group of speakers are located within a restricted range of L2 English proficiency level. No conclusion should be drawn as significant/insignificant correlation exists when a group of L2 English speakers at a different proficiency level are involved.

Correlational results in Table 4 suggest that PCA is needed to reduce fluency and lexical variables, so that components to be used for cluster analysis will not induce collinearity. Two components are expected to be created after PCA, where the three fluency variables would load on one component and the two vocabulary features would load on another. The two new components will later be used for Hierarchical Cluster Analysis (HCA).

Speakers rated 50 and 60 were pooled together for PCA, so that common coefficients of linear combination could be obtained. Before conducting PCA, all data were checked for Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity. KMO is a statistic indicating the proportion of variance that might be caused by underlying factors. While KMO close to 1 suggests inadequate sampling, values lower than 0.5 would lead to an unmeaningful interpretation of PCA results. As for Bartlett's test of sphericity, significant values smaller than 0.05 mean that PCA would be beneficial to data explanation. For the pooled group of data, Bartlett's test of sphericity is <0.01. The KMO measure of sampling adequacy is close to 0.6, which is above the minimum value recommended for PCA. Oblique (Promax) rotation is used for PCA, as fluency and vocabulary measures are assumed to be related in explaining language proficiency test performance.

The scree plot in Figure 3 suggests that two components can be extracted. As shown in Table 5, fluency measures are all significantly loaded on Component 1, while Component 2 includes the two vocabulary measures. Component 1 is thus named as fluency features, and Component 2 is named as vocabulary features. The two extracted components account for 68.68% of the variance among the five features. Correlation between the two components, which were used for the subsequent cluster analysis, is reduced to 0.20 after PCA.

HCA was applied to data analysis with Ward's method of minimum within-group variance. In this study, two main techniques were consulted to decide the number of clusters: (a) Dendrogram observation and (b) scree plot of coefficient change. Figure 4 shows the dendrogram generated for agglomerative HCA, and the scree plot for coefficient change is presented in Figure 5. Both the dendrogram and scree plot are references for deciding the number of clusters. The dendrogram in Figure 4 demonstrates a preliminary view of different clusters along the branches. The scree plot in Figure 5 shows a bending point following a sharp decline of coefficients. Additional new cases are not creating new clusters after the bending point, suggesting that a four-cluster solution is optimal.

Actual results from the cluster analysis are presented in Table 6, which includes descriptive statistics of Component 1 (fluency features) and Component 2 (vocabulary features) across clusters.

Each of the four clusters generated from HCA represents a profile. The cluster mean was transformed to an ordinal scale before the mean value of the five fluency and vocabulary features of each cluster are

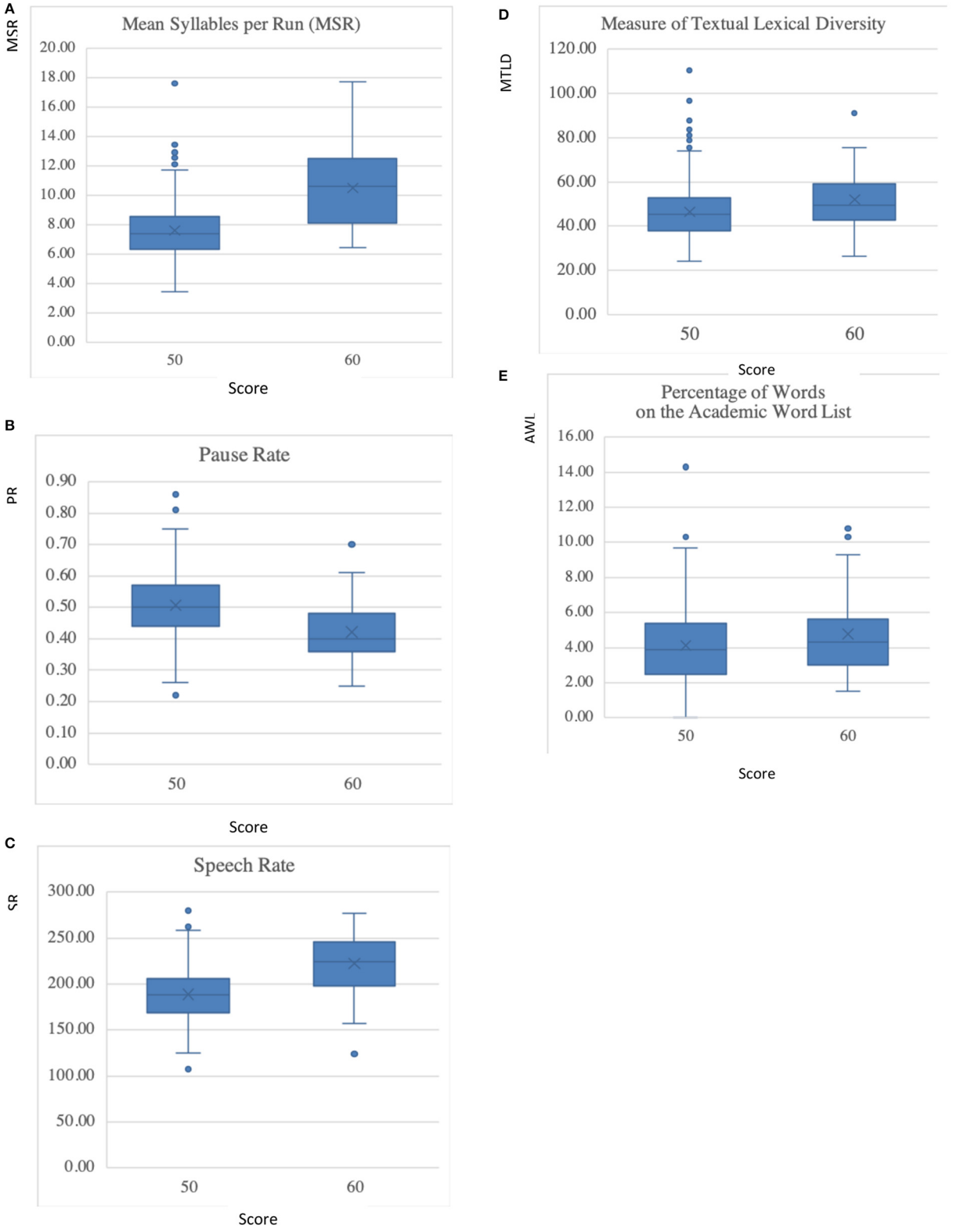


FIGURE 2
Boxplots of fluency and vocabulary measures across proficiency levels. **(A)** Mean Syllables per Run (MSR). **(B)** Pause Rate. **(C)** Speech Rate. **(D)** Measure of Textual Lexical Diversity (MTLD). **(E)** Percentage of Words in the Academic Word List.

TABLE 4 Correlation between variables for all speakers.

	MSR	SR	PR	MTLD	AWL
MSR	1				
SR	0.75**	1			
PR	−0.72**	−0.41**	1		
MTLD	0.11*	0.12*	−0.19*	1	
AWL	0.18**	0.12*	−0.14**	0.16*	1

**Correlation is significant at the 0.01 level (two-tailed).
*Correlation is significant at the 0.05 level (two-tailed).

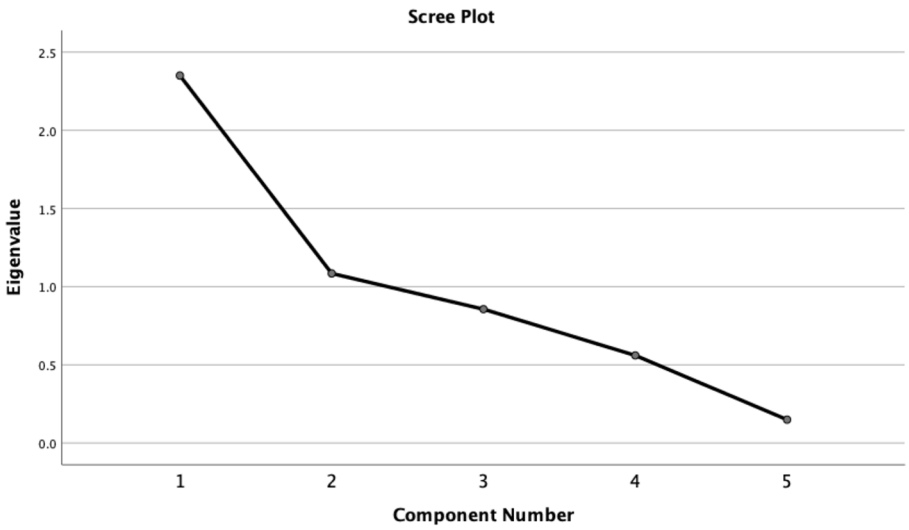


FIGURE 3
Scree plot for principal component analysis of fluency and vocabulary measures.

TABLE 5 Component loadings for speakers rated as 50 (after promax rotation).

	Component 1 Fluency features	Component 2 Vocabulary features
Mean Syllables per Run (MSR)	0.95	
Speech Rate (SR)	0.84	
Pause Rate (PR)	−0.80	
Measure of Textual Lexical Diversity (MTLD)		0.80
Percentage of Words on the Academic Word List (AWL)		0.70

reported. Table 7 lists more detailed information about the numerical range of each variable and its corresponding ordinal value. Table 8 demonstrates a closer examination at the five individual fluency and vocabulary variables, including mean values of each measure across the four clustered profiles.

Four profiles can be categorized based on the clustering results, and Table 10 provides a direct view of the characteristics demonstrated in these different profiles.

Profile 1: Low fluency measures + medium vocabulary measures

Low Mean Syllables per Run (MSR), low Speech Rate (SR), very high Pause Rate (PR), medium Measure of Textual Lexical Diversity, and medium percentage of words in the Academic Word List (AWL).

Profile 2: Medium fluency measure + low vocabulary measures

Medium Mean Syllables per Run (MSR), medium Speech Rate (SR), high Pause Rate (PR), low Measure of Textual Lexical Diversity, and low percentage of words in the Academic Word List (AWL).

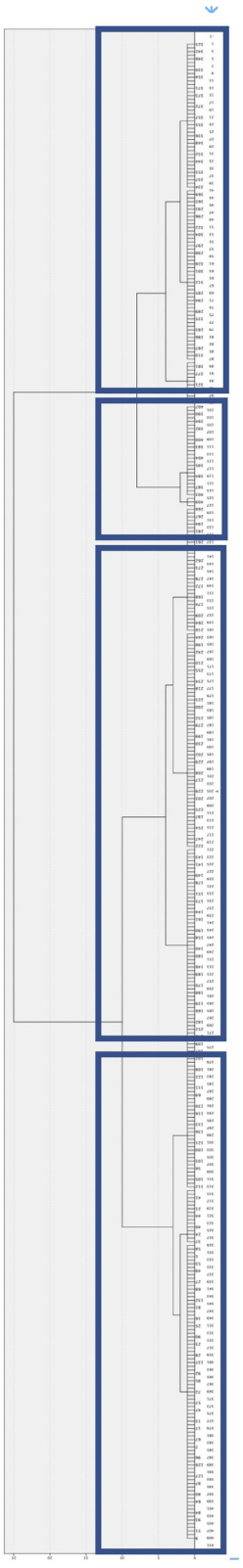


FIGURE 4
Hierarchical cluster analysis dendrogram.

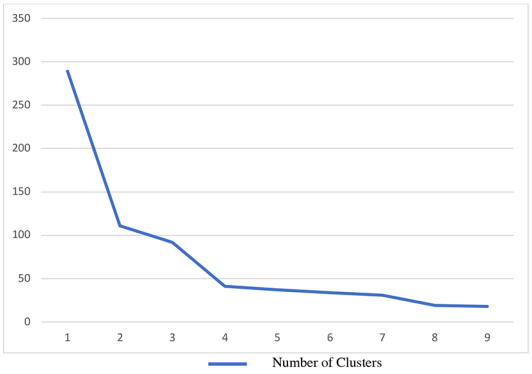


FIGURE 5
Hierarchical cluster analysis scree plot.

Profile 3: High fluency measures + medium vocabulary measures

High Mean Syllables per Run (MSR), high Speech Rate (SR), low Pause Rate (PR), medium Measure of Textual Lexical Diversity, and medium percentage of words in the Academic Word List (AWL).

Profile 4: Medium fluency measures + very high vocabulary measures

Medium Mean Syllables per Run (MSR), medium Speech Rate (SR), low Pause Rate (PR), very high Measure of Textual Lexical Diversity, and very high percentage level of words in the Academic Word List (AWL).

Profiles, L1 background, and L2 proficiency level

Further investigation into each profile with Chi-square test shows that profile membership is associated with speakers' L1 background ($\chi^2 = 49.84$, $p < 0.01$) and their overall oral proficiency level ($\chi^2 = 36.99$, $p < 0.01$). Table 9 presents general profile information characterized by fluency and vocabulary measures. Table 10 lists the number of speakers in each cluster grouped by their L1 background and oral English test scores. Although most of the L1 Hindi speakers rated 50 concentrated in Profile 3, the same profile also contains a large number of speakers rated 60.

The relationship between speakers' L1 background and their profile membership is displayed in Figures 6–9. Figure 6 is a percentage pie chart illustrating the profile membership of L1 Hindi speakers rated 50. Among all the L1 Hindi speakers who were rated 50, 27.5% of the speakers are in Profile 1, 20% of the speakers are in Profile 2, 40% of the speakers

TABLE 6 Descriptive statistics of component 1 (fluency features) score and component 2 (vocabulary features) score across clusters.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Total
Component 1					
Fluency features	−0.80	−0.13	1.17	0.47	0
Component 2					
Vocabulary features	0.22	−0.84	0.30	2.23	0

TABLE 7 Ordinal scale conversion of fluency and vocabulary features.

Ordinal scale	Low	Medium	High	Very high
Mean Syllables per Run (MSR)	MSR < 7	7 ≤ MSR ≤ 9	9 < MSR ≤ 11	MSR > 11
Speech Rate (SR)	SR < 180	180 ≤SR ≤ 200	200 < SR ≤ 220	SR > 220
Pause Rate (PR)	PR ≤ 0.45	0.45 ≤ PR ≤ 0.50	0.50< PR ≤ 0.55	PR > 0.55
Measure of Textual Lexical Diversity (MTLD)	MTLD <45	45 ≤ MTLD ≤ 55	55 < MTLD≤65	MTLD >65
Academic Word List (AWL)	AWL <4	4≤ AWL ≤5	5 < AWL≤6	AWL > 6

TABLE 8 Descriptive statistics of fluency and vocabulary measures across clusters.

	Number	Mean MSR	Mean SR	Mean PR	Mean MTLD	Mean AWL
Cluster 1	137	6.36	170.03	0.56	50.63	4.25
		Low	Low	Very High	Medium	Medium
Cluster 2	145	7.63	191.77	0.52	38.02	3.01
		Medium	Medium	High	Low	Low
Cluster 3	99	10.36	222.62	0.40	49.74	4.86
		High	High	Low	Medium	Medium
Cluster 4	28	8.63	199.91	0.42	69.61	7.67
		Medium	Medium	Low	Very High	Very High

TABLE 9 Distribution of fluency and vocabulary measures for each profile.

	Vocabulary measures			
	Low	**Medium**	**High**	**Very high**
Fluency measures				
Low		Profile 1		
Medium	Profile 2			Profile 4
High		Profile 3		

TABLE 10 Profile membership information.

	Profile 1	Profile 2	Profile 3	Profile 4	Total
Hindi 50	22	16	32	10	80
Hindi 60	4	6	19	3	32
Mandarin 50	109	122	41	14	286
Mandarin 60	2	1	7	1	11
Total	137	145	99	28	409

are in Profile 3, and 12.5% of the speakers are in Profile 4. The majority of L1 Hindi speakers are located in Profile 3 based on the measurement results of the five fluency and vocabulary features.

Figure 7 shows the percentage for L1 Mandarin speakers who were rated 50. According to the pie chart, 38.11% percent of the L1 Mandarin speakers are in Profile 1, 42.66% of the speakers are in Profile 2, 14.34% of the speakers are in Profile 3, and 4.9% of the speakers are in Profile 4. In comparison to L1 Hindi speakers rated as 50, most of the L1 Mandarin speakers rated as 50 are in Profile 1 and Profile 2 rather than Profile 3.

The pattern of distribution for the two groups of speakers rated 60, however, does not exhibit as much of a difference as the speakers rated 50. As is shown in Figures 8, 9, most of the L1 Hindi speakers (59.38%) and L1 Mandarin speakers (63.55%) are in Profile 3. However, more L1 Hindi speakers (18.75%) are in Profile 2 when compared with L1 Mandarin speakers (9.08%).

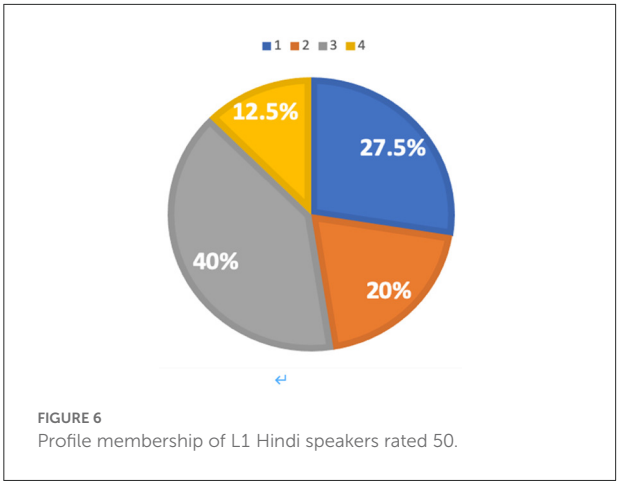


FIGURE 6
Profile membership of L1 Hindi speakers rated 50.

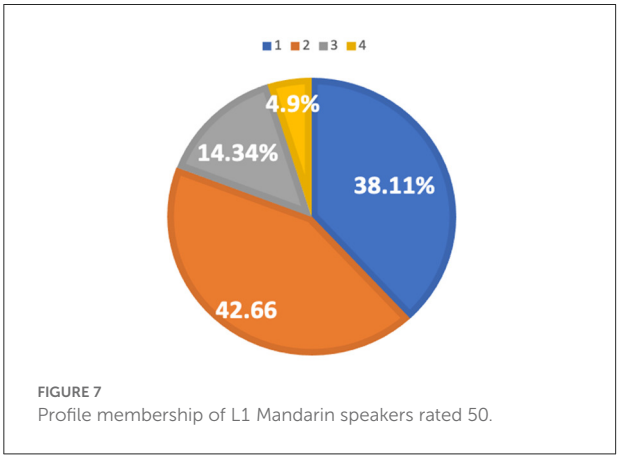


FIGURE 7
Profile membership of L1 Mandarin speakers rated 50.

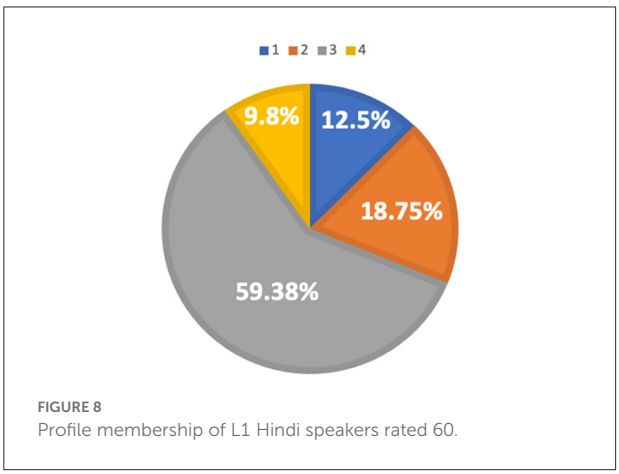


FIGURE 8
Profile membership of L1 Hindi speakers rated 60.

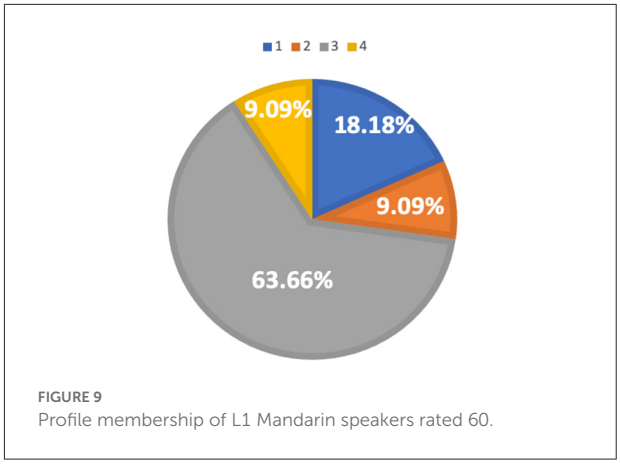


FIGURE 9
Profile membership of L1 Mandarin speakers rated 60.

Discussion of linguistic profiles

After cluster analysis, four different profiles emerged from the 409 speech samples, which have showcased different combinations of fluency and vocabulary features. The individuality of each profile indicates the connection between profiling analysis and L2 speaking pedagogy.

A straightforward observation of the profiling results is the reverse relationship between speakers' delivery speed and their use of vocabulary. More specifically, high delivery speed does not coexist with speakers' use of more diverse and complex vocabulary. Speakers in Profile 1 demonstrate medium values in fluency features and low values in vocabulary features, while the situation is the opposite for Profile 2. Speakers located in Profile 2 have enhanced values of vocabulary measures but deliver at a slower speed. A similar phenomenon can also be found in Profile 3 and Profile 4. Speakers of Profile 3 manifest high values in fluency features, in combination with medium values of vocabulary features. Speakers in Profile 4, however, showcase medium-level fluency features but very high values in vocabulary features.

The combination of fluency and vocabulary features in Profile 4 requires more detailed examination. Values for fluency measures (Speech Rate and Mean Syllables per Run) in Profile 4 are lower than Profile 3, and are closer to those of Profile 2. The vocabulary measures of Profile 4, however, are noticeably higher than in any other profiles. It is possible that speakers who use more diverse vocabulary and more academic words control their delivery speed on purpose. In this occasion, lower measures of Speech Rate and Mean Syllables per Run may indicate higher proficiency.

In addition, Profile 4 includes speakers across both proficiency levels and L1 backgrounds: 12.5% of the L1 Hindi speakers rated as 50, 9.8% of the L1 Hindi speakers rated 60, 9.8% of the L1 Mandarin speakers rated 50, and 9.09% of the L1 Mandarin speakers rated 60. Speakers in Profile 4 ranked above average with respect to both delivery speed and vocabulary use, and should have received a score of 60 based on their performance in these two dimensions. Contrary to this hypothesis, Profile 4 contains a relatively high number of speakers rated 50. Other linguistic features, such as discourse

structure, rhetorical patterns, and grammatical accuracy, might be influential factors worthy of exploration in future studies.

The association between speakers' L1 background and their profile membership may reflect the disparities in the English instruction received by the L2 English speakers. A large number of L1 Mandarin speakers rated 50 were located in Profile 1 and Profile 2, whereas L2 Hindi speakers rated 50 mostly fell in Profile 3. Profile 1 and Profile 2 is characterized by either slower delivery speed or less diverse vocabulary, which might suggest that instructions on fluency and vocabulary need to be prioritized in the refinement of L1 Mandarin speakers' oral English delivery. It could also be observed that both groups of L2 English speakers were rated 50 on a holistic rubric regardless of their differences in delivery speed and use of vocabulary. These two dimensions are important building blocks for linguistic profiles of L2 spoken English. However, raters' perception of speakers' overall English proficiency level may still be affected by other factors, such as accent, syntactic complexity, and grammatical accuracy.

This study also offers insights into the use of holistic rubrics in L2 speaking testing and assessment. Profiling results have indicated that speakers rated at the same level still fully demonstrate individuality regarding vocabulary and delivery fluency. Holistic rubrics, which ask raters to make the scoring decision based on an overall evaluation of all the key dimensions involved, could reduce possible disagreement and help reach consensus in an efficient manner. It is important for language researchers and teachers to keep in mind, however, that holistic rubrics are not in denial of the existence of individual differences. Learners' L1 background, L2 proficiency level, and delivery style are intricately intertwined, which requires more careful observation and analysis before explanations of individual differences can be provided.

Profiling, pedagogical implications, and possible concerns

From the perspective of language teaching and learning, linguistic profiling presents an opportunity for language teachers to adopt a more accommodating approach to pedagogy design. This study, for example, provided L2 speakers with possible guidance for delivery skill refinement. Speakers in Profile 3 demonstrated the fastest delivery speed, while speakers in Profile 4 used more diverse and complex vocabulary. Both profiles contained L2 speakers who scored 50 as well as L2 speakers who scored 60. In other words, delivering content at a fast speed and using diverse and complex vocabulary often offset one another. Achieving high measurement results at one dimension does not necessarily lead to an increase in the overall test score. It is possible that vocabulary use and fluency work together in a balanced way to reach the goal of effective communication. In spite of the fact that examinees who were rated 50 and 60 are qualified to teach undergraduate-level courses, mapping out their different linguistic profiles is still of great benefit for L2 English speakers if further progress is desired. It could also be hypothesized that speakers make use of different strategies intentionally when delivering in a second/foreign language, which might lead to the individuality demonstrated in their responses. More exploration is thus needed to investigate the reasons for individual differences in speech production. Possible research designs include connecting L2 speakers' language performance with cognitive tasks, or conducting qualitative interviews to inquire the strategies L2 speakers adopt for test-taking purposes.

Interpretation made from linguistic profiling results has been integrated in research fields such as law, criminology, and social justice (Welch, 2007; Legewie, 2016; Baugh, 2018; Minhas and Walsh, 2018). Equity is advocated through disconnecting linguistic profiling outcome from stereotypical attributes of the language users. To avoid misunderstanding linguistic profiling as one of the causes of stereotyping, researchers need to articulate the method and primary purpose. The construction of linguistic profiles is realized through the selection of key dimensions and the measurement of linguistic features. The research goal, however, is not to fit individual learners into a fixed category or pin labels on them based on anecdotal snippets.

Linguistic profiling research presents another important issue for consideration: How multi-faceted should linguistic profiles be? Researchers need to identify the major dimensions for establishing profiles, which could be represented by a myriad of quantitative indices. Index selection, however, often leads to a conflict between interpretability and parsimony. A limited volume of indices would result in an incomplete presentation of the major dimensions, causing inaccurate profile extraction. Various computational tools are indeed of great assistance for automatically calculating the quantitative indices' numerical values, but the involvement of a large number of indices may increase the risk of collinearity. It is not difficult to predict that methodological justifications and statistical interpretations would remain critical concerns for linguistic profiling research in the future.

Data availability statement

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

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Funding

This paper represents independent research part-funded by the Chenguang Scholar Grant (Grant Number 20CG04), which is jointly provided by Shanghai Municipal Education Committee and Shanghai Education Development Foundation.

Acknowledgments

The author would like to thank Dr. April Ginther and Dr. Nancy Kauper for kindly providing access to the testing data. Valuable comments and feedback from the journal editor and reviewers are also deeply appreciated.

Conflict of interest

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

Publisher's note

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

References

Ågren, M., Granfeldt, J., and Schlyter, S. (2012). "The growth of complexity and accuracy in L2 French: past observations and recent applications of developmental stages," in *Dimensions of L2 Proficiency and Performance: Complexity, Accuracy and Fluency in SLA*, eds A. Housen, F. Kuiken, and I. Vedder (Amsterdam and Philadelphia: John Benjamins Publishing Company), 95–119. doi: 10.1075/lllt.32.05agr

Alarcón, I. (2010). Advanced heritage learners of Spanish: a sociolinguistic profile for pedagogical purposes. *Foreign Lang. Ann.* 43, 269–288. doi: 10.1111/j.1944-9720.2010.01078.x

Banat, H. (2021). The status and functions of English in contemporary Lebanon. *World Englishes* 40, 268–279. doi: 10.1111/weng.12513

Bartening, I. (2000). Gender agreement in L2 French: pre-advanced vs advanced learners. *Stud. Linguist.* 54, 225–237. doi: 10.1111/1467-9582.00062

Baugh, J. (2018). *Linguistics in Pursuit of Justice*. Cambridge, UK: Cambridge University Press. doi: 10.1017/9781316597750

Brindley, G. (1998). "Describing language development? Rating scales and SLA," in *Interface Between Second Language Acquisition and Language Testing*, eds L. F. Bachman and A. D. Cohens (Cambridge, UK: Cambridge University Press), 112–140. doi: 10.1017/CBO9781139524711.007

Ciszér, K., and Dornyei, Z. (2005). Language learner" motivational profiles and their motivated learning behavior. *Lang. Learn.* 55, 613–659. doi: 10.1111/j.0023-8333.2005.00319.x

Clahsen, H. (1985). "Profiling second language development: a procedure for assessing second language proficiency," in *Modelling and Assessing Second Language Acquisition*, eds K. Hiltenstam and M. Piennemann (Clevedon, UK: Multilingual Matters), 283–331.

Custódio, J., and Paraboni, I. (2021). Stacked authorship attribution of digital texts. *Expert Syst. Appl.* 176, 114866. doi: 10.1016/j.eswa.2021.114866

Delaere, I., De Sutter, G., and Plevoets, K. (2012). Is translated language more standardized than non-translated language?: using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target Int. J. Transl. Stud.* 24, 203–224. doi: 10.1075/target.24.2.01del

Delmondes Neto, J., and Paraboni, I. (2022). Multi-source BERT stack ensemble for cross-domain author profiling. *Expert Syst.* 39, 1–19. doi: 10.1111/exsy.12869

Deutsch, C., and Paraboni, I. (2022). Authorship attribution using author profiling classifiers. *Nat. Lang. Eng.* 1—28. doi: 10.1017/S1351324921000383. [Epub ahead of print].

Divjak, D. T., and Gries, S. (2006). Ways of trying in Russian: clustering. *Corpus Linguist. Linguist. Theory* 2, 23–60. doi: 10.1515/CLLT.2006.002

Esseili, F. (2017). A sociolinguistic profile of English in Lebanon. *World Englishes* 36, 684–704. doi: 10.1111/weng.12262

Everitt, B. S., Landau, S., Leese, M., and Stahl, D. (2011). *Cluster Analysis*. Chichester, UK: Wiley.

Fabricius, A. (2006). The "vivid sociolinguistic profiling" of received pronunciation: responses to gendered dialect-in-discourse. *J. Sociolinguist.* 10, 111–122. doi: 10.1111/j.1360-6441.2006.00320.x

Friginal, E., Li, M., and Weigle, S. (2014). Revisiting multiple profiles of learner compositions: a comparison of highly rated NS and NNS essays. *J. Second Lang. Writ.* 23, 1–16. doi: 10.1016/j.jslw.2013.10.001

Friginal, E., and Weigle, S. (2014). Exploring multiple profiles of L2 writing using multi-dimensional analysis. *J. Second Lang. Writ.* 26, 80–95. doi: 10.1016/j.jslw.2014.09.007

Ginther, A., Dimova, S., and Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implication for automated scoring. *Lang. Test,* 27, 377–399.

Graham, S., Woore, R., Porter, A., Courtney, L., and Savory, C. (2020). Navigating the challenges of L2 reading: self-efficacy, self-regulatory reading strategies, and learner profiles. *Mod. Lang. J.* 104, 693–714. doi: 10.1111/modl.12670

Grandfeldt, J., and Ågren, M. (2014). SLA developmental stages and teachiers' assessment of written French: exploring Direkt Profil as a diagnostic assessment tool. *Lang. Test.* 31, 285–305. doi: 10.1177/0265532214526178

Gries, S. T. (2010). Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. *Ment. Lex.* 5, 323–346. doi: 10.1075/ml.5.3.04gri

Halteren, H. (2007). Author verification by linguistic profiling. *ACM Trans. Speech Lang. Process.* 4, 1–17. doi: 10.1145/1187415.1187416

Hoffmann, C. (2012). *Cohesive Profiling: Meaning and Interaction in Personal Weblogs*. Amsterdam/Philadelphia: John Benjamins. doi: 10.1075/pbns.219

Jarvis, S., Grant, L., Bikowski, D., and Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *J. Second Lang. Writ.* 12, 377–403. doi: 10.1016/j.jslw.2003.09.001

John, P., Brooks, B., and Schriever, U. (2017). Profiling maritime communication by non-native speakers: a quantitative comparison between the baseline and standard marine communication phraseology. *English Specif. Purp.* 47, 1–14. doi: 10.1016/j.esp.2017.03.002

Keβler, J. U., and Liebner, M. (2011). "Diagnosing L2 development: rapid profile," in *Studying Processability Theory: An Introductory Textbook*, eds M. Pienemann and J. U. Keβler (Amsterdam/Philadelphia: John Benjamins), 133–148. doi: 10.1075/palart.1.11dia

Kowsari, K., Heidarysafa, M., Odukoya, T., Potter, P., Barnes, L., Brown, D., et al. (2020). Gender detection on social networks using ensemble deep learning. *arXiv [Preprint]*. arXiv:2004.0651.

Kuznetsova, J. (2015). *Linguistic Profiles: Going from Form to Meaning via Statistics*. The Hague: Morton. doi: 10.1515/9783110361858

Legewie, J. (2016). Racial profiling and the use of force in police stops. *Am. J. Sociol.* 122, 379–424. doi: 10.1086/687518

Lundberg, J., and Laitinen, M. (2020). Twitter trolls: a linguistic profile of anti-democratic discourse. *Lang. Sci.* 79, 101268–101214. doi: 10.1016/j.langsci.2019.101268

Manna, R., Pascucci, A., and Monti, J. (2019). "Gender detection and stylistic differences and similarities between males and females in a dream tales blog," in *CLiC-it 2019 Sixth Italian Conference on Computational Linguistics* (Vol. 2481) (Bari).

McCarthy, P. M., and Jarvis, S. (2010). MTLD, vocd-D and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behav. Res. Methods* 42, 381–392.

Minhas, R., and Walsh, D. (2018). Influence of racial stereotypes on investigative decision making in criminal investigations: a qualitative comparative analysis. *Cogent Soc. Sci.* 4, 1538588. doi: 10.1080/23311886.2018.1538588

Moreno-Sandoval, L., Pomares-Quimbaya, A., and Alvarado-Valencia, J. (2021). Celebrity profiling through linguistic analysis of digital social networks. *Comput. Soc. Netw.* 8, 1–36. doi: 10.1186/s40649-021-00097-w

Park, S. (2016). *Measuring fluency: Temporal variables and pausing patterns in L2 English speech. (Unpublished Doctoral Dissertation)*. Purdue University, West Lafayette, IN.

Peng, J., Choo, K., and Ashman, H. (2016). Bit-level n-gram based forensic authorship analysis on social media: identifying individuals from linguistic profiles. *J. Netw. Comput. Appl.* 70, 171–182. doi: 10.1016/j.jnca.2016.04.001

Phakiti, A. (2018). "Exploratory factor analysis," in *The Palgrave Handbook of Applied Linguistics Research Methodology*, eds A. Phakiti, P. De Costa, L. Plonsky, and S. Starfield, S. (London, UK: Palgrave Macmillan), 423–457.

Pienemann, M. (1998). *Language Processing and Second Language Development: Processability Theory*. Amsterdam/Philadelphia: John Benjamins. doi: 10.1075/sibil.15

Pienemann, M., ed (2005). *Cross-linguistic Aspects of Processability Theory*. Amsterdam/Philadelphia: John Benjamins. doi: 10.1075/sibil.30

Pienemann, M. and Keβler, J. U., eds (2011). *Studying Processability Theory: An Introductory Textbook*. Amsterdam/Philadelphia: John Benjamins. doi: 10.1075/palart.1

Potocki, A., Ecalle, J., and Magnan, A. (2017). Early cognitive and linguistic profiles of different types of 7- to 8-year-old readers. *J. Res. Read.* 40, S125–S140. doi: 10.1111/1467-9817.12076

Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B., et al. (2014). "Improving the reproducibility of PAN's shared tasks. Plagiarism detection, author identification, and author profiling, in *Information Access Evaluation. Multilinguality, Multimodality, and Interaction*. CLEF 2014. Lecture Notes in Computer Science, 8685. Cham: Springer. doi: 10.1007/978-3-319-11382-1_22

Prentza, A., and Kaltsa, M. (2020). Linguistic profiling of heritage speakers of an endangered language: the case of Vlach Aromanian-Greek bilingual. *Open Linguist.* 6, 626–641. doi: 10.1515/opli-2020-0034

Ryslewicz, J. (2008). Cognitive profiles of (un)successful FL learners: A cluster analytical study. *Mod. Lang. J.* 92, 87–99.

Saito, K., Ilkan, M., Magne, V., Tran, M., and Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid- and high-level second language fluency. *Appl. Psycholinguist.* 39, 593–617. doi: 10.1017/S0142716417000571

Staples, S., and Biber, D. (2015). "Cluster analysis," in L. Plonsky (Ed.), *Advanced Quantitative Methods in Second Language Research* (New York, NY: Routledge), 243–74. doi: 10.4324/9781315870908-11

Van Compernolle, R. A. (2014). Profiling second language sociolinguistic development through dynamically administered strategic interaction scenarios. *Lang. Commun.* 37, 86–99. doi: 10.1016/j.langcom.2014.01.003

Wang, C., Verdon, S., McLeod, S., and Tran, V. (2021). Profiles of linguistic multicompetence in Vietnamese–English speakers. *Am. J. Speech Lang. Pathol.* 30, 1711–1727. doi: 10.1044/2021_AJSLP-20-00296

Welch, K. (2007). Black criminal stereotypes and racial profiling. *J. Contemp. Crim. Justice* 23, 276–288. doi: 10.1177/1043986207306870

Zheng, Y., Lu, X., and Ren, W. (2019). Profiling Chinese university students' motivation to learn multiple languages. *J. Multiling. Multicult. Dev. 40,* 590–604. doi: 10.1080/01434632.2019.1571074

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 29 November 2022
DOI 10.3389/fpsyg.2022.954217

OPEN ACCESS

EDITED BY
Hye Pae,
University of Cincinnati,
United States

REVIEWED BY
Hyunwoo Kim,
Yonsei University,
South Korea
Jacee Cho,
University of Wisconsin-Madison,
United States

*CORRESPONDENCE
Boping Yuan
by10001@sjtu.edu.cn;
by10001@cam.ac.uk

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

RECEIVED 27 May 2022
ACCEPTED 07 October 2022
PUBLISHED 29 November 2022

CITATION
Zhang L and Yuan B (2022) The role of L1 in L2 speech production at different stages of L2 development: Evidence from L2 Chinese oral production of verb-phrase ellipsis by English and Korean speakers.
Front. Psychol. 13:954217.
doi: 10.3389/fpsyg.2022.954217

The role of L1 in L2 speech production at different stages of L2 development: Evidence from L2 Chinese oral production of verb-phrase ellipsis by English and Korean speakers

Lulu Zhang[1] and Boping Yuan[2,3]*

[1]College of Foreign Languages and Literature, Fudan University, Shanghai, China, [2]School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China, [3]Faculty of Asian and Middle Eastern Studies, University of Cambridge, Cambridge, United Kingdom

The article reports on an empirical study investigating the role of L1 at the initial and developmental stages of L2 speech production. It examines two types of Chinese verb- phrase-ellipsis, ΣP-ellipsis licensed by the auxiliary *shi* 'BE' and vP-ellipsis licensed by the other auxiliaries, in 45 English and 45 Korean adult speakers' L2 Chinese speech production. An elicited imitation task was administered to L2 learners at beginner, intermediate and advanced Chinese proficiency levels. L1 influence is not observed at beginner levels, but it appears at intermediate and advanced levels, L1 influence disappears at different time in English and Korean learners' oral production of verb-ellipsis and ΣP-ellipsis. It is proposed that the absence of L1 influence at beginner levels is due to a breakdown of syntax-stylistics interface and beginners' difficulty in implementing checking and deleting operations in their L2 oral production. The different timings of the disappearance of L1 influence in the two language groups at advanced levels is attributed to interactions between the persistence of L1 influence and the computational complexity involved in the target elliptical structures.

KEYWORDS

L2 Chinese, English- and Korean-speaking learners, oral production, L1 influence, verb-phrase ellipsis

Introduction

It is widely observed that in contrast to the uniform success of children acquiring their first language (L1), few adult learners can achieve native-like competence in their acquisition of a second language (L2). Obviously, L2 learners already have a language, i.e., their L1, in their mind, which can exert influence on their L2 acquisition. (FTFA, Schwartz and Sprouse, 1994,

1996) propose an influential model called Full Transfer (FT) Hypothesis[1] in L2 acquisition research, which assumes that L1 grammar is transferred in its entirety to the initial state of L2 grammar. According to this hypothesis, the final state of grammatical properties of speakers' L1 constitutes the initial state of their L2 grammars, and the development of L2 grammars is failure-driven; that is, when the L2 grammar is not able to accommodate data in the target language input, it is restructured on the basis of the input. This hypothesis has been supported by a substantial amount of evidence in L2 acquisition research (e.g., Hawkins, 2001; Haznedar, 2001; Slabakova, 2013) and few researchers would deny the fact that L1 does play a role in L2 acquisition. It is also well-documented that positive L1 transfer can facilitate L2 acquisition, and learners with L1 structures similar to or the same as those in the target language can acquire the target structures easier than those without (e.g., Inagaki, 2002; Slabakova, 2015; Zufferey et al., 2015). However, some researchers have also noticed that L1 influence is not inevitable, and it can be overridden in L2 acquisition (e.g., Montrul, 2010; Scheidnes and Tuller, 2010; Prévost et al., 2014). For instance, in Yuan (2015), which investigates the acquisition of attitude-bearing *daodi*…*wh*-questions in L1 English learners' L2 Chinese, it is argued that L1 influence in L2 acquisition can be overridden by computational complexity. Specifically, unlike English *wh*-questions, where a *wh*-word is required to be raised from its base-generated position to the initial position of a sentence, a *wh*-word in Chinese *wh*-questions remains *in situ*. However, his study finds no evidence in the results that *wh*-movement in English is transferred into L1 English learners' L2 Chinese *wh*-questions and causes problem in this aspect of their L2 Chinese grammars. Also, Chinese and English share the same restriction on attitude-bearing *wh*-questions, which regulates that a question cannot have more than one attitude. Yuan's study shows that English speakers are unable to rule out ungrammatical Chinese *wh*-questions with two attitude features embedded in them, indicating that the similarities between English and Chinese in attitude-bearing *wh*-questions have very limited facilitation to L1 English learners' handling of L2 Chinese *wh*-questions with more than one attitude feature. On the basis of Prévost et al. (2014) and Scheidnes and Tuller (2010), Yuan (2015) argues that L1 transfer is a relative phenomenon rather than an absolute phenomenon in L2 acquisition, and it can be overridden by the computational complexity involved in a construction.

The present study is an attempt to track the role of L1 in L2 speech production at different stages of L2 development. It aims to examine whether L1 grammar is transferred to L2 oral production at initial stages of L2 development, as predicted by the FT Hypothesis (FTFA, Schwartz and Sprouse, 1994, 1996), how L1 influence varies in the L2 development and whether L1 influence is subject to constraints such as computational complexity of a grammatical structure in the development of the L2. The study focuses on L1 English and L1 Korean learners' L2 Chinese oral production of two types of verb-phrase ellipsis in an elicited oral production task. Chinese, English and Korean differ from each other in allowing certain types of verb-phrase ellipsis, which enables us to scrutinise the role that L1 plays in L1 English and L1 Korean learners' L2 Chinese oral production.

The article is structured as follows. Section "Cross-linguistic differences of verb-phrase ellipsis in Chinese, Korean and English" discusses syntactic analyses of verb-phrase ellipsis in Chinese, English and Korean, and Section "Prior studies of L1 influence on L2 oral production" briefly reviews prior studies on the L2 production of Chinese elliptical structures and outlines the research questions. Section "Present study" introduces the methodology of the present study, and Section "Results" reports the scoring methods and results. The results are discussed in Sections "Discussion" and "Conclusion" contains our conclusions.

Cross-linguistic differences of verb-phrase ellipsis in Chinese, Korean, and English

Chinese allows two types of verb-phrase ellipsis: a verb-phrase ellipsis licensed by the auxiliary *shi*是 "BE", as exemplified in (1), and a verb-phrase ellipsis licensed by auxiliaries other than *shi* "BE", as illustrated in (2). As noted in Soh (2007), the scope of ellipsis licensed by *shi* 'BE' is larger than that licensed by the other auxiliaries like *hui* 会 'will'. As shown in the contrast between (1) and (2), the elided constituent in the latter includes the verb phrase *likai yingguo* "leave the UK", whereas that in the former includes the auxiliary *hui* 'will' as well as the verb phrase *likai yingguo* 'leave the UK'. Also, as can be seen in the contrast between (3) and (4), when containing the negator *bu* 不 'not' in the antecedent clause, the scope of ellipsis licensed by *shi* 'BE', as shown in (3), includes the negator *bu* 'not', but that licensed by the auxiliary *hui* 'will' does not, as shown in (4).

1. 张三 会 离开 英国， 李四 也 是 会 离开 英国。
Zhangsan hui likai yingguo, Lisi ye shi hui likai yingguo.
Zhangsan will leave the UK Lisi also BE will leave the UK.
'Zhangsan will leave the UK, and Lisi will (leave the UK) too.'

2. 张三 会 离开 英国， 李四 也 会 ~~离开 英国~~。
Zhangsan hui likai yingguo, Lisi ye hui ~~likai yingguo~~.
Zhangsan will leave the UK Lisi also will leave the UK.
'Zhangsan will leave the UK, and Lisi will (leave the UK) too.'

3. 张三 不 会 离开 英国， 李四 也 是 ~~不 会 离开 英国~~。
Zhangsan bu. hui likai yingguo, Lisi ye shi ~~bu hui likai yingguo~~.
Zhangsan not will leave the UK Lisi also BE not will leave the UK.

1 The full name of the model proposed by Schwartz and Sprouse (1994, 1996) is Full Transfer and Full Access Hypothesis. As we are only concerned with the transfer part, but not the access part, of the model in this article, we, hereafter, simply call the model Full Transfer (FT) Hypothesis to highlight our focus on the transfer part of the model.

'Zhangsan will leave the UK, and Lisi will (leave the UK) too.'

4. 张三 不 会 离开 英国， 李四 也 不 会 ~~离开 英国~~。
Zhangsan bu. hui likai yingguo, Lisi ye bu hui ~~likai yingguo~~.
Zhangsan not will leave the UK Lisi also not will leave the UK.
'Zhangsan will not leave the UK, and Lisi will not (leave the UK) either.'

Based on the above observations, the present study follows Soh (2007) by assuming that *shi* 'BE', a dummy auxiliary in Chinese, occupies the head of TP, a position higher than the other auxiliaries such as *hui* 'will' in the hierarchy. Following Chomsky's (1995) proposal that English auxiliaries are generated under Mod(al)P in the hierarchy, Soh (2007) argues that the auxiliaries in Chinese such as *hui* 'will' are generated under a Mod(al) node, which is lower than T, where *shi* 'BE' is located. The positions of the auxiliary *shi* 'BE' and the other auxiliaries in the hierarchy in Soh's (2007) proposal are demonstrated in Figure 1. As can be seen, the auxiliaries exemplified by *hui* 'will' occupy the head of ModP, lower than the category Σ, which can be realized by the negator *bu* 'not' to express negative meaning; in contrast, *shi* 'BE' occupies the head of TP, which is higher than ΣP.

The fact that *shi* 'BE' and the other auxiliaries occupy different structural positions implies that they also differ in terms of the constituent they license. Thus, in line with Soh's (2007) proposal, the scope of ellipsis licensed by *shi* 'BE' is a ΣP, whilst that licensed by the other auxiliaries like *hui* 'will' is a vP.

According to Li (2014), verb-phrase ellipsis in Chinese is a result of PF deletion. That is, when deriving a sentence with verb-phrase ellipsis, the elliptical verb phrase is first fully spelled out on the surface and then a deletion applies at the PF, resulting in the inaudibility of the verb phrase. A verb phrase can be deleted when verbal identity condition is met. That is, a vP or a ΣP can undergo deletion when it is identical to the antecedent in the first coordinate sentence (Chung, 2013; Liu, 2014; *cf.* Merchant, 2001).

It should be noted that complete sentences without verb-phrase ellipsis, as shown in (5), are perfectly grammatical in Chinese. However, they are stylistically heavy and unconcise, and thus are less preferred than the elliptical counterparts. Then, combining the above facts, we assume that in the derivation of a sentence with verb-phrase ellipsis, the deletion of the vP or ΣP is triggered at the syntax-stylistics interface, on the premise that the verbal identity condition is met.

5. 张三 会 离开 英国， 李四 也 会 离开 英国。
Zhangsan hui likai yingguo, Lisi ye hui likai yingguo.
Zhangsan will leave the UK Lisi also will leave the UK.
'Zhangsan will leave the UK, and Lisi will leave the UK too.'

It is widely observed that English allows its verb-phrase ellipsis to be licensed by auxiliaries like *will* and *can* (Adger, 2003; *cf.* Johnson, 2001; Merchant, 2001, 2004). As shown in (6), sentences with ellipsis licensed by the auxiliary *will* in English seem to behave analogously to the Chinese vP-ellipsis licensed by auxiliaries in (2) and (4) respectively. Here we adopt Soh's (2007) proposal that English auxiliaries are generated under a Modal node, occupying the head of ModP in the hierarchy. In this sense, like Chinese vP-ellipsis, what is elided in English elliptical sentences licensed by auxiliaries, like *will* in (6), is also a vP.

6. a. John will leave the UK and Bill will ~~leave the UK~~ too.
b. John will not leave the UK, and Bill will not ~~leave the UK~~ either.

It is worth mentioning that a verb-phrase ellipsis in English can also be licensed by the dummy *do*, which is believed to behave similarly to the Chinese *shi* 'BE' in licensing some ΣP-ellipsis in English (Xu, 2003; Soh, 2007; Li and Wei, 2013). As can be seen in the Chinese sentence in (7) and its English translation, when the verb phrase *like Xiaohong* is elided, the dummy *do* is inserted to

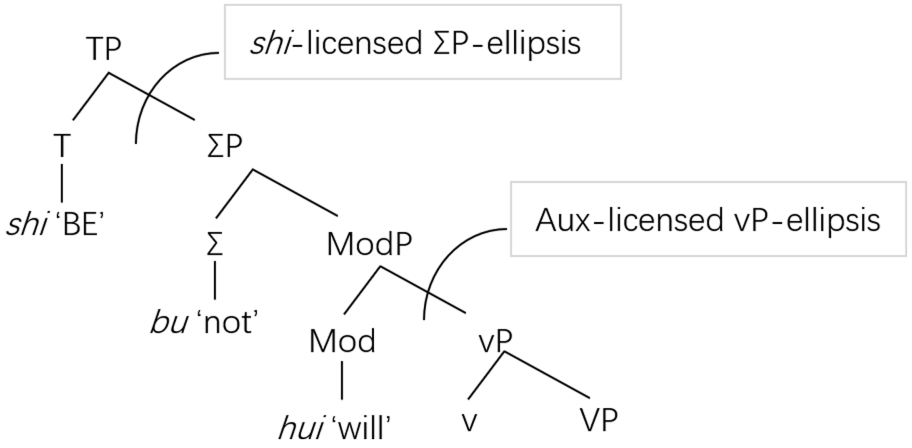


FIGURE 1
Positions of *shi* "BE" and auxiliaries in the hierarchy. Drawn using data from Soh (2007).

TABLE 1 Summary of the availability of vP-ellipsis and ΣP-ellipsis in Chinese, English and Korean.

	Chinese	English	Korean
vP-ellipsis	+ (Aux-licensed)	+ (Aux-licensed & *do*-licensed)	–
ΣP-ellipsis	+ (*shi*-licensed)	–	+ (*ya*-licensed)

license the ellipsis in English, just like what *shi* 'BE' does in Chinese. However, unlike ΣP-ellipsis in Chinese, the scope of *do*-licensed verb-phrase ellipsis in English cannot include an auxiliary or negator, as exemplified in the sentences in (8). According to Soh (2007), this is because the English dummy *do* is positioned at the head of ModP, which is the same as the other auxiliaries in English but different from *shi* 'BE' in Chinese, which is at the head of TP. Consequently, what is elided after the dummy *do* is a vP rather than a ΣP, leading to the fact that English allows vP-ellipsis but not ΣP-ellipsis.

7. 张三 喜欢 小红， 李四 也 是 ~~喜欢 小红~~
Zhangsan xihuan Xiaohong, Lisi ye shi ~~xihuan Xiaohong~~.
Zhangsan like Xiaohong Lisi also BE like Xiaohong.
'Zhangsan likes Xiaohong, and Lisi does (like Xiaohong) too.'

8. a. *Zhangsan will leave the UK, and Lisi does ~~will leave the UK~~ too.
b. *Zhangsan will not leave the UK, and Lisi does ~~will not leave the UK~~ either.

It has been observed in the literature that languages like Japanese and Korean do not allow auxiliaries to license a verb-phrase ellipsis. This is because auxiliaries in these languages, like *-eul* 'will' in the Korean sentence in (9), are verbal suffixes, and thus deleting a verb phrase in a sentence and leaving the auxiliary alone would produce a 'stray affix'[2], rendering the remnant of the sentence ungrammatical.

9. 존이 영국을 떠날 거야. 빌도 영국을 떠날 거야.
Jon-i yeonggug-eul tteona-l geoya. Bil-do yeonggug-eul tteona-l geoya.
John-NOM England-ACC leave will Bill-too England-ACC leave will.
'John will leave the UK, and Bill will (leave the UK) too.'

However, it is found that a verb-phrase ellipsis in Korean can be licensed by a lexical item *ya*야 'BE'[3] (Kim and Sohn, 1998). As illustrated in sentences in (10), the scope of the ellipsis licensed by *ya* 'BE' in Korean can include an auxiliary (e.g., *yongkuk-eul donal-koeaya* 'will leave the UK' in (10a)) and a negator (e.g., *yongkuk-eul an donal-koeaya* 'will not leave the UK' in (10b)). According to Kim and Sohn (1998), the lexical item *ya* is inserted to the head of TP after a focus-movement process and the deletion of a ModP. Based on the derivation analysis and the examples above, it can be summarised that the lexical item *ya* occupies the head of TP, a position higher than auxiliaries in Korean. On the basis of this analysis, we can assume that in Korean, the scope of the ellipsis licensed by *ya* is a ΣP, which can include an auxiliary and a negator, and that Korean allows ΣP-ellipsis, but not vP-ellipsis.

10. a. 존이 영국을 떠날 거야. 빌도 ~~영국을 떠날 거~~야.
Jon-i yeonggug-eul tteona-l geoya. Bil-do ~~yeonggug-eul tteona-l geo~~ya.
John-NOM England-ACC leave-will Bill-too England-ACC leave-will.
'John will leave the UK, and Bill will (leave the UK) too.'

b. 존이 영국을 떠나지 않을 거야. 빌도 ~~영국을~~.
Jon-i yeonggug-eul tteona-ji an-heul geoya. Bil-do ~~yeonggug-eul~~.
John-NOM England-ACC leave not will Bill-too England-ACC.
~~떠나지 않을 거~~야.
~~tteona-ji an-heul geo~~ya.
leave not will.
'John will not leave the UK, and Bill will not (leave the UK) either.'

Cross-linguistic differences with regard to the availability of vP-ellipsis and ΣP-ellipsis are summarised in Table 1.

Prior studies of L1 influence on L2 oral production

Much evidence of L1 influence on adult L2 speech production has been reported in the literature, and many features in L2 oral production find their origin in the.

L1 (see overviews in Kellerman and Sharwood Smith, 1986; Gass and Selinker, 1992; Odlin, 1989, 2003). However, evidence has been emerging in the literature that L2 influence in L2 speech production is not inevitable, and it is argued in 2008Yuan (2001) that L1 influence is not everywhere. His argument is based on oral production data and judgment data

2 The term *stray affix* in Lasnik (1981)is used to describe a situation where a verbal suffix is illicitly used independently in languages like Japanese and Korean.

3 As pointed out by the handling editor, unlike *shi* 是 'BE' in Chinese, which is a free morpheme, *ya*야 in Korean is a bound morpheme which has to be attached to a verbal element. What is in common between the two, though, is that both of them can license a ΣP-ellipsis in their respective languages. Also, as mentioned by a Korean native speaker, both *ya*야 and *yo*요 can license a ΣP-ellipsis. *Ya* is used in informal sentences while *yo* is a formal form. The current article follows Kim and Sohn (1998) and only uses *ya* in Korean examples.

concerning thematic-verb raising collected from adult French-, German- and English-speaking learners of L2 Chinese. Thematic verbs are allowed to raise in French and German, but not in English and Chinese, and Yuan's findings show that neither French- nor German-speaking adult learners' L2 oral production of Chinese is influenced by the thematic-verb raising in their L1 French and German, which shows clear absence of L1 transfer in L2 speech production and provides evidence against the FT Hypothesis (FTFA, Schwartz and Sprouse, 1994, 1996).

Absence of L1 influence is also reported in Hawkins and Casillas (2008), although their study is not to confirm or disconfirm the FT Hypothesis. In their study, adult Chinese- and Spanish-speaking learners of English are examined for their use of subject-verb agreement in their L2 English. Subject-verb agreement is realised in.

Spanish but not in Chinese, and if properties of L1 verb morphology are influential in the acquisition of English, this difference should show up in the performance of the.

two groups. Results of an oral completion task show that both groups perform strikingly similarly; (i) the copula */is/* is supplied more than the 3rd person singular */s/* with simple subjects, and there is no overgeneralisation of */is/* or */s/* when the subject is plural; (ii) there is no decrease in the suppliance of */is/* or */s/* when there is a complex subject; (iii) suppliance of */s/* with a complex subject is disrupted when there is an intervening prepositional phrase (PP); however, suppliance of the copula */is/* is only disrupted where the PP contains a plural N, not when both Ns are singular. The similar behaviours of the two groups in Hawkins and Casillas (2008) suggest again that the L1 is unlikely to be influential in determining the knowledge that gives rise to these patterns of behaviours.

While evidence for the absence of L1 influence in L2 speech production is emerging, it is still not as robust as evidence for such influence. In addition, the variety of language phenomena tested for the former is still rather limited. More importantly, answers are yet to be found as to why there is absence of L1 influence on some L2 structures given that L1 transfer is a rather pervasive phenomenon in L2 speech production.

There has been considerable linguistic research on the syntactic mechanism underlying ellipsis (e.g., Grinder and Postal, 1971; Lobeck, 1995; Kehler, 2000; Johnson, 2001; Hendriks, 2004; Kertz, 2013), as well as psycholinguistic research examining parallelism effects on ellipsis (Arregui et al., 2006; Matsuo, 2007; Frazier, 2008; e.g., Matsuo and Duffield, 2001), but only a few studies have investigated elliptical structures in L2 speech production, one of which is Yuan and Zhang's (2020) study, which investigates object ellipsis in L2 Chinese speech production by adult L1 Korean and L1 English learners at various L2 Chinese proficiency levels. They adopt an analysis of Chinese object-ellipsis structures on the basis of topicalization and topic deletion (Li and Thompson, 1981; Huang, 1984), and argue that the equivalent of object ellipsis is allowed in Korean but not in English. An elicited imitation task[4] was used to test L2 speech production of the target elliptical structures. In the study, both Korean- and English-speaking beginner learners of L2 Chinese are found to overwhelmingly produce utterances with overt objects after they hear sentences with object ellipsis. The authors' explanation for the absence of the object ellipsis in L2 Chinese beginners' oral production is based on an incremental model for speech production (adapted from Bock and Levelt, 1994). The model proposes four stages during the grammatical encoding for speech production. Specifically, lexical concepts and lemmas are selected for conveying the message at the first stage, and are assigned grammatical functions at the second stage. At the third stage, the constituents are assembled in a word order suitable for the target sentence, while at the final derivation stage some procedures such as movement and deletion take place before the sentence is spelt out. The authors argue that beginner learners encounter problems in handling the movement and deletion procedures at the derivation stage, rendering the overwhelming production of non-ellipsis responses in L1 English and L1 Korean beginners' L2 Chinese oral production. Another finding of the study is that results of both the elicited imitation task and an acceptability judgment task suggest no L1 influence on speech production throughout L2 Chinese developmental stages, as no significant difference is found between L1 English groups and proficiency-matched L1 Korean groups. However, no specific account is provided in Yuan and Zhang (2020) as to why no L1 influence is found in their study. Another study reported in Zhang (2020) yields an inconsistent finding. This study explores the role of L1 in L2 acquisition of verb-phrase ellipsis, and the results are discussed from the perspective of structural priming effect, i.e., whether language users tend to reuse the same grammatical structure as the one in recent discourse (Bock, 1986). Specifically, Zhang (2020) examines data from an elicited imitation task by 77 intermediate L1 English and L1 Korean learners of L2 Chinese. The data shows an obvious difference between L1 English and L1 Korean groups; when primed for a certain type of verb-phrase ellipsis structure, learners whose L1 has the equivalent of the ellipsis type produce significantly more responses with the primed ellipsis structure, displaying a significantly stronger priming effect than those whose L1 does not have the equivalent. The author attributes the between-group difference to L1 influence, and concludes that at intermediate levels, learners' L2 speech production is affected by the presence or absence of the equivalent of the primed structure in their L1s. This finding supports the language-nonspecific account in Flett et al. (2013), which argues that the magnitude of a structure's priming effect in L2 speech production is influenced by both the speaker's L2 and L1, rather than by their L2 only. The

4 The elicited imitation task is called *utterance-recall task* in Yuan and Zhang (2020).

finding, however, left a question unanswered as to why L1 difference is found in L2 Chinese production of verb-phrase ellipsis, but is absent in L2 Chinese production of object ellipsis, as observed by Yuan and Zhang (2020). More importantly, since Zhang (2020) focuses on intermediate learners of L2 Chinese in her study, it remains unclear whether the significant difference between different L1 groups' L2 Chinese oral production of verb-phrase ellipsis occurs at stages before the intermediate level, particularly at beginner levels, and whether it can be overcome beyond the intermediate level. Thus, the unexplored questions become the aims of the present study, which is to provide a full picture about the role of L1 in L2 Chinese speech production during the L2 development, from beginner to advanced levels. Attempts are to be made to account for the occurrence and disappearance of L1 influence in L2 Chinese oral production of vP- and ΣP-ellipsis.

Present study

Research questions and predictions

On the basis of the cross-linguistic differences with regard to the (un)availability of vP-ellipsis and ΣP-ellipsis in English, Korean as well as Chinese, the following research questions are asked in this study.

Research Question 1. Is English- and Korean-speaking L2 Chinese beginners' oral production of vP- and ΣP-ellipses influenced by their L1s?

Predictions: On the basis of the Full Transfer (FT) Hypothesis (FTFA, Schwartz and Sprouse, 1994, 1996) and given the fact that English allows vP-ellipsis but not ΣP-ellipsis while Korean allows the latter but not the former, it is predicted that L1 influence will occur in beginners' L2 Chinese speech production and that the L1 influence will lead to differences between English and Korean in (dis)allowing vP-ellipsis and ΣP-ellipsis at L2 Chinese beginner levels. Specifically, (i) L1 English beginners are predicted to produce more vP-ellipsis sentences in their L2 Chinese speech production than L1 Korean beginners; (ii) L1 Korean beginners are predicted to produce more ΣP-ellipsis sentences in their L2 Chinese speech production than L1 English beginners.

Research Question 2. To what extent does L1 play a role in L2 Chinese oral production of vP- and ΣP-ellipsis at different stages of the L2 Chinese development? Does L1 influence persist or disappear at advanced levels? Specifically, do L1 English and L1 Korean advanced learners of L2 Chinese behave similarly in their oral production of sentences with vP- or ΣP-ellipsis?

Predictions: If L1 influence persists at the advanced level, L1 English and L1 Korean advanced learners will behave differently to each other in their oral production of target sentences with vP- or ΣP-ellipsis; if L1 influence can be overcome, advanced learners from different L1 backgrounds will not differ significantly in their oral production of target sentences.

Participants

The total number of participants in the empirical study is 105, which includes 45 adult L1 English and 45 adult L1 Korean learners of L2 Chinese as well as 15 adult native Chinese speakers as a control group. They were mainly students from universities in Britain and China at the time of data collection. The L1 English and L1 Korean participants all had previously received classroom instruction in Chinese language, and most of them had spent a certain period of time in China by the time of the experiment. Native English speakers who had learned any East Asian languages other than Chinese, such as Korean or Japanese, were excluded. For native Korean speakers, as English is a compulsory course in universities in South Korea, it is unavoidable that all of them have learned English for some time. Those who had not been to any English-speaking country and self-rated their English as lower than advanced level (i.e., elementary level or intermediate level) were selected. Payments were given to every participant as a token of thanks for their participation in the study.

Participants' working memory capacity is also controlled. The task chosen in the current study to test participants' working memory capacity is the backward digit span task, which is one of the subtests of Wechsler Adult Intelligence Scale–Fourth UK Edition (Wechsler, 2010), and has been used in recent literature (Gathercole et al., 2004, 2008; Gathercole and Alloway, 2007; Hsieh, 2015). In the task, participants first listen to a digit span (one digit per second) read in their native languages and then are required to repeat the span backwards. The score of the task is the highest number of digits that a participant is able to correctly repeat. To ensure that participants have similar working memory capacity, those who are only able to correctly repeat fewer than 6 digits are excluded from the study. The statistical data and the results of one-way ANOVA of the backward digit span test scores of different L1 groups are shown in Table 2.

The remaining participants are divided into seven groups based on their native languages and their performance in a cloze test. The cloze test is adopted from Mai and Yuan (2016), which consists of 3 passages and contains 40 gaps in total. Participants are required to fill in the gaps using correct Chinese characters or Pinyin (an alphabetical system for Chinese pronunciation). The maximum number of correct responses in this test is 40. Information of the participants and results of the cloze test for

TABLE 2 Results of the backward digit span test.

English speakers		Korean speakers		Chinese speakers			
Mean	***SD***	**Mean**	***SD***	**Mean**	***SD***	***F* (2,102)**	***p***
6.71	0.46	6.62	0.49	6.53	0.52	0.880	0.418

TABLE 3 Information about participants in each group.

Groups	n (male/female)	Age		Onset age of learning Chinese		Time spent learning Chinese (months)		Duration of stay in China (months)		Cloze test	
		Mean (range)	*SD*	Mean (range)	*SD*	Mean (range)	*SD*	Mean (range)	*SD*	Mean (range)	*SD*
EB	15 (8/7)	22 (17–27)	3.23	20 (17–25)	2.76	17 (4–48)	14.65	2 (0–16)	5.40	6 (6–13)	1.86
EI	15 (8/7)	21 (19–27)	2.27	18 (17–25)	2.05	34 (5–96)	31.08	7 (0–15)	5.97	19 (15–24)	2.88
EA	15 (7/8)	23 (21–29)	2.72	17 (15–22)	1.63	58 (38–108)	17.82	17 (10–30)	6.67	33 (30–37)	2.43
KB	15 (7/8)	22 (18–25)	2.35	21 (17–25)	2.58	7 (1–24)	6.83	2 (1–8)	2.13	6 (6–13)	2.06
KI	15 (8/7)	22 (19–25)	1.99	19 (17–23)	1.76	36 (3–84)	25.07	27 (1–72)	20.92	19 (16–24)	2.59
KA	15 (6/9)	22 (18–28)	2.53	17 (15–20)	1.41	61 (36–96)	22.47	47 (6–72)	18.19	32 (29–37)	2.37
NS	15 (9/6)	24 (18–30)	3.75	N/A	N/A	N/A	N/A	N/A	N/A	39 (36–40)	1.28

EB, English Beginner Group; EI, English Intermediate Group; EA, English Advanced Group; KB, Korean Beginner Group; KI, Korean Intermediate Group; KA, Korean Advanced Group; NS, Native Speaker Group.

each group are given in Table 3. A one-way ANOVA is administrated on the cloze test scores between the learner groups and the NS Group, and the results reveal a significant difference between the groups in their performance in the cloze test ($F(6, 98) = 465.763$, $p < 0.001$). *Post hoc* Scheffé tests indicate that all learner groups are significantly different from the NS group. The results also show that there is no significant difference between any of the two corresponding language groups in their scores in the cloze test; that is, no significant difference is found between the EB Group and the KB Group ($p > 0.05$), between the EI Group and the KI Group ($p > 0.05$), or between the EA Group and the KA Group ($p > 0.05$). These indicate that all of the English groups are compatible with their corresponding Korean groups with regard to their Chinese language proficiency.

Instruments

Participants are required to complete a language background questionnaire and the working memory test prior to the main experiment, which includes an elicited imitation task[5] for eliciting L2 learners' Chinese speech production of sentences with vP- or ΣP-ellipsis.

Before the experiment begins, each participant is required to read aloud the words and phrases on a vocabulary list for the experiment and tell the administrator the meaning of each character/phrase. This is to make sure that their performance in the task is not to be affected by vocabulary issues. Both written and oral instructions are provided in participants' L1s, and five practice trials are given to the participant before the experiment starts. In the experiment, recorded utterances are presented to the participant auditorily one by one, and then the participant is prompted to recall the utterance orally. On each trial in the experiment, participants first read contextual information conveyed by a picture on the computer screen and a sentence or phrases under or around the picture, and then click a speaker icon on the upper left corner of the screen to listen to an audio file that contains the eliciting utterance. Each eliciting utterance is preceded by a chiming sound to alert participants to listen. After the audio presentation of the eliciting utterance, the participant would hear an instruction in Chinese *qing huida* 'please answer'. Participants are then required to make a decision about whether the sentence they have just heard matches the contextual information on the screen, by selecting an option of "Match," "Mismatch" or "I do not know" on an answer sheet provided. This serves as a comprehension task to draw participants' attention to the meaning rather than the form of the eliciting utterance. This also provides a way to measure participants' comprehension of the utterance. Obviously, without correct comprehension, it would be difficult for the participant to recall the utterance. These procedures are also to ensure that there will be a time interval of at least 3 s between the presentation of the eliciting utterance and the start of the recalling. All this helps to make sure that the utterance produced by the participant is reconstructive, "requiring participants to process, rather than repeat verbatim, language stimuli" (Erlam, 2009, p. 488). Participants are then required to orally recall the utterance they have heard in Chinese immediately, which is to force participants to perform the recalling with time pressure instead of being self-paced, and to ensure that participants have little time to plan or monitor their responses.

This design is adapted from the methods used in Erlam's (2006, 2009) and Chrabaszcz and Jiang's (2014) studies. The rationale behind the elicited imitation task is the requirement for a participant to "decode the sentence they hear through syntactic and semantic parsing, retain the meaning, and reconstruct the sentence for subsequent production" (Chrabaszcz and Jiang, 2014, p. 359).

5 An acceptability judgement task and a picture-matching judgment task were also conducted, but the results of these two tasks are not to be reported here as they are irrelevant to the research questions discussed in the current paper.

这是小明和小丽明天的计划。
This is Xiaoming and Xiaoli's plan for tomorrow.

FIGURE 2
Picture and contextual information for (11).

这是小明和小丽明天的计划。
This is Xiaoming and Xiaoli's plan for tomorrow.

FIGURE 3
Picture and contextual information for (12).

Materials

There are 70 sentences in audio files during the experiment, out of which 12 are related to the investigation of L2 Chinese vP-ellipsis and ΣP-ellipsis, whilst the rest serve as distracters in the experiment. The 12 sentences are in two conditions, i.e., vP- and ΣP-ellipsis conditions (as illustrated in (11) and (12)), with each condition having six test sentences. Each test sentence consists of three clauses, and each sentence contains 22 or 23 Chinese characters. The pictures and contextual information for (11) and (12) are provided in Figures 2, 3 respectively. In the experiment, the contextual information only uses Chinese characters, and the English translation in Figures 2, 3 is provided for readers of this article.

11. Example of the vP-ellipsis condition:
Mingtian Xiaoming hui qu Beijing, Xiaoli ye **hui**, tamen hui yiqi qu.
Tomorrow Xiaoming will go Beijing Xiaoli also will they will together go.
kan Tian'an Men.
see Tian'an Men.
'Xiaoming will go to Beijing tomorrow, and Xiaoli will too. They will go to see Tian'an Men together.'

12. Example of the ΣP-ellipsis condition:
Mingtian Xiaoming hui qu Lundun, Xiaoli ye shi. Tamen dou shi diyi-ci.
tomorrow Xiaoming will go London Xiaoli also BE they both be first-CL.
chuguo lvxing.
go abroad travel.
'Xiaoming will go to London tomorrow, and Xiaoli will too. This is their first time to travel abroad.'

Categorizing and scoring participants' responses

Participants' responses in the elicited imitation task were recorded, transcribed and analysed. In the data trimming process, unclear responses, responses not immediately produced, and responses where the second clause is not produced or largely incomplete were removed and treated as missing values. Note that the choice of incorrect names (e.g., *Xiao Wei* instead of *Xiao Li*) and replacement of verbs or nouns with synonyms, were not treated as incorrect responses, as this kind of mistake does not affect the use of ellipsis. We categorized the responses into four categories, as described in (13).

13. Categorization of responses:

a. Responses with vP-ellipsis, where the second clause of the sentence does not have a phonetically realized verb phrase following an auxiliary;
b. Responses with ΣP-ellipsis, where the second clause of the sentence does not have a phonetically realized model verb phrase following *shi* 'BE';
c. Responses with object ellipsis, where the second clause of the sentence has a transitive verb, but does not have a phonetically realized object;
d. Responses with no ellipsis, where no vP, or ΣP or object in the second clause is elided.

Next, we gave 1 to each response representing one of the categories described above. The score and percentage of each response category were calculated for each participant group, respectively. Recall that six test sentences are contained in each condition, and each group contains 15 participants. A group's maximum the accumulated score for a response category in a condition is 90 and the minimum is 0.

Results

Results of the comprehension test

Recall that a comprehension test is included in the elicited imitation task, which requires participants to choose an option of "Match," "Mismatch" or "I do not know" on the answer sheet. Their comprehension of the sentences is checked by analysing the accuracy of their Match/Mismatch choices. Participants who correctly judged more than 63 items out of 70 (i.e., accuracy above 90%) were considered to have attended to meanings of the eliciting sentences. The results show that all groups' accuracy rates are higher than 93%, indicating that they have good comprehensions of the eliciting sentences in the task.

Data analysis

Native speakers of Chinese

As can be seen in Table 4, in both vP-ellipsis and ΣP-ellipsis conditions, native Chinese speakers produce an overwhelming number of target structures in their responses. Specifically, their vP-ellipsis responses account for 79% of the responses in the vP-ellipsis condition and ΣP-ellipsis responses 69% of the responses in the ΣP-ellipsis condition, whist non-target responses in both conditions are all under 15%. As the present study focuses on the L1 influence on English and Korean speakers' oral production of sentences with vP or ΣP-ellipsis, the native speakers' data in the study only serves as the baseline for response choices and will not be discussed further.

L2 groups

As shown in Table 4, great variations can be found in L2 groups' responses. In the vP-ellipsis condition, the EB Group behave similarly to the KB Group, producing very few target responses (the EB Group: 8% and the KB Group: 2%), even though vP-ellipsis is allowed in English; instead, both beginner groups produce a large proportion of responses with overt vP (the EB Group: 88% and the KB Group: 93%) in spite of the fact that no overt vP is included in the eliciting utterance. This provides us with evidence that no L1 transfer takes place at beginning levels of L2 Chinese oral production of vP-ellipsis. As their Chinese proficiency improves, the L2 learners produce increasingly more target responses with vP-ellipsis (the EI Group: 42%, the EA Group: 60%, the KI Group:15%, and the KA Group: 43%); at the same time, the frequencies of responses with non-ellipsis dramatically decrease (the EI Group: 41%, the EA Group: 14%, the KI Group: 48%, and the KA Group: 21%).

In the ΣP-ellipsis condition, the two beginner groups again behave similarly; they rarely produce target responses with ΣP-ellipsis (the EB Group: 0% and the KB Group: 1%) even though ΣP-ellipsis is allowed in Korean. In contrast, they produce responses with overt ΣP at very high rates (the EB Group: 90% and the KB Group: 96%), in spite of that fact that the eliciting utterances contain ΣP-ellipsis. Again, absence of L1 transfer is observed in beginners' oral production of ΣP-ellipsis. With the increase of their Chinese proficiency at intermediate and advanced levels, L1 Korean groups produce an increasingly higher proportions of target responses with ΣP-ellipsis (the KI Group: 41%, and the KA Group: 50%) than L1 English groups (the EI Group: 6%, and the EA Group: 17%).

The number of L2 learners' target responses were submitted to a linear mixed-effect models under the lme4 package in R version 4.1.0 (R Development Core Team, 2021). The fixed predictors include Proficiency (categorical factor, sum coded: beginner = −1, intermediate = 0, and advanced = 1), L1 (categorical factor, sum coded: English = −1 and Korean = 1) and Condition (categorical factor, sum coded: vP-ellipsis = −1 and ΣP-ellipsis = 1), and the interactions of Proficiency * L1, Proficiency * Condition, Condition * L1, and Condition * L1 * Proficiency. Participant and test items were entered as random factors for intercepts and slopes. A maximal model was first established, based on which the optimal model was found by backword elimination procedure. The formula of the optimal model is Score ~ Condition x L1 x Proficiency + (1 + L1 + Condition | Participant) + (1 + Proficiency + L1 | Item).

TABLE 4 The number of each response category in vP-ellipsis and ΣP-ellipsis conditions across groups (percentages in parentheses).

	vP-ellipsis Condition				ΣP-ellipsis Condition			
	Target	Non-Target			Target	Non-Target		
	vP-E	Non-E	Object-E	ΣP-E	ΣP-E	Non-E	Object-E	vP-E
EB	7 (8%)	79 (88%)	4 (4%)	0 (0%)	0 (0%)	81 (90%)	6 (7%)	3 (3%)
EI	40 (42%)	39 (41%)	17 (18%)	0 (0%)	6 (6%)	42 (44%)	23 (24%)	25 (26%)
EA	50 (60%)	12 (14%)	20 (24%)	2 (2%)	14 (17%)	14 (17%)	23 (28%)	32 (38%)
KB	2 (2%)	78 (93%)	4 (5%)	0 (0%)	1 (1%)	81 (96%)	2 (2%)	0 (0%)
KI	14 (15%)	46 (48%)	14 (15%)	22 (23%)	39 (41%)	47 (49%)	8 (8%)	2 (2%)
KA	39 (43%)	19 (21%)	15 (17%)	17 (19%)	45 (50%)	17 (19%)	13 (14%)	15 (17%)
NS	71 (79%)	7 (8%)	10 (11%)	2 (2%)	62 (69%)	7 (8%)	8 (9%)	13 (14%)

EB, English Beginner Group; EI, English Intermediate Group; EA, English Advanced Group; KB, Korean Beginner Group; KI, Korean Intermediate Group; KA, Korean Advanced Group; NS, Native Speaker Group; vP-E, responses with vP-ellipsis; Non-E, responses with no ellipsis; Object-E, responses with object ellipsis; ΣP-E, responses with ΣP-ellipsis.

The model output is presented in Table 5 and Figure 4. The results reveal a significant three-way Condition * L1 * Proficiency interaction, which indicates that the interaction of Condition and L1 differed across three proficiency levels. Specifically, from Figure 4A, it can be observed that the score is very close between ΣP-ellipsis condition and vP-ellipsis condition, although the score is slightly higher for vP-ellipsis condition than ΣP-ellipsis condition. The statistical data in Table 5 confirm that the effects of Condition is non-significant ($p = 0.149$). Similarly, Figure 4B shows that the score of English learners of Chinese is only slightly lower than that of Korean learners of Chinese, and the data in Table 5 reveal that the effect of L1 is non-significant ($p = 0.444$). In contrast, Figure 4D shows that the difference in the score between English and Korean learners is clearly different in ΣP-ellipsis condition than it is in vP-ellipsis condition (one difference is positive, the other negative), and this significant difference is confirmed by the statistical result of interaction between Condition and L1 ($p < 0.001$). Consequently, there is no overall effect of either L1 or condition, but there is a crossover interaction. From Figure 4C, it can be observed that the score is proportionate with proficiency, and the statistical data in Table 5 reveal that the effect of Proficiency is significant ($p < 0.001$). This indicates that the number of target responses varied across different proficiency groups. Figure 4E shows that the effect of proficiency is similar in

TABLE 5 Summary of the linear mixed-effect models for target responses.

	Score				
Predictors	**Estimates**	**std. Error**	**CI**	**Statistic**	***p***
(Intercept)	0.24	0.03	0.18–0.30	7.92	<0.001
Condition	−0.09	0.06	−0.21 – 0.03	−1.44	0.149
L1	0.03	0.04	−0.05 – 0.12	0.77	0.444
Proficiency	0.40	0.06	0.27–0.52	6.39	<0.001
Condition * L1	0.40	0.09	0.22–0.58	4.31	<0.001
Condition * Proficiency	−0.14	0.13	−0.39 – 0.11	−1.08	0.279
L1 * Proficiency	0.11	0.07	−0.03 – 0.24	1.52	0.128
Condition * L1 * Proficiency	0.43	0.16	0.12–0.73	2.75	0.006
Observations			1,080		

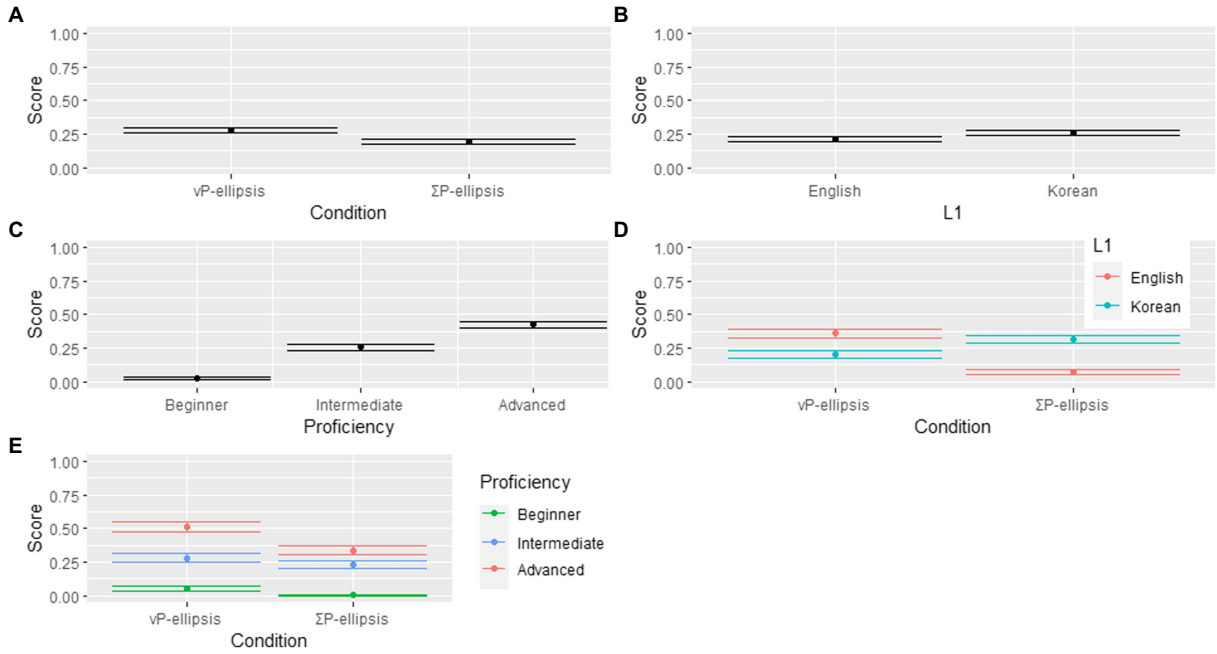


FIGURE 4
Score as a function of Condition, Proficiency, and L1. **(A)** Score as a function of Condition. **(B)** Score as a function of L1. **(C)** Score as a function of Proficiency. **(D)** Score as a function of Condition by L1. **(E)** Score as a function of Condition by Proficiency.

ΣP-ellipsis condition and vP-ellipsis condition. This echoed the patten in Figure 4E that the difference among three proficiency groups' scores across the two conditions are very similar, and thus the effect of Condition and Proficiency is not significant ($p = 0.279$).

In order to explore difference between target responses of proficiency-matched L1 English and L1 Korean groups, thereby answering the questions about the occurrence and disappearance of L1 influence at different L2 stages, Turkey *post hoc* comparisons were conducted through the emmeans package (Lenth, 2020) in a simplified model using dummy coding (formula: Score ~ Condition × L1 × Proficiency + (1 + L1 | Participant) + (1 + Proficiency + L1 | Item)). To explore the answer to the first research question about the role of L1 in L2 Chinese beginners' oral production, beginner learner groups' results were examined. The results suggest that there is no L1-related difference between L1 English and L1 Korean beginner learners' L2 Chinese production of utterances with vP- and ΣP-ellipsis (vP-ellipsis condition, EB vs. KB: $\beta\hat{} = 0.0518$, $SE = 0.0797$, $t = 0.649$, $p > 0.05$; ΣP-ellipsis condition, EB vs. KB: $\beta\hat{} = -0.0098$, $SE = 0.0798$, $t = -0.123$, $p > 0.05$). This suggests that L1 English beginners do not produce more vP-ellipsis sentences in their L2 Chinese speech production than L1 Korean beginners, and L1 Korean beginners do not produce more ΣP-ellipsis sentences in their L2 Chinese speech production than L1 English beginners.

To explore the answer to the second research question about the role of L1 in L2 developmental stages, intermediate and advanced learner groups' results were examined. In contrast, L1 difference is found to be a significant factor in L1 English and L1 Korean intermediate learners' L2 Chinese production of vP- and ΣP-ellipsis (vP-ellipsis condition, EI vs. KI: $\beta\hat{} = 0.2692$, $SE = 0.0779$, $t = 3.458$, $p < 0.05$; ΣP-ellipsis condition, EI vs. KI: $\beta\hat{} = -0.3421$, $SE = 0.0779$, $t = -4.394$, $p < 0.01$), indicating that L1 influence is absent at the beginner levels, but occurs at the intermediate level. At the advanced level, L1 English and L1 Korean groups' frequencies of target responses in the ΣP-ellipsis condition differ significantly from each other (EA vs. KA: $\beta\hat{} = -0.3318$, $SE = 0.0801$, $t = 4.141$, $p < 0.01$), indicating the persistence of L1 influence in the two advanced groups' L2 Chinese oral production of ΣP-ellipsis. In contrast, the advanced groups' frequencies of target responses in the vP-ellipsis condition are not significantly different (EA vs. KA: $\beta\hat{} = 0.1603$, $SE = 0.0801$, $t = 2.001$, $p > 0.05$), suggesting disappearance of L1 influence in the two advanced groups' L2 Chinese oral production of vP-ellipsis.

Discussion

As the present study aims to investigate the role of L1 in L2 speech production of vP- and ΣP-ellipsis at different stages of L2 Chinese development, this section is to discuss findings of the investigation with answers to the research questions concerning the role of L1 in different developmental stages of L2 Chinese vP- and ΣP-ellipsis.

Answers to research questions

Research Question 1. Is English- and Korean-speaking L2 Chinese beginners' oral production of vP- and ΣP-ellipses influenced by their L1s?

The results suggest that L1 influence is absent at beginner levels, which provides us with evidence against the FT Hypothesis (FTFA, Schwartz and Sprouse, 1994, 1996), which argues that the L2 initial state is entirely based on the final state of learners' L1. In our study, both L1 English and L1 Korean beginner learners of Chinese produce few utterances with vP- or ΣP-ellipsis even though the former is allowed in English and the latter in Korean; instead, they have overwhelming production of utterances with overt vP and ΣP in spite of the fact that the eliciting sentences contain vP- and ΣP-ellipsis. In contrast, L1 influence is found in the two intermediate groups' L2 Chinese production of utterances with vP- and ΣP-ellipsis.

Research Question 2. To what extent does L1 play a role in L2 Chinese oral production of vP- and ΣP-ellipsis at different stages of the L2 Chinese development? Does L1 influence persist or disappear at advanced levels? Specifically, do L1 English and L1 Korean advanced learners of L2 Chinese behave similarly in their oral production of sentences with vP- or ΣP-ellipsis?

The results indicate an asymmetry in the persistence of L1 influence with regard to vP-ellipsis and ΣP-ellipsis at advanced learners' L2 Chinese oral production. Specifically, the results reveal that at advanced levels, the difference between the L1 English and L1 Korean learners in producing utterances with vP-ellipsis disappears, as no significant difference is found between the frequencies of vP-ellipsis utterances in the EA and KA Groups. This is in contrast to the ΣP-ellipsis condition, where the EA Group still produce significantly fewer ΣP-ellipsis utterances than the KA Group, suggesting that the L1 influence concerning the ΣP-ellipsis continues to exist at advanced levels.

Absence of L1 influence at beginner levels of L2 Chinese oral production

It seems rather unexpected that L1 influence is absent in L1 English and L1 Korean beginners' L2 Chinese oral production of vP-ellipsis and ΣP-ellipsis, given the Full Transfer (FT) Hypothesis (FTFA, Schwartz and Sprouse, 1994, 1996), which proposes that the initial state of the L2 grammar is entirely based on the final state of learners' L1 grammar. Assuming that L1 English and L1 Korean beginners in our study are representatives of initial states of L2 Chinese grammars, it would be predicted, on the basis of FT Hypothesis, that beginner learners of L2 Chinese whose L1 disallows a certain target language structure would lag behind those whose L1 has an equivalent of it. However, neither L1 English

beginners have any advantage over L1 Korean beginners in their oral production of vP ellipsis, nor L1 Korean beginners have any advantage over L1 English beginners in ΣP-ellipsis, even though vP-ellipsis is allowed in English but disallowed in Koran and ΣP-ellipsis is allowed in Korean but disallowed English. The two groups behave similarly in our study and neither of the groups produce any substantial number of utterances with vP-ellipsis and ΣP-ellipsis in spite of the vP-ellipsis and ΣP-ellipsis in eliciting sentences in the study. These findings do not support the FT Hypothesis. According to this hypothesis, any failure in accommodating target language input will trigger restructuring of the L2 grammar. If L2 initial states are entirely based learners' L1 grammar, the abundant evidence of ΣP-ellipsis and vP-ellipsis in the Chinese input is expected to trigger changes to learners' L2 Chinese grammars so that ΣP-ellipsis and vP-ellipsis can be accommodated. However, no change seems to occur in either L1 English or L1 Korean beginners' L2 Chinese grammars. One may wonder whether the absence of L1 transfer is due to the beginner learners' difficulty with the basic sentence structures or vocabulary involved in the study. Recall that the participants' high accuracy in the comprehension task reported in Section "Results of the comprehension test". Above suggests that learners in all groups have no problem understanding the sentences involved in the study. More importantly, both L1 English and L1 Korean beginner groups produce an overwhelming number of "complete" utterances with no ellipsis in the experiment, which suggests their mastery of the underlying structures involved in the study. These facts indicate that the basic sentence structures and vocabulary involved in the study are available in beginner learners' L2 Chinese.

Recall that in comparison with utterances with vP- or ΣP-ellipsis, "complete" sentences with no ellipsis are grammatical but stylistically heavy and unconcise in Chinese. It seems likely that no syntax-stylistics interface is established in beginner learners' L2 Chinese, leading to a breakdown at a syntax-stylistics interface in their handling of sentences with vP- or ΣP-ellipsis. As beginner learners have limited L2 knowledge and unsophisticated coordination between information from different cognitive domains, such as syntax and stylistics, the mechanisms for their L2 oral production tend to be geared for syntactic "completeness" and are unlikely to be susceptible to any stylistic requirement, even though the syntax-stylistics interface is available in their L1s. As a result, this insensitivity to stylistic requirements at L2 initial stages leads to the absence of vP- and ΣP-ellipsis observed in L1 English and L1 Korean beginners' L2 Chinese oral production. That is, L2 learners' production at the beginner level is governed exclusively by basic essential syntactic computations and it is immune to stylistic requirements, which overrides L1 transfer of the syntax-stylistics interface from their L1s to their L2 speech production, leading to the absence of the influence of their L1 vP- or ΣP-ellipsis on their L2 Chinese oral production of utterances with vP- and ΣP-ellipsis in the current study.

If the analysis above is on the right track, it is reasonable to assume that, with improved L2 Chinese language proficiency and their increased automaticity in L2 Chinese oral production, and with more exposure to vP- and ΣP-ellipsis in their L2 Chinese input, they are more likely to produce utterances with vP- and ΣP-ellipsis, and this tendency is indeed observed in our intermediate and advanced learners' data. At the same time, L1 influences are detected at intermediate and advanced levels as well, where English speakers are found to produce more utterances with vP-ellipsis in their L2 Chinese oral production than Korean speakers, and in contrast, Korean speakers produce more utterances with ΣP-ellipsis than English speakers. This finding is on a par with what is reported in Zhang (2020), who focuses on English- and Korean-speaking learners of L2 Chinese at intermediate levels only, which is why the absence of L1 influence is observed in the current study and in Yuan and Zhang (2020), but not in Zhang (2020), where no L2 Chinese speakers at beginner levels are involved. Anyway, we argue that the different behaviours in their L2 Chinese oral production between English and Korean speakers at intermediate and advanced levels are a manifestation of what is allowed and disallowed in their respective L1s. That is, their oral production of these syntactically complicated but stylistically concise utterances is facilitated by the availability of the syntax-stylistic interface in their respective L1s, English and Korean.

The absence of L1 influence at beginner levels found in the present study is also in conformity with the finding concerning beginner learners' L2 Chinese oral production of utterances with object-ellipsis in Yuan and Zhang (2020), where an incremental model is adapted from Bock and Levelt (1994) for the findings in their study. The model is designed for the planning of speech production, and is assumed to have four stages: (a) lexical selection; (b) functional assignment; (c) constituent assembly; and (d) derivation, which includes checking and deleting. According to this model, the grammatical coding and operations are expected to be implemented before a sentence is phonetically spelt out. On the basis of the incremental model, Yuan and Zhang believe that beginner learners of L2 Chinese in their study have no problem with the first three stages. That is, they are able to select lexical items from their mental lexicon for the meaning to be expressed; they can assign grammatical functions, such as subject, object, etc., to the lexical items selected from their mental lexicon, and they are also able to assemble the lexical items in a word order appropriate to the target language. However, what they are unable to do at beginner levels is implement the derivation, such as checking and deleting, which require additional operations and are therefore more costly and taxing. Although the absence of L1 influence is not specifically addressed in Yuan and Zhang (2020), it seems possible to use their analysis to account for English and Korean beginners' overwhelming production of L2 Chinese utterance with no vP- and ΣP-ellipsis. We can argue that L2 beginners in our study also encounter problems in dealing with operations at the derivation stage. Production of utterances with vP-ellipsis or ΣP-ellipsis requires verbal

identity checking of whether the vP or ΣP in the second sentence is identical to that in the first coordinate sentence before the vP or ΣP in the second sentence is elided. If beginner learners are unable to implement the operations of the required checking and deleting, this would naturally lead to the overproduction of utterances of non-ellipsis and rare production of utterances with vP- and ΣP-ellipsis in their L2 Chinese oral production. Furthermore, the interface issue discussed above is likely to insert an additional layer of complication in the model. That is, as Chinese vP- and ΣP-ellipses operate at a syntax-stylistic interface, L2 Chinese speakers are unable to simultaneously handle, among other things, information from different sources at beginners' levels, which further reduces their ability to rely on their L1s in their oral production of vP- and ΣP-ellipses in their L2 Chinese oral production.

It should be acknowledged that with the data from and the design of our experiment, we are not in a position to pinpoint whether and to what extent the breakdown at the syntax-stylistics interface or the operations of checking and deleting is the main reason behind the absences of L1 transfer in beginner learners' L2 Chinese oral production of utterances with vP- and ΣP-ellipsis. It is likely to be a joint effect of the two. We have to leave this issue for future research.

Asymmetry in disappearance of L1 influence

Another interesting finding in the current study is that L1 influence disappears earlier in L2 Chinese production of vP-ellipsis than ΣP-ellipsis. The two advanced groups perform similarly in their Chinese oral production of vP-ellipsis, but the Korean group seems to continue to have the advantage of the ΣP-ellipsis in their L1 Korean and produce significantly more L2 Chinese utterances with ΣP-ellipsis than English speakers. In order to explain the asymmetry, it seems necessary to take into account differences in structural complexity between vP-ellipsis and ΣP-ellipsis. As discussed in "Cross-linguistic differences of verb-phrase ellipsis in Chinese, Korean and English", in comparison to a vP, a ΣP involves an ellipsis of a bigger constituent and its scope can include a negator, a model verb as well as a vP, as shown in Chinese sentences like (3) in "Cross-linguistic differences of verb-phrase ellipsis in Chinese, Korean and English". We believe that what is elided can be measured on the basis of computational complexity involved, which, in turn, can affect early or late disappearance of L1 influence. The asymmetry in the disappearance of L1 influence in speech production of vP- and ΣP-ellipsis at advanced levels can be accounted for with the help of the measurement of computation complexity in feature checking, as in Yuan (2015), who proposes that "Feature checking of α gives rise to a less complex computation than feature checking of α+ β" (Yuan, 2015, p.8). We can adapt his proposal and apply it to the analysis of the late disappearance of L1 influence on ΣP-ellipsis in English speakers' L2 Chinese, by assuming that verb identity checking of α alone gives rise to a less complex computation than verb identity checking of α+β and that deleting only α gives rise to a less complex computation than deleting α+β.

According to this metric, the more items a verbal identity checking operation involves, the more computational complexity the operation has. L2 structures with less computational complexity are expected to be acquired more easily than those with more computational complexity. In the current case, the operations on the vP-ellipsis are computationally less complex than the ΣP-ellipsis, because the former involves identity checking and deleting of only a vP and an NP object, but the latter requires identity checking and deleting of not only a vP and an NP object, but also a model and a negator. Although Korean does not have the vP-ellipsis, L1 Korean learners are able to overcome the disadvantage of not having vP-ellipsis in their L1 and acquire the less complex vP-ellipsis construction at a relatively early stage in their L2 Chinese acquisition. In contrast, L1 English learners do not allow the ΣP-ellipsis in their L1. In addition, the ΣP-ellipsis in the target language Chinese is computationally more complex than the vP-ellipsis, requiring identity checking and deleting of more items than vP-ellipsis, and as a result, their L1 English grammar is relied upon more when Chinese sentences with ΣP-ellipsis is dealt with, delaying the disappearance of L1 influence in their L2 Chinese oral production of utterances with ΣP-ellipsis. This explains why L1 influence is shorter-lived in L2 Chinese oral production of vP-ellipsis than ΣP-ellipsis.[6]

Conclusion

The current study tracks the role of L1 in L2 speech production of Chinese verb phrase-ellipsis structures at different stages of L2 development. One finding of the study is the absence of L1 influence on L2 Chinese speech production until intermediate and advanced levels. Both L1 English and L1 Korean learners of L2 Chinese at beginner levels tend to produce complete responses with no ellipsis, in spite of the fact that vP-ellipsis is allowed in English and that ΣP-ellipsis in Korean. This finding provides us with evidence against the FT Hypothesis proposed by (FTFA, Schwartz and Sprouse, 1994, 1996), which proposes that the initial state of

6 This analysis seems to tie in well with the finding in Yuan and Zhang (2020) that no L1 influence is observed in L2 Chinese production of object-ellipsis by L1 English and L1 Korean learners. Identity checking and deleting of just the NP object is computationally even less complex than vP- and ΣP-ellipsis. Consequently, the L1 influence involved is expected to disappear even earlier. The reason that no L1 influence is observed in L2 Chinese production of object-ellipsis is probably because the L1 influence involved is too short-lived to be caught in Yuan and Zhang's (2020) study.

the L2 grammar is entirely based on the final state of learners' L1 grammar. At intermediate levels, English- and Korean-speaking learners produce significantly more utterances with the type of ellipsis allowed in their L1s. The different behaviours between L2 learners at beginner and intermediate levels are attributed to a breakdown at the syntax-stylistics interface and to the difficulty caused by the identity checking and deleting operations involved in the derivation stage in beginner learners' L2 Chinese speech production. L2 learners at beginner levels are believed to strive for syntactic completeness and derivational simplicity before implementing syntactic approaches to stylistic modification, which overrides L1 transfer in beginners' L2 Chinese speech production. Another finding is the difference in the persistence of L1 influence on the two types of ellipsis in English and Korean speakers' L2 Chinese oral production; with regard to vP-ellipsis, L1 influence can be caught at intermediate levels but disappears at advanced levels; with regard to ΣP-ellipsis, L1 influence seems to be longer-lived, as it continues to exist at advanced levels. This is accounted for on the basis of a modified version of the computational complexity metric in Yuan (2015). Based on the finding in the current study, we argue that L1 influence should be considered a relative phenomenon in L2 speech production, and its presence and absence can be related to a number of factors, including learners' ability in handling information from different cognitive domains at interface levels, the availability of operations at derivation stages in their L2, the computational complexity involved, etc. Of course, it deserves further research as to which of these factors plays a more important or decisive role in L1 influence in L2 oral production.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

LZ did the data collection and statistical analyses. BY made substantial contributions to the theoretical framework for the findings of the study. LZ and BY made equal contributions to the design of the experiment, data interpretation, and the writing of the manuscript. Both authors contributed to the article and approved the submitted version.

Funding

The work was supported by Shanghai Planning Office of Philosophy and Social Science (Grant No. 2022EYY009).

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

References

Adger, D. (2003). *Core Syntax: A Minimalist Approach*. Oxford Oxford University Press.

Arregui, A., Clifton, C., Frazier, L., and Moulton, K. (2006). Processing elided verb phrases with flawed antecedents: the recycling hypothesis. *J. Mem. Lang.* 55, 232–246. doi: 10.1016/j.jml.2006.02.005

Bock, J. K. (1986). Syntactic persistence in language production. *Cogn. Psychol.* 18, 355–387. doi: 10.1016/0010-0285(86)90004-6

Bock, J. K., and Levelt, W. J. M. (1994). "Language production: grammatical encoding," in *Handbook of Psycholinguistics*. ed. M. A. Gernsbacher (San Diego, CA: Academic Press)

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT Press.

Chrabaszcz, A., and Jiang, N. (2014). The role of the native language in the use of the English nongeneric definite article by L2 learners: a cross-linguistic comparison. *Second. Lang. Res.* 30, 351–379. doi: 10.1177/0267658313493432

Chung, S. (2013). Syntactic identity in sluicing: how much and why. *Linguis. Inq.* 44, 1–44. doi: 10.1162/LING_a_00118

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: an empirical validation study. *Appl. Linguis.* 27, 464–491. doi: 10.1093/applin/aml001

Erlam, R. (2009). "The elicited oral imitation test as a measure of implicit knowledge" in *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. eds. R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp and H. Reinders (Bristol: Multilingual Matters)

Flett, S., Branigan, H. P., and Pickering, M. J. (2013). Are non-native structural preferences affected by native language preferences? *Biling. Lang. Congn.* 16, 751–760. doi: 10.1017/S1366728912000594

Frazier, L. (2008). "Processing ellipsis: a processing solution to the undergeneration problem?" in *Proceedings of the 26th West Coast Conference on Formal Linguistics*. eds. C. B. Chang and H. J. Haynie (Somerville, MA: Cascadilla Proceedings Project), 21–32.

Gathercole, S. E., and Alloway, T. P. (2007). *Understanding working memory: A classroom guide*. Lontoo: Harcourt Assessment.

Gathercole, S. E., Alloway, T. P., Kirkwood, H. J., Elliott, J. G., Holmes, J., and Hilton, K. A. (2008). Attentional and executive function behaviours in children with poor working memory. *Learn. Individ. Differ.* 18, 214–223.

Gathercole, S. E., Pickering, S. J., Ambridge, B., and Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Dev. Psychol.* 40, 177–190.

Gass, S., and Selinker, L. (1992). *Language transfer in language learning*. Amsterdam: John Benjamins.

Grinder, J., and Postal, P. M. (1971). Missing antecedents. *Linguis. Inq.* 2, 269–312.

Hawkins, R. (2001). *Second Language Syntax: A Generative Introduction*. Oxford Wiley Blackwell.

Hawkins, R., and Casillas, G. (2008). Explaining frequency of verb morphology in early L2 speech'. *Lingua* 118, 595–612.

Haznedar, B. (2001). The acquisition of the IP system in child L2 English. *Stud. Second. Lang. Acquis.* 23, 1–39. doi: 10.1017/S0272263101001012

Hendriks, P. (2004). Coherence relations, ellipsis and contrastive topics. *J. Semant.* 21, 133–153. doi: 10.1093/jos/21.2.133

Hsieh, F. A. (2015). *Relative clause acquisition in second language Chinese and second Language English (Unpublished doctoral dissertation)*. University of Cambridge.

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguis. Inq.* 15, 531–574.

Inagaki, S. (2002). Japanese learners' acquisition of English manner-of-motion verbs with locational/directional PPs. *Second. Lang. Res.* 18, 3–27. doi: 10.1191/0267658302sr196oa

Johnson, K. (2001). "What VP ellipsis can do, and what it can't, but not why," in *The Handbook of Contemporary Syntactic Theory*. eds. M. Baltin and C. Collins (Oxford: Wiley Blackwell), 439–479.

Kehler, A. (2000). Coherence and the resolution of ellipsis. *Linguis. Philos.* 23, 533–575. doi: 10.1023/A:1005677819813

Kellerman, E., and Sharwood Smith, M. (eds.) (1986). *Crosslinguistic influence in second language acquisition*. New York: Pergamon.

Kertz, L. (2013). Verb phrase ellipsis: the view from information structure. *Language* 89, 390–428. doi: 10.1353/lan.2013.0051

Kim, J.-S., and Sohn, K.-W. (1998). Focusing effects in Korean/Japanese ellipsis. *Japan. Korean Linguis.* 8, 459–470.

Lasnik, H. (1981). "Restricting the theory of transformations: A case study," in *Explanations in Linguistics*. eds. N. Hornstein and D. Lightfoot (London: Longman), 152–173.

Lenth, R. (2020). *Emmeans: Estimated marginal means, aka least-squares means (Version 1.5.2-1) [R package]*.

Li, Y.-H. A. (2014). Born empty. *Lingua* 151, 43–68. doi: 10.1016/j.lingua.2013.10.013

Li, C., and Thompson, S. (1981). *Mandarin Chinese: A Functional Reference Grammar*. Berkeley University of California Press.

Li, Y.-H. A., and Wei, T. (2013). "Ellipsis" in *The New Handbook of CHINESE Linguistics*. eds. C.-T. J. Huang, Y. Li and A. Simpson (Oxford: Wiley Blackwell)

Liu, C. M. L. (2014). *A Modular Theory of Radical Pro drop*. Cambridge Harvard University.

Lobeck, A. (1995). *Ellipsis: Functional Heads, Licensing, and Identification*. Oxford: Oxford University Press.

Mai, Z., and Yuan, B. (2016). Uneven reassembly of tense, telicity and discourse features in L2 acquisition of the Chinese shì … de cleft construction by adult English speakers. *Second. Lang. Res.* 32, 247–276. doi: 10.1177/0267658315623323

Matsuo, A. (2007). Differing interpretations of empty categories in English and Japanese VP ellipsis contexts. *Lang. Acquis.* 14, 3–29. doi: 10.1080/10489220701331805

Matsuo, A., and Duffield, N. (2001). VP-ellipsis and anaphora in child language acquisition. *Lang. Acquis.* 9, 301–327. doi: 10.1207/S15327817LA0904_02

Merchant, J. (2001). *The Syntax of Silence: Sluicing, Islands, and the Theory of Ellipsis*. Oxford: Oxford University Press.

Merchant, J. (2004). Fragments and ellipsis. *Linguis. Philos.* 27, 661–738. doi: 10.1007/s10988-005-7378-3

Montrul, S. (2010). Dominant language transfer in adult second language learners and heritage speakers. *Second. Lang. Res.* 26, 293–327. doi: 10.1177/0267658310365768

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. New York: Cambridge University Press.

Odlin, T. (2003). "Crosslinguistic influence," in *Handbook of second language acquisition*. eds. C. Doughty and M. Long (Oxford: Blackwell), 436–486.

Prévost, P., Strik, N., and Tuller, L. (2014). Wh-questions in child L2 French: derivational complexity and its interactions with L1 properties, length of exposure, age of exposure, and the input. *Second. Lang. Res.* 30, 225–250. doi: 10.1177/0267658313519814

R Core Team (2021). *R: A language and environment for statistical computing [software]*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Scheidnes, M., and Tuller, L. (2010). "Syntactic movement in the production of French wh-questions: the role of computational complexity versus L1 transfer in adult L2 acquisition" in *Movement and Clitics*. eds. V. Torrens, L. Escobar, A. Gavarro and G. Juncal (Newcastle upon Tyne, UK: Cambridge Scholars Publishing)

Schwartz, B. D., and Sprouse, R. A. (1994). Word order and nominative case in non-native language acquisition: a longitudinal study of (L1 Turkish) German interlanguage. Language Acquisition Studies in Generative Grammar: Papers in Honor of Kenneth Wexler from the 1991 GLOW Workshops, 317–368. doi: 10.1075/lald.8.14sch

Schwartz, B. D., and Sprouse, R. A. (1996). L2 cognitive states and the full transfer/full access model. *Second. Lang. Res.* 12, 40–72. doi: 10.1177/026765839601200103

Slabakova, R. (2013). The effect of construction frequency and native transfer on second language knowledge of the syntax-discourse interface. *Appl. Psycholinguist.* 36, 671–699. doi: 10.1017/S0142716413000386

Slabakova, R. (2015). The effect of construction frequency and native transfer on second language knowledge of the syntax-discourse interface. *Appl. Psycholinguist.* 36, 671–699. doi: 10.1017/S0142716413000386

Soh, H. L. (2007). Ellipsis, last resort, and the dummy auxiliary shi "be" in mandarin Chinese. *Linguis. Inq.* 38, 178–188. doi: 10.1162/ling.2007.38.1.178

Wechsler, D. (2010). *Wechsler Adult Intelligence Scale UK 4th Edn*. London: Pearson.

Xu, L. (2003). Remarks on VP-ellipsis in disguise. *Linguis. Inq.* 34, 163–171. doi: 10.1162/ling.2003.34.1.163

Yuan, B. (2001). The status of thematic verbs in the second language acquisition of Chinese: against inevitability of thematic-verb raising in second language acquisition. *Second Lang. Res.* 17, 248–272.

Yuan, B. (2015). The effect of computational complexity on L1 transfer: evidence from L2 Chinese attitude-bearing wh-questions. *Lingua* 167, 1–18. doi: 10.1016/j.lingua.2015.09.001

Yuan, B., and Zhang, L. (2020). An incremental model of second language speech production mechanisms: developmental evidence from object ellipsis in second language Chinese speech production. *Int. J. Biling.* 24, 783–810. doi: 10.1177/1367006919886666

Zhang, L. (2020). The existence of L1 influence on the course of development of L2 representation: evidence from L2 acquisition of verb phrase ellipsis. *Lingua* 237, 102823–102829. doi: 10.1016/j.lingua.2020.102823

Zufferey, S., Mak, W., Degand, L., and Sanders, T. (2015). Advanced learners' comprehension of discourse connectives: the role of L1 transfer across on-line and off-line tasks. *Second. Lang. Res.* 31, 389–411. doi: 10.1177/0267658315573349

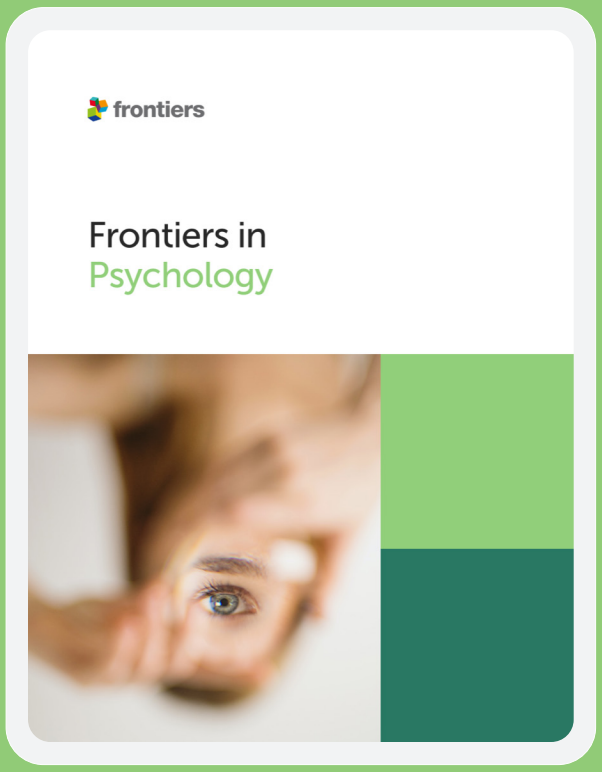
Frontiers in Psychology

Paving the way for a greater understanding of human behavior

The most cited journal in its field, exploring psychological sciences - from clinical research to cognitive science, from imaging studies to human factors, and from animal cognition to social psychology.

Discover the latest Research Topics

See more →



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact