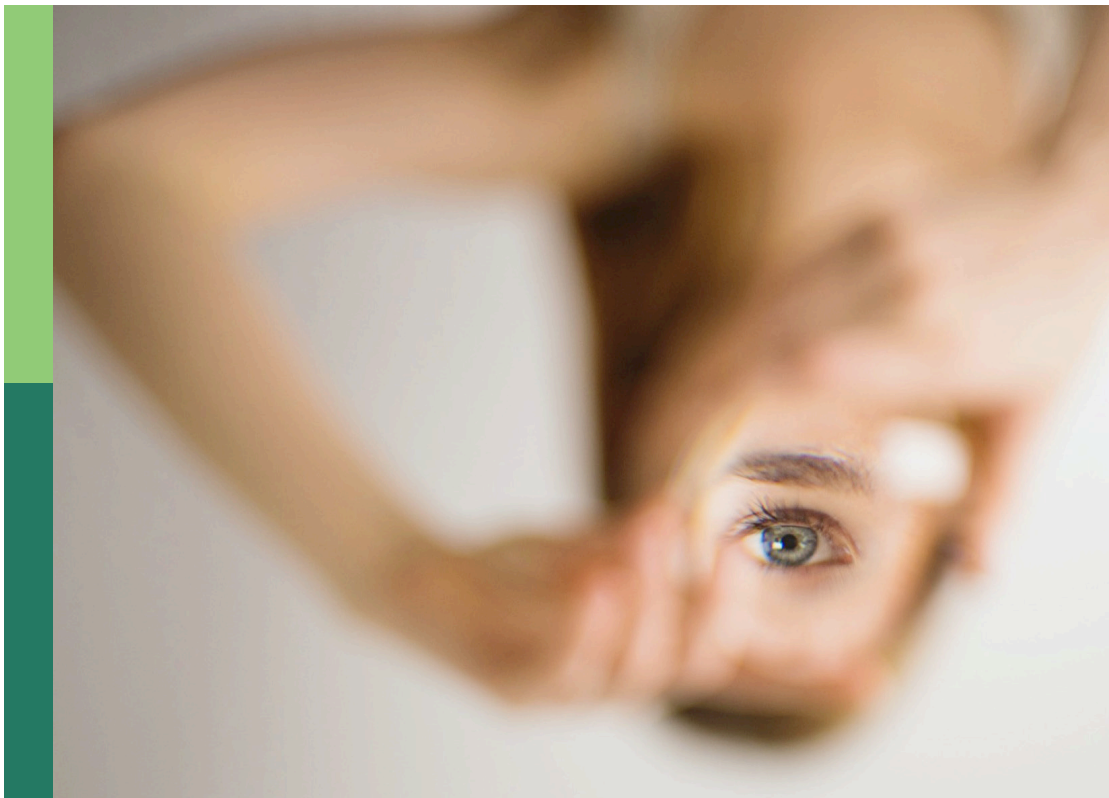
frontiers | Research Topics

# Second or foreign language learning and cognitive development

**Edited by**
Dingfang Shu, Jinfen Xu and Hui Zhang

**Coordinated by**
Zhen Tian

**Published in**
Frontiers in Psychology



ISSN 1664-8714
ISBN 978-2-8325-4217-0
DOI 10.3389/978-2-8325-4217-0

## 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

# Second or foreign language learning and cognitive development

**Topic editors**
Dingfang Shu — Shanghai International Studies University, China
Jinfen Xu — Huazhong University of Science and Technology, China
Hui Zhang — Nanjing Normal University, China

**Topic Coordinator**
Zhen Tian — Shanghai International Studies University, China

**Citation**
Shu, D., Xu, J., Zhang, H., Tian, Z., eds. (2024). *Second or foreign language learning and cognitive development*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-4217-0

# Table of contents

frontiers | Frontiers in Psychology

TYPE Editorial
PUBLISHED 04 January 2024
DOI 10.3389/fpsyg.2023.1354329



OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Dingfang Shu
✉ shudfk@yahoo.com

RECEIVED 12 December 2023
ACCEPTED 20 December 2023
PUBLISHED 04 January 2024

CITATION
Shu D, Xu J, Zhang H and Tian Z (2024) Editorial: Second or foreign language learning and cognitive development.
*Front. Psychol.* 14:1354329.
doi: 10.3389/fpsyg.2023.1354329



# Editorial: Second or foreign language learning and cognitive development

Dingfang Shu[1]*, Jinfen Xu[2], Hui Zhang[3] and Zhen Tian[1]

[1]Shanghai Center for Research in English Language Education, Shanghai International Studies University, Shanghai, China, [2]School of Foreign Languages, Huazhong University of Science and Technology, Wuhan, Hubei, China, [3]School of Foreign Languages, Nanjing Normal University, Nanjing, Jiangsu, China

KEYWORDS

**second language (L2) acquisition, foreign language (English) learning and teaching, cognitive development, language teaching, language learning**

Editorial on the Research Topic
Second or foreign language learning and cognitive development

Research on bilingualism has shown that acquiring a second language enhances a learner's executive function and metalinguistic awareness within the cognitive development domain (Bialystok, 2001; Bialystok and Luk, 2012; Kroll and Bialystok, 2013). Further investigation is necessary to understand the impact of individual differences on the learning process and its result, including the influencing factors like age, gender, gender, first language, learning style, input and feedback types, and teaching methods. Perhaps more crucially, however, is the potential for the learning process to interact with learners' metalinguistic, affective, cognitive, and metacognitive abilities (Dörnyei, 2009; Bylund and Jarvis, 2011; Gass and Mackey, 2015; VanPatten and Williams, 2015).

A total of 29 manuscripts were submitted on the Research Topic. A total of 15 have been accepted, which fall into five broad categories: (1) the interaction between L1 and L2; (2) second or foreign language learning and cognitive controls; (3) second or foreign language learning and social skills and empathy; (4) second or foreign language learning and metacognitive skills; and (5) teaching methodology and L2 and foreign language learning.

One paper in particular, "*Australian English listeners' perception of Japanese vowel length reveals underlying phonological knowledge*," by Yazawa et al., examines how native speakers of Australian English, who typically emphasize vowel length compared with most other English varieties, perceive Japanese vowel length contrasts. In a forced-choice study, twenty monolingual Australian English speakers were asked to rank the Japanese long and short vowels based on their resemblance to their native vowel categories. The findings indicated a general tendency for Australian English long and short vowels (such as/i:, I/as in "heed," "hid") to be classified as Japanese long and short vowels (e.g.,/ii, i/). This contrasts with the literature-reported categorization of all Japanese vowels as tense by American English listeners, regardless of length (e.g.,/ii, i/as both "heed"). The result is consistent with a feature-based speech perception approach.

Research on the shared-dialect effect, which suggests that raters who share a candidate's dialect may provide higher scores on English speaking examinations, was conducted by Xu et al.. Oral performance in the recounting task of the computer-based English Listening and Speaking Test was evaluated by raters proficient in Cantonese and Mandarin. No statistically significant interaction was found between the raters' and candidates' dialects, nor were there any significant variations in the ratings given by either group in the quantitative data.

The understanding and scoring process of raters were influenced by their awareness and familiarity with accents, according to the qualitative data.

Wang D. et al. investigate the efficacy and diversity of translation strategies employed by Chinese English as a foreign language (EFL) learners when addressing light verb constructions (LVCs), a significant distinction between Chinese and English, in a different study titled "Walking out of the light verb jungle: Exploring the translation strategies of light verb constructions in Chinese–English consecutive interpreting". The study examines the methods used by 66 Chinese EFL learners to interpret 12 target LVCs using a theory-driven, context-based interpreting problem. The outcomes demonstrate the typical structural trends in LVC translation as well as the overall preferences for strategy selection among Chinese EFL learners. Additionally, the study reveals a positive relationship between vocabulary knowledge and the acceptability rates of LVCs, indicating the necessity of integrating constructional teaching into EFL instruction.

The research paper titled "*Non-adjacent dependency learning from variable input: investigating the effects of bilingualism, phonological memory, and cognitive control*" by Verhagen and de Bree delves into L2 learning and cognitive control. It sheds new light on the correlation between bilingualism and statistical learning, and compares the effects of consistent and variable input on statistical learning in both monolingual and bilingual children and adults. The study also investigates whether phonological memory and cognitive control play a role in potential group differences. The results indicate that bilinguals have a limited advantage in statistical learning, which is not consistently linked to enhanced cognitive abilities associated with bilingualism.

In recent years, there has been a surge in research on the relationship between emotion and L2 learning, with a particular focus on social skills and empathy. One such study, "*Understanding foreign language writing anxiety and its correlates*" by Li, conducted a quantitative meta-analysis of 84 effect sizes from 22 primary studies to investigate the connections between foreign language writing anxiety and its high and low-evidence correlates. The study revealed moderate correlations between foreign language writing anxiety and writing self-efficacy and performance, as well as moderately positive effects with listening, speaking, and reading anxiety. Additionally, the study found that age and language proficiency have significant moderating effects. The findings have important pedagogical implications, which were discussed based on the results.

Wang H. et al.'s article titled "*Unpacking the relationships between emotions and achievement of EFL learners in China: Engagement as a mediator*" explores the connections between learners' emotions, such as foreign language enjoyment (FLE), foreign language classroom anxiety (FLCA), and foreign language learning boredom (FLLB), and engagement, as well as their English achievement. The study involved 907 English as a foreign language (EFL) learners from a university in China who completed an online questionnaire, and structural equation modeling was used to test the hypothesized relations among the variables. The results showed correlations between learners' FLE, FLCA, and FLLB, and that learners' engagement mediated the relationships between their emotions and English achievement. The study provides evidence for the mechanism underlying the relationships between emotions, engagement, and achievement, and sheds light on EFL teaching and learning at the tertiary level in China.

A third article that falls into this category, *Measuring Chinese English-as-a-foreign-language learners' resilience: development and validation of the foreign language learning resilience scale* by Guo and Li, aimed to develop the Foreign Language Learning Resilience Scale (FLLRS) to measure the psychometric scale reliability and validity of foreign language learning resilience in Chinese English-as-a-foreign-language contexts. Data was collected from 313 Chinese college students, and the FLLRS was validated based on reliability and validity tests. The FLLRS consisted of three factors: ego resilience, metacognitive resilience, and social resilience, all contributing to foreign language learning resilience. Metacognitive resilience had the highest path coefficient, followed by social resilience and ego resilience. The validated scale could advance knowledge in second language acquisition regarding the factors that affect foreign language learning resilience.

Second or foreign language learning and metacognitive skills have been of great interest among researchers. In a study by Qin et al., entitled "*Validation of metacognitive strategies in writing and their predictive effects on the writing performance of English as foreign language student writers*," the metacognitive writing strategies of EFL college students in China were examined through a survey and a writing test. The study utilized exploratory factor analysis and confirmatory factor analysis to analyze the data, and multiple regression analysis was employed to understand the predictive effects of metacognitive strategies on writing performance. The findings suggest that writing instruction can enhance students' awareness and ability to acquire metacognitive writing strategies, particularly those related to planning, monitoring, and evaluating.

Wang's study, "*Text memorization: an effective strategy to improve Chinese EFL learners' argumentative writing proficiency*," explored the impact of text memorization strategies on the argumentative writing proficiency of EFL learners in China. The study focused on the text memorization process and the strategies used by learners to enhance memorization. Thirty-three Chinese English majors participated in seven text memorization tests, a pre-test, and a post-test to evaluate their memorization outcomes and writing proficiency before and after memorizing seven model English writings. Additionally, twelve top scorers in the memorization tests were interviewed. The results indicated that text memorization significantly improved learners' writing proficiency. Moreover, a new system of text memorization strategies was developed to assist scholars and teachers in enhancing the writing skills of EFL learners.

Peng and Bao's article, "*Effects of reasoning demands triggered by genre on Chinese EFL learners' writing performance*," examined the impact of cognitive complexity on the writing performance of advanced Chinese EFL learners in two different genres: expository writing and argumentative writing. The study involved 76 EFL learners who completed two writing tasks with varying levels of reasoning demands. Multiple measure indices, including lexical complexity, syntactic complexity, accuracy, fluency, and cohesion, were used to assess the differences in production dimensions between the two tasks. The results indicated that

cognitive complexity significantly enhanced lexical complexity, clausal complexity, and cohesion, but there was a trade-off effect for phrasal and clausal structures within syntactic complexity. The findings of this study have important implications for the sequencing and design of L2 writing tasks.

The success of foreign students' academic and life skills in the Northern Cyprus region is heavily reliant on the importance given to Turkish language teaching. "*Teaching the Turkish language to foreigners at higher education level in northern Cyprus: an evaluation based on self-perceived dominant intelligence types, twenty-first century skills and learning technologies*" by Kurt and Güneyli aimed to investigate how college students use learning technology, 21st-century skills, and perceive intelligence categories in learning a foreign language. The study utilized purposeful and convenience sampling, selecting the institution with the largest number of international students in Northern Cyprus. The results indicated a statistically significant correlation between 21st-century skills and foreign language-learning technology usage, highlighting the importance of modern methodologies and social learning in foreign language education.

Zhao and Huang's article, "*A comparative study of frequency effect on acquisition of grammar and meaning of words between Chinese and foreign learners of English language*," investigated the impact of frequency on L2 vocabulary acquisition. The study explored the frequency effect on the acquisition of grammar and meaning of alphabetic words between Chinese learners of hieroglyphic language and foreign learners of alphabetic language. The results indicated that mother tongue type may not be the factor causing differences in grammar and meaning acquisition of vocabulary, while exposure frequency of vocabulary plays a determining role. Furthermore, learner types, language types, frequency, and part of speech of a word have an interaction effect on word acquisition. The study sheds light on the importance of frequency in L2 vocabulary acquisition and highlights the need for tailored teaching methods to facilitate this process.

Meng et al.'s article, "*Cognitive diagnostic assessment of EFL learners' listening barriers through incorrect responses*," utilized Cognitive Diagnostic Models (CDMs) for bugs, or Bug-CDMs, to diagnose EFL learners' listening barriers through incorrect responses. The study found that Bug-GDINA was the optimal model, and semantic understanding and vocabulary recognition were the most prevalent barriers. The findings demonstrate the feasibility of using Bug-GDINA to diagnose listening barriers from incorrect responses.

The majority of research on collocations in L2 acquisition and cognitive psychology has focused on phonographic languages, giving scant attention to ideographic languages such as Chinese and Japanese. In "*The lexical processing of Japanese collocations by Chinese Japanese-as-a-Foreign-Language learners: an experimental study by manipulating the presentation modality, semantic transparency, and translational congruency*," Song et al. investigated the processing of Japanese collocations by 36 Chinese Japanese-as-a-Foreign-Language learners. The study manipulated presentation modality, semantic transparency, and translational congruency in a lexical judgment task. The results indicated longer reaction times for auditory presentation than visual presentation, and longer reaction times for high semantic transparency and congruent translation in auditory presentation. These findings support the dual-route model of Japanese collocational processing and demonstrate that presentation modality, semantic transparency, and translational congruency have an impact on processing.

He and Gao's study, "*Explicating peer feedback quality and its impact on feedback implementation in EFL writing*," investigated the impact of peer feedback quality on EFL students' feedback implementation in argumentative writing tasks. The researchers developed a measuring scale with two dimensions to assess feedback quality, including accuracy and revision potential. The results indicated that feedback accuracy was at a medium level, while revision potential was at a low level, with accuracy having a stronger predictive power on implementation. Furthermore, feedback quality had the strongest predictive power when feedback features and focus were considered. The study highlights the importance of training students to provide and implement high-quality feedback marked by good accuracy and high revision potential in future instructions.

For future studies on second or foreign language learning and learners' cognitive development, it would be valuable to explore how the learning process, and learning multiple foreign languages, either simultaneously or consecutively, can impact learners' emotional and cognitive development, and how the new development, in turn, affects the learning process and outcome. This research can provide insights into the relationship between language learning and cognitive development, as well as the potential benefits of multilingualism. Additionally, it may be interesting to investigate how individual differences, such as age, gender, and learning styles, affect this relationship. Understanding the impact of language learning on cognitive development can inform language education and help educators design more effective language learning programs.

## Author contributions

DS: Writing—review & editing. JX: Writing—review & editing. HZ: Writing—review & editing. ZT: Writing—review & editing.

## Funding

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

## Conflict of interest

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

## Publisher's note

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

## References

Bialystok, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511605963

Bialystok, E., and Luk, G. (2012). The cognitive benefits of bilingualism. *Res. Lang. Soc. Inter.* 45, 108–118.

Bylund, E., and Jarvis, S. (2011). L2 effects on L1 event conceptualization. *Lang. Lear.* 61, 446–486. doi: 10.1017/S1366728910000180

Dörnyei, Z. (2009). *The Psychology of Second Language Acquisition*. Oxford: Oxford University Press.

Gass, S. M., and Mackey, A. (2015). "Input, interaction, and output in second language acquisition," in *The Routledge Handbook of Second Language Acquisition* (London: Routledge), 64–79.

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

VanPatten, B., and Williams, J. (2015). *Theories in Second Language Acquisition: An Introduction.* London: Routledge. doi: 10.4324/9780203628942

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 03 November 2022
DOI 10.3389/fpsyg.2022.1031514



OPEN ACCESS

EDITED BY
Hui Zhang,
Nanjing Normal University,
China

REVIEWED BY
Davoud Amini,
Azarbaijan Shahid Madani University, Iran
Eleniva Silaki,
University of Crete,
Greece

*CORRESPONDENCE
Rui Li
liruidianzi@hotmail.com

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

RECEIVED 30 August 2022
ACCEPTED 03 October 2022
PUBLISHED 03 November 2022

CITATION
Li R (2022) Understanding foreign language writing anxiety and its correlates. *Front. Psychol.* 13:1031514. doi: 10.3389/fpsyg.2022.1031514



# Understanding foreign language writing anxiety and its correlates

Rui Li*

School of Foreign Languages, Hunan University, Changsha, China

Despite the increasing number of empirical studies that investigated foreign language writing anxiety and its correlates, there is still a lack of quantitative meta-analytic attempt on the effect sizes among these studies. To bridge the gap, this study identified 84 effect sizes from 22 primary studies to meta-analyze the correlations of foreign language writing anxiety and several key high-and low-evidence correlates. For the two high-evidence correlates, moderator analyses were also conducted, which demonstrated that foreign language writing anxiety has a moderate correlation with foreign language writing self-efficacy and foreign language writing performance. The three low-evidence correlates have positively moderate effects of foreign language listening anxiety, foreign language speaking anxiety and foreign language reading anxiety. The significant moderating effects of learners' age and language proficiency were obtained. With respect to the results, pedagogical implications were discussed as well.

KEYWORDS

**correlate, foreign language, meta-analysis, writing anxiety, self-efficacy**

## Introduction

As one of the important productive language skills, writing skill receives considerable attention in second language acquisition (SLA) and evaluation (Rakedzon and Baram-Tsabari, 2017). However, due to second or foreign language writing anxiety/apprehension, language learners may encounter writing difficulties and feel cognitively and physiologically nervous when writing in a foreign language, as reflected from the decreased writing performance and negative writing affects (Abdel Latif, 2015, 2019; He, 2018; Russell-Pinson and Harris, 2019).

Foreign language writing anxiety is often defined as "the dysfunctional anxiety that many individuals suffer when confronted with foreign language writing tasks" (Cheng, 2002, p. 647). In other words, highly anxious learners are documented to achieve lower foreign language performance (Abdel Latif, 2015), poorer foreign language writing performance (Cheng, 2002, 2004), and lower foreign language writing affects, such as motivation (e.g., Alico, 2016; Tsao et al., 2017; Abdel Latif, 2019), writing self-efficacy (e.g., Cheng, 2004; Woodrow, 2011; Abdel Latif, 2019), writing attitude (Sarkhoush, 2013), and writing strategies (Wu and Lin, 2016; Tsiriotakis et al., 2017). Despite these numerous empirical studies, the accumulation of these studies necessitates research on the related factors of foreign language writing anxiety from a more generalizable meta-analytic approach.

# Literature review

## Related studies of foreign language writing anxiety

In the literature, subsequent to preliminary conceptual work of first language (L1) Writing Apprehension Test (WAT) produced by Daly and Miller (1975), an emerging array of second language (L2) studies (Cheng, 1998, 2002, 2004, 2017; Cheng et al., 1999) have begun to offer empirical insights into foreign language writing anxiety. For instance, Cheng (2004) aimed to develop and measure the reliability and validity of the Second Language Writing Anxiety Inventory (SLWAI) among 421 Chinese English-as-a-foreign-language (EFL) learners. Result of explanatory factor analysis supports a three-factor constructs: avoidance behavior, cognitive anxiety and somatic anxiety. More specifically, avoidance behavior represents "an indicative of avoidance behavior," cognitive anxiety is defined as "anxiety related to fear of negative evaluation or worrisome perceptions," and somatic anxiety refers to "anxiety related to increased physiological arousal" (Cheng, 2004, p. 325).

Apart from the empirical attempts, qualitative literature reviews on foreign language writing anxiety have also been recently presented (Ma and Dong, 2018; Abdel Latif, 2019). For instance, Ma and Dong (2018) performed a review of foreign language writing anxiety by retrieving all the related studies published in Chinese key journals from 2001 to 2015, and identified two major findings pertinent to the study: First, most of the existing studies focused on exploring the relationships between foreign language writing anxiety and its related correlates, *viz.* foreign language writing performance, affects and other related anxieties. Second, those studies published to date also obtained that the relationships were modulated by some contextual-related and learner-related variables, such as types of anxiety, language distance, target language, learners' age and foreign language proficiency. In a more recent study, Abdel Latif (2019, p. 8) critically and systematically reviewed key writing motivational constructs from the literature, and highlighted the need to "make the results of a particular study more generalizable."

## Related meta-analyses

While these empirical attempts and qualitative literature reviews may shed some light on foreign language writing anxiety research, the aggregated effects regarding correlates of foreign language writing anxiety remain largely unidentified. As such, a closer look into the research domain has been made, which reveals no meta-analysis of foreign language writing anxiety published to date, but similar meta-analyses (e.g., Teimouri et al., 2019; Zhang, 2019; Botes et al., 2020; Li, 2022a) pertinent to the study. For instance, a meta-analysis of 97 effect sizes conducted by Teimouri et al. (2019) indicated a moderate and negative correlation ($r = -0.360$) between foreign language anxiety and foreign language performance. Moderating effects regarding types of language performance, educational level, types of anxiety were also achieved. Likewise, Zhang (2019) also obtained the moderate, negative correlation ($r = -0.340$) between foreign language anxiety and foreign language performance. Apart from the moderators mentioned in Teimouri et al. (2019), language distance has also been found to significantly moderate the foreign language anxiety–foreign language performance correlation. Similarly, another meta-analysis reported by Botes et al. (2020) dealt with the correlation of Foreign Language Classroom Anxiety (FLCA) and five types of academic performance, including general language performance and four skill-specific (listening, speaking, reading and writing) performance. Concerning the result, moderately negative correlations have been achieved regarding FLCA and all types of academic performance. To the best of our knowledge, only Li's team (Li, 2022a,b) began to meta-analyze correlates of foreign language reading and listening anxiety, warranting a fresh look at other skill-specific anxieties, e.g., foreign language writing anxiety. Because it would be of vital importance to gain a deeper understanding of the correlates of foreign language writing anxiety, and earlier studies (Pae, 2013; Chen, 2019) on foreign language writing anxiety also argue that it is distinguishable from the domain-general foreign language anxiety. For instance, Pae (2013) aimed to revisit the relationship between four skill-specific anxieties and the domain-general foreign language anxiety. A multiple regression analysis indicated that foreign language writing anxiety could only explain 9.5% variance of foreign language anxiety ($\beta = 0.095$, $p = 0.041$, cf. Table 1, Pae, 2013, p. 248), suggesting that both are statistically distinguishable from each other. On the other hand, existing studies on foreign language anxiety focus too much on "test anxiety and general trait anxiety" (Chen, 2019, p. 314), which may fail to assess language learners' responses to the skill-specific foreign language writing anxiety.

Taken together, although the related meta-analytic studies have been valuable to gain an understanding, yet indirect, of fundamental aspects of foreign language writing anxiety, little is still known about its main correlates (e.g., foreign language writing

TABLE 1 Overall average correlations and publication bias test for the low-evidence correlates.

| Correlates | $k$ | $r$ [95% CI] | $Q$ | $I^2$ | $N_{fs}$ | $N_{observed}$ | $r_{adjusted}$ |
|---|---|---|---|---|---|---|---|
| Listening anxiety | 5 | 0.485 [0.415, 0.549] | 5.289 | 24.366 | 251 | 5 | 0.481 [0.430, 0.528] |
| Speaking anxiety | 9 | 0.455 [0.377, 0.526] | 21.358** | 62.544 | 685 | 9 | 0.469 [0.426, 0.509] |
| Reading anxiety | 5 | 0.489 [0.376, 0.588] | 12.258 | 67.367 | 277 | 5 | 0.532 [0.482, 0.578] |

**$p < 0.010$.

performance, foreign language writing self-efficacy, foreign language listening anxiety, foreign language speaking anxiety and foreign language reading anxiety), and potential moderators (e.g., types of anxiety, age, target language, language proficiency and language distance), calling for research into the possible correlates and moderators of foreign language writing anxiety.

## Related correlates and moderators

Currently, foreign language writing anxiety studies have focused on identifying its correlates, including foreign language writing performance (Guo and Fan, 2009), foreign language writing self-efficacy (Tola and Sree, 2016; Guo, 2018), foreign language listening anxiety (Xiao and Wong, 2014), foreign language speaking anxiety (Gkonou, 2011) and foreign language reading anxiety (Cheng, 2004), respectively.

Drawing on Li (2022a), the identification of correlates should observe the following steps: First, prior to meta-analysis, an initial search should be conducted to exhaustively identify all the related correlates. Second, those correlates of very low numbers of effect sizes ($k < 3$, see also Li, 2022a) should be removed, as they are insufficient for generating trustworthy interpretations. Normally, to make the meta-analysis more operational and reliable, correlates should be further divided into high-and low-evidence correlates. The high-evidence correlates are defined as correlates of high investigation frequency (beyond 10 effect sizes), and the low-evidence correlates are of low investigation frequency (5–9 effect sizes). Third, it is premature to execute moderator analysis for the low-evidence correlates thus far, potential moderators are identified from the literature and moderator analysis should only be done for the high-evidence correlates (Li, 2022a). In this study, we have identified two high-evidence correlates (foreign language writing performance and foreign language writing self-efficacy) and three low-evidence correlates (foreign language listening anxiety, foreign language speaking anxiety and foreign language reading anxiety) of foreign language writing anxiety. The main correlates and potential moderators of the high-evidence correlates are defined in the remainder of this section.

## High-evidence correlates and moderators

### Foreign language writing performance and correlates

As a frequently examined correlate, foreign language writing performance (*writing performance* hereafter) refers to those studies that reported foreign language learners' writing scores or grades. For instance, Liu and Ni (2015) investigated the foreign language writing anxiety–writing performance correlation among 1,174 first-year Chinese university EFL learners of intermediate level, and obtained a weak and negative correlation ($r = [-0.136, -0.091]$, $p < 0.001$), corroborating the interview result that "around one third of the learners did not report having anxiety when writing in a foreign language" (p. 55), suggesting that writing anxiety might not be so influential to learners' writing performance. Zhang (2011) examined the foreign language writing anxiety–writing performance correlation among Chinese English majors of high proficiency level, and reported a moderately negative effect ($r = [-0.879, -0.838]$, $p < 0.001$). The non-consensual results may be explained by such potential moderators as language proficiency, target language, age and types of anxiety (e.g., Zhang, 2019; Li, 2022a). Consequently, this study first calculates the aggregated foreign language writing anxiety–writing performance correlation, and then reports the moderator results of language proficiency, target language, age and types of anxiety.

### Foreign language writing self-efficacy and correlates

The operational definition of foreign language writing self-efficacy (*writing self-efficacy* hereafter) refers to learners' self-confidence in the ability to succeed in foreign language writing (Cheng, 2004). Since Cheng (2004), the foreign language writing anxiety–writing self-efficacy correlation has caught the attention of many researchers (Li and Liu, 2013; Tola and Sree, 2016; Guo, 2018). These studies regarding the significant foreign language writing anxiety–writing self-efficacy correlation have been confirmed in some (e.g., $r = [-0.760, -0.382]$, $p < 0.001$, Li et al., 2013; and $r = [-0.420, -0.360]$, $p < 0.001$, Cheng, 2004), but not in others (e.g., $r = 0.186$, $p > 0.050$, Singh and Rajalingam, 2012), which may give rise to potential moderators, including language distance, target language, language proficiency and types of anxiety (e.g., Teimouri et al., 2019; Botes et al., 2020; Li, 2022a).

## Low-evidence correlates

### Foreign language listening anxiety

Foreign language listening anxiety (*listening anxiety* hereafter) refers to the "fear of misunderstanding what language learners listen to and being embarrassed by interpreting the message wrongly" (Serraj and Noordin, 2013, p. 3). Language learners who are anxious about their listening comprehension might experience the lack of confidence and worry over foreign language listening tasks, or even "failure to recognize spoken foreign language words" (Bekleyen, 2009, p. 664). The significantly positive foreign language writing anxiety–listening anxiety correlation suggests that foreign language learners with high writing anxiety are likely to feel higher listening anxiety, and the vice versa (Xiao and Wong, 2014; Cheng, 2017), justifying the needs to have a fresh look at the role of listening anxiety by exploring the foreign language writing anxiety–listening anxiety correlation.

### Foreign language speaking anxiety

Foreign language speaking anxiety (*speaking anxiety* hereafter) is defined as a sense of fear or anxiety that language learners would feel when using, speaking or communicating in a foreign language (Woodrow, 2006). The foreign language writing anxiety–speaking anxiety correlation has been gaining attention among researchers (e.g., Gkonou, 2011; Xiao and Wong, 2014; Cheng, 2017). For instance, Cheng (2017) recruited 523 Chinese college students to measure the correlations among four foreign language-skill-specific anxieties, and found the foreign language writing

anxiety–speaking anxiety correlation was $r = 0.510$, $p < 0.050$. In another study, Gkonou (2011) also surveyed the correlation and found $r = [0.340, 0.543]$, $p < 0.050$, While these primary studies sheds light on the important role of speaking anxiety, little is known about the average correlation, necessitating a meta-analysis to aggregate the effects with larger sample sizes.

### Foreign language reading anxiety

Foreign language reading anxiety (*reading anxiety* hereafter) is defined as the "perceptions of uneasiness, apprehension or stress from which an individual might suffer when reading a foreign language text" (Capan and Karaca, 2013, p. 1362). Researchers investigated the foreign language writing anxiety–reading anxiety correlation, and found that the correlation was moderate and positive: $r = [0.272, 0.546]$, $p < 0.050$ (Xiao and Wong, 2014) and $r = 0.580$, $p < 0.050$ (Cheng, 2017). The moderate foreign language writing anxiety–reading anxiety correlation in these primary studies calls for more investigations. For this reason, we take reading anxiety as the correlate of writing anxiety.

## Research statements and questions

The current study aims to achieve two research purposes. First, we carried out a meta-analysis based on a systematic review of existing primary studies that explored the correlations of foreign language writing anxiety and its two high-evidence correlates (foreign language writing self-efficacy and foreign language writing performance) along with three low-evidence correlates (foreign language listening anxiety, foreign language speaking anxiety and foreign language reading anxiety). Second, apart from the correlations under investigation, we also examined the moderating effects of learners' age, language proficiency, target language, types of anxiety and language distance for the high-evidence correlates. To this end, the following research questions are to be addressed.

***Research question 1***: What are the correlations of foreign language writing anxiety and two high-evidence correlates (writing performance and writing self-efficacy)?

***Research question 2***: How do age, language proficiency, target language, types of anxiety and language distance moderate the correlations of writing anxiety and its two high-evidence correlates (writing performance and writing self-efficacy)?

***Research question 3***: What are the correlations of foreign language writing anxiety and three low-evidence correlates (listening anxiety, speaking anxiety and reading anxiety)?

# Research method

## Literature search and inclusion criteria

The study attempts to retrieve the currently available literature of writing anxiety in second and/or foreign language learning published during 2000 to 2021, because foreign language writing research remained few in number before 2000 (Cheng, 2002, 2004). Several electronic databases (e.g., Chinese CNKI, ERIC, ProQuest, ScienceDirect, Springer, web of science, Wiley) and search engines (Chinese Baidu Scholar and Google Scholar) were retrieved with a combination of the following key words: *affect*, *foreign language*, *second language (L2)*, *self-efficacy*, *(writing) score*, *(writing) grade*, *(writing) achievement*, *(writing) proficiency*, *listening anxiety*, *speaking anxiety*, *reading anxiety*, *writing anxiety* and *writing apprehension*. Moreover, to ensure the comprehensiveness of the literature, we conducted backward and forward citation searches based on seminal article (Daly and Miller, 1975; Cheng, 2004) and "snowballing technique" (Biernacki and Waldorf, 1981) by scanning references in the identified articles. To ensure the quality of primary literature during the selection process, only the peer reviewed journal articles, dissertations, and conference proceedings were included. The inclusion/exclusion criteria were proposed as follows:

1. The study should investigate the correlations of second or foreign language writing anxiety, writing performance (writing test, score or grade), writing affects and other skill-specific anxieties, resulting in 30 primary studies included.
2. The study should contain the statistics (e.g., correlation, sample sizes, standard error and variance, etc.) sufficient for the transformation or calculation of effect sizes. Eighteen articles were included by excluding 12 publications that failed to provide the sufficient statistics for calculation.
3. The backward and forward together with snow-balling searches from the existing studies (e.g., Daly and Miller, 1975; Cheng, 2004; Woodrow, 2011) on the section of literature review, together with specific search of each correlates yielded another three journal articles and one conference proceeding on foreign language writing anxiety needed for the forthcoming analysis.
4. Both peer-reviewed journal articles or unpublished materials (e.g., conference proceedings, master's or doctoral dissertations) were retrieved, which resulted in 22 primary studies.

## Variables coded for each study

According to Wilson (2019, p. 154), a coding scheme should "capture the pertinent information suitable for meta-analysis." Thus, the selected studies were coded in terms of related correlates (writing performance, writing self-efficacy, listening anxiety, speaking anxiety and reading anxiety) and moderators (types of anxiety, age, target language, language distance, and language proficiency). The code scheme proposed was presented in Table 2, including the following major categories:

Coding procedures were followed to ensure the methodological quality (e.g., Valentine, 2019): On the one hand, as issue of data dependencies should be considered first (Plonsky and Oswald,

TABLE 2 Coding scheme.

| Coding types | Subtypes | Operational definitions | References |
|---|---|---|---|
| Correlates | | | |
| Foreign language writing performance | Writing grade/score | Studies that reported learners' writing score or grade | Brown et al. (2018); Cheng (2004) |
| Foreign language writing self-efficacy | Self-efficacy in writing | Students' self-confidence in the ability to succeed in foreign language writing | Cheng (2004) |
| Foreign language listening anxiety | Listening anxiety | Studies that reported listening anxiety as a correlate of writing anxiety. | Kim (2000) |
| Foreign language speaking anxiety | Speaking anxiety | Studies that reported speaking anxiety as a correlate of writing anxiety | Woodrow (2006) |
| Foreign language reading anxiety | Reading anxiety | Studies that reported reading anxiety as a correlate of writing anxiety | Saito et al. (1999) |
| Moderators | | | |
| Types of anxiety | Overall anxiety | Studies that reported anxiety in general | Cheng (2004) |
| | Avoidance behavior | Students' avoidance of writing in a foreign language | |
| | Cognitive anxiety | Students' perceptual arousal to write in a foreign language | |
| | Somatic anxiety | Students' physiological arousal to write in a foreign language. | |
| Age | Child/Adolescent | Less than grade twelve (age 18) | Researcher-designed |
| | Adult | At and over grade twelve (18 or older) | |
| Target language | English | English as a foreign language (EFL) | Levine (2003) |
| | Mixed languages | Other mixed languages | |
| Language distance | Near | Indo-European L1 and Indo-European L2 | Lervåg and Lervåg (2011) |
| | Distant | Indo-European L1 and non-Indo-European L2 or Non-Indo-European L1 and Indo-European L2 | |
| Language proficiency | High | Studies that reported high (highly proficient learners), intermediate (intermediate learners) and low (foreign language beginners) level | Li (2022a) |
| | Intermediate | | |
| | Low | | |

2014), multiple studies reported in a single paper involving different types of measurement or participants were coded separately as independent studies. On the other hand, to ensure the reliability of coding scheme, two coders who had a consistent understanding of coding types, subtypes and operational definitions were required to independently code the items. They should also negotiate with each other when discrepancies occurred.

## Calculation and analysis of the effect sizes

For data calculation and interpretations, correlation coefficients, sample sizes and effect directionality were first converted to Fisher's z, and the aggregated coefficients, standard error and confidence interval were then calculated. According to Plonsky and Oswald (2014), the interpretations of the effect size were indexed as 0.25 (small), 0.40 (moderate), and 0.60 (large), respectively.

For the data analysis, both fixed and random model were utilized to compute all the aggregated correlations, depending on the different sources of variation in effect sizes. For the fixed model, all studies were assumed to share a common true effect and the between-study variation of the effect sizes is sampling error. By contrast, for the random model, the true effects were assumed to have been sampled from a between-study variation across studies (Borenstein et al., 2009; Plonsky and Oswald, 2012). As such, a random model was consulted, and the heterogeneity was located in respect to moderators including age, language proficiency, target language, types of anxiety and language distance.

# Results

Results were reported based on the 84 effect sizes with a total of 24,290 participants involved ($M \pm SD = 289.167 \pm 333.380$, range = 50–1,635). In the rest of this section, results of high-evidence correlates and moderator analysis were first reported, and then results of three low-evidence correlates followed suit. As number of low-evidence correlates was too small to analyze the moderating effects, moderator analysis for the low-evidence correlates was not executed accordingly (*cf.* Lervåg and Lervåg, 2011).

## Results of high-evidence correlates and moderator analysis

### Foreign language writing anxiety and writing performance

Forty effect sizes consisting of 14,918 participants ($M \pm SD = 372.950 \pm 436.800$, range = 50–1,635) examined the foreign language writing anxiety–writing performance correlation.

As presented in Table 3, the foreign language writing anxiety–writing performance correlation was significantly moderate, $r = -0.298$, 95% CI [−0.353, −0.240], $z(39) = -9.667$, $p < 0.001$. No any publication bias was observed, $N_{fs} = 7{,}503 > N_{observed} = 40$, $p < 0.001$, which did not affect the results.

As both significance tests [$Q(39) = 520.779$, $p < 0.001$, $I^2 = 92.511$] were significantly heterogenous, several moderator analyses should be further conducted. Results of the moderator analysis regarding language proficiency, target language, age and types of anxiety were reported in the rest of this section, respectively.

### Language proficiency

As shown in Table 4, language proficiency significantly moderates the foreign language writing anxiety–writing performance correlation ($Q_{between} = 11.408$, $p = 0.003$). Pairwise comparisons showed that, the low proficiency learners had the weakest foreign language writing anxiety–writing performance correlation ($r = -0.129$, 95% CI [−0.219, −0.037]), which was statistically lower than high proficiency learners ($Q_{between} = 7.553$, $p = 0.006$), and intermediate proficiency learners ($Q_{between} = 8.939$, $p = 0.003$). No significant difference was found between high proficiency learners and intermediate proficiency learners ($Q_{between} = 0.225$, $p = 0.635$).

### Target language

Target language involves two types, *viz.* English and mixed languages. It could be found in Table 4, target language obtained no statistically significant moderating effect on the foreign language writing anxiety–writing performance correlation ($Q_{between} = 0.225$, $p = 0.635$).

### Age

Age was found to significantly moderate the foreign language writing anxiety–writing performance correlation ($Q_{between} = 3.793$, $p = 0.051$), indicating that children and adolescents ($r = -0.393$, 95% CI [−0.481, −0.297]) tend to have more negative foreign language writing anxiety–writing performance correlation than adults ($r = -0.280$, 95% CI [−0.340, −0.218]).

### Types of anxiety

According to Cheng (2004), foreign language writing anxiety could be further classified into three types, i.e., avoidance behavior, cognitive anxiety and somatic anxiety. As apparent in Table 4, there is no significant moderating effect of types of anxiety ($Q_{between} = 1.134$, $p = 0.567$).

## Foreign language writing anxiety and writing self-efficacy

Nineteen effect sizes involving 5,626 participants ($M \pm SD = 296.105 \pm 152.918$, range = 50–738) explored the foreign language writing anxiety–writing self-efficacy correlation.

As apparent in Table 3, the foreign language writing anxiety–writing self-efficacy correlation was moderate, $r = -0.382$, 95% CI

TABLE 3 Overall average correlation and publication bias test for the high-evidence correlates.

| Correlates | *r* [95% CI] | $N_{fs}$ | $N_{observed}$ | $r_{adjusted}$ |
|---|---|---|---|---|
| Writing performance | −0.298 [−0.353, −0.240] | 7,503 | 40 | −0.185 [−0.200, −0.169] |
| Writing self-efficacy | −0.382 [−0.456, −0.302] | 3,715 | 19 | −0.348 [−0.371, −0.324] |

$N_{fs}$ = number of missing studies that would bring $p > 0.05$; $N_{observed}$ = number of observed studies.

TABLE 4 Moderator analyses for the foreign language writing anxiety–writing performance correlation.

| Moderators | *k* | *r* [95% CI] | $I^2$ | Heterogeneity | | |
|---|---|---|---|---|---|---|
| | | | | *Q* | *df* | *p* |
| *Language proficiency* | 40 | −0.252 [−0.302, −0.201] | 92.511 | 11.408** | 2 | 0.003 |
| High | 10 | −0.337 [−0.444, −0.219] | 83.799 | | | |
| Intermediate | 27 | −0.304 [−0.373, −0.232] | 94.245 | | | |
| Low | 3 | −0.129 [−0.219, −0.037] | 48.383 | | | |
| *Target language* | 19 | −0.366 [−0.432, −0.295] | 91.137 | 0.731 | 1 | 0.393 |
| English | 12 | −0.409 [−0.521, −0.284] | 93.341 | | | |
| Mixed languages | 7 | −0.345 [−0.427, −0.259] | 82.748 | | | |
| *Age* | 40 | −0.312 [−0.362, −0.259] | 92.511 | 3.793△ | 1 | 0.051 |
| Adult | 34 | −0.280 [−0.340, −0.218] | 93.113 | | | |
| Child/Adolescent | 6 | −0.393 [−0.481, −0.297] | 49.695 | | | |
| *Types of anxiety* | 27 | −0.249 [−0.337, −0.171] | 92.511 | 1.134 | 2 | 0.567 |
| avoidance behavior | 13 | −0.256 [−0.337, −0.171] | 80.303 | | | |
| cognitive anxiety | 7 | −0.222 [−0.298, −0.143] | 66.879 | | | |
| somatic anxiety | 7 | −0.188 [−0.280, −0.092] | 76.654 | | | |

△$p < 0.100$; *$p < 0.050$; **$p < 0.010$. Unreported information is not included.

[−0.456, −0.302], $z(18) = -8.717$, $p < 0.001$. No publication bias could be observed, $N_{fs} = 3{,}715 > N_{observed} = 19$, $p < 0.001$, which did not affect the results.

As both significance tests [$Q(18) = 203.084$, $p < 0.001$, $I^2 = 91.137$] were reported, necessitating further moderator analyses. Moderator analyses regarding language distance, target language, language proficiency and types of anxiety were reported in Table 5, which indicated that no moderator surveyed above could reliably explain the variation of foreign language writing anxiety–writing self-efficacy correlation.

## Results of low-evidence correlates

Results of three low-evidence correlates of foreign language writing anxiety (listening anxiety, speaking anxiety and reading anxiety) were reported in Table 1, but moderator analyses were performed as there was no sufficient data for aggregation (Li, 2022a).

Five effect sizes comprising 871 participants dealt with the foreign language writing anxiety–listening anxiety correlation. As shown in Table 1, the correlation result was significantly moderate and positive, $r = 0.481$, 95% CI [0.430, 0.528], $z(4) = 11.849$, $p < 0.001$.

Nine effect sizes comprising a total of 1,383 participants investigated the foreign language writing anxiety–speaking anxiety correlation. As shown in Table 1, the correlation was significantly moderate and positive, $r = 0.455$, 95% CI [0.377, 0.526], $z(8) = 10.213$, $p < 0.001$.

Likewise, five effect sizes consisting of 871 participants explored the foreign language writing anxiety–reading anxiety correlation. As shown in Table 1, the correlation was significantly moderate and positive, $r = 0.489$, 95% CI [0.376, 0.588], $z(4) = 7.506$, $p < 0.001$.

# Discussion

The present study endeavored to quantitatively meta-analyze the two high-evidence correlates (writing performance and writing self-efficacy) and the three low-evidence correlates (listening anxiety, speaking anxiety and reading anxiety) of foreign language writing anxiety identified in the primary literature. Simultaneously, it also dealt with moderator analyses for the two high-evidence correlates, including learners' age, foreign language proficiency, target language and language distance.

Research question 1 explored the correlations of foreign language writing anxiety and its two high-evidence correlates (writing performance and writing self-efficacy). As noted, the results demonstrated that foreign language writing anxiety has a moderate correlation with writing performance and writing self-efficacy, suggesting that worse writing performance is likely to be accompanied with higher writing anxiety, and those with higher writing anxiety tend to have a lower writing self-efficacy, mirroring an increasing number of studies maintaining the detrimental or debilitative effects of foreign language writing anxiety (e.g., Horwitz, 2017; MacIntyre, 2017). A plausible explanation might be that, those learners with higher writing anxiety might lead to the lack of self-confidence and ability to retrieve linguistic (i.e., lexical, semantic and syntactic) knowledge from the mental lexicon, choose appropriate language structure, use appropriate rhetorical devices and adopt other writing skills, then perfectly organize and output ideas as required, which would also result in unsatisfactory writing performance and a low sense of self-efficacy in writing tasks in turn (Öztürk and Saydam, 2014; Kırmızı and Kırmızı, 2015; Tola and Sree, 2016).

Research question 2 concerned the moderating effects of age, language proficiency, target language, types of anxiety and language distance on the correlations of foreign language writing anxiety and two high-evidence correlates. The results of moderator

TABLE 5 Moderator analyses for the foreign language writing anxiety–writing self-efficacy correlation.

| Moderators | *k* | *r* [95% CI] | $I^2$ | Heterogeneity | | |
|---|---|---|---|---|---|---|
| | | | | *Q* | *df* | *p* |
| *Language distance* | 19 | −0.378 [−0.431, −0.322] | 91.137 | 0.018 | 1 | 0.893 |
| Distant | 16 | −0.377 [−0.430, −0.322] | 79.254 | | | |
| Similar | 3 | −0.415 [−0.795, 0.201] | 98.371 | | | |
| *Target language* | 19 | −0.366 [−0.432, −0.295] | 91.137 | 0.731 | 1 | 0.393 |
| English | 12 | −0.409 [−0.521, −0.284] | 93.341 | | | |
| Mixed languages | 7 | −0.345 [−0.427, −0.259] | 82.748 | | | |
| *FL proficiency* | 19 | −0.367 [−0.444, −0.284] | 91.137 | 0.898 | 1 | 0.343 |
| High | 2 | −0.588 [−0.857, −0.068] | 94.276 | | | |
| Intermediate | 17 | −0.361 [−0.439, −0.277] | 91.164 | | | |
| *Types of anxiety* | 8 | −0.400 [−0.439, −0.360] | 68.527 | 3.068 | 2 | 0.216 |
| avoidance behavior | 4 | −0.490 [−0.599, −0.363] | 83.504 | | | |
| cognitive anxiety | 2 | −0.411 [−0.467, −0.350] | 0.000 | | | |
| somatic anxiety | 2 | −0.370 [−0.429, −0.308] | 0.000 | | | |

Unreported information is not included.

analysis suggested that learners' age and language proficiency, rather than target language, types of anxiety and language distance, are found to be significant moderators. Specifically, compared with the two higher proficiency learners, those low proficiency learners had the weakest foreign language writing anxiety–writing performance correlation, and no difference was found between the intermediate and high proficiency learners, resonating the argument that writing anxiety may change as a function of language proficiency (Zhang, 2019). In other words, compared with relatively high proficiency peers (*viz*. intermediate and high proficiency learners), low proficiency learners might not perform competently in writing tasks that invoke high loads of working memory and are unlikely to be actively involved in foreign language writing, hence their writing anxiety might not be triggered. A further support could be found in Horwitz (1996) who asserted that even the highly proficient language learners may experience anxiety when using a foreign language. The finding, however, is inconsistent with that of Zhang (2019) who did not find the moderating effects of language proficiency on foreign language anxiety–language performance correlation. A possible explanation for the inconsistence may rest on the difference in anxiety: domain-general language anxiety vs. skill-specific writing anxiety. In other words, while Zhang (2019) deals with the relationship between domain-general language anxiety and language proficiency, writing tasks in this meta-analysis that involves the skill-specific writing anxiety would be more cognitive resources demanding compared to language-general tasks or other receptive tasks in terms of different degree of task difficulty (He, 2018), adding to the emerging body of literature by showing the significant moderating effect of language proficiency.

Meanwhile, our meta-analysis provides another piece of evidence that the writing anxiety–writing performance correlation is sensitive to age effect, as reflected in the results that children and adolescents tended to have more negative correlations between foreign language writing anxiety and writing performance than adults. This finding, however, is not in line with Zhang (2019, p. 12) who claims "the language anxiety–language performance correlation became stronger as age increased." The discrepancy might reside in the different measures of anxiety, since Zhang's (2019) meta-analysis involves language anxiety–language performance correlation, while our study deals with writing anxiety–writing performance correlation. Another explanation for the significant age effect is that, adults' cognitive or metacognitive skills tend to be more mature to reduce foreign language writing anxiety compared to children and adolescents (Li, 2022a).

Some nonsignificant moderating effects of target language, types of anxiety and language distance should be noteworthy as well. First, regarding target language (English vs. mixed languages), no moderating effect was found on the writing anxiety–writing performance correlation, indicating that learners whose writing anxiety–writing performance correlation might not vary across different target languages. This result could be explained by the complex and demanding nature of the writing process, that is, it is not the target language, be it English or other languages, that matters, rather a complex writing process that matters (Kim and Kim, 2020). Second, the moderating effect regarding types of anxiety was not found on both the writing anxiety–writing performance correlation together with the writing anxiety–writing self-efficacy correlation, suggesting that somatic anxiety, cognitive anxiety and avoidance behavior might play a somewhat equal role in the aforementioned correlations. Third, inspired by Lervåg and Lervåg (2011) who meta-analyzed reading comprehension and its correlates, this study also examined the moderating effect of language distance regarding the orthographic difference/similarity between first and foreign language. Contrary to Lervåg and Lervåg (2011), our study obtained no significant moderating effect of language distance. A plausible explanation for such a discrepancy might lie in the difference between reading and writing modal. As for reading modal, the materials would be visually presented first. In this case, the orthographic difference/similarity of the visually presented materials might play a moderating role, whereas writing modal involves a series of complex activities, e.g., how to generate, organize ideas and how to organize the ideas in written forms. As such, the moderating effect regarding the distance between first and foreign language visually presented might not be the same case as in reading comprehension (Lervåg and Lervåg, 2011).

Research question 3 dealt with the correlations between foreign language writing anxiety and three low-evidence correlates (listening anxiety, speaking anxiety and reading anxiety). The moderate and positive correlations of the three low-evidence correlates have been obtained, resonating the previous studies that investigated different types of language learners, e.g., Korean EFL learners (Pae, 2013), Chinese-as-a-heritage-language (CHL) learners (Xiao and Wong, 2014) and Korean-as-a-heritage-language (KHL) learners (Jee, 2016), confirming the "interdependence among the four skill-specific anxieties" (Pae, 2013, p. 250).

Taken together, the pedagogical implications both for researchers and teachers in the field to help alleviate learners' writing anxiety are also inferred as follows. First, as the debilitative effects of foreign language writing anxiety have been found with regard to writing performance and writing self-efficacy, teachers should try to locate the sources of learners' writing anxiety. One implication is that teachers could establish a relaxed learning environment, design relaxation writing activities and encourage students to express their fears (Li, 2022a). Another way is to seek for some automated writing evaluation (AWE) tools that enable to provide foreign language learners with timely and supportive writing feedback (Li et al., 2019). By introducing online pedagogical intervention along with face-to-face instruction (Li, 2022c), those shy learners might be likely to feel less anxious to express their anxieties and personalized needs. Second, as the moderating effects of learners' individuality (e.g., age and language proficiency) were found to be significant, when teaching how to write well, teachers should try to alleviate learners' writing anxiety with a particular eye on learners' personalized needs. For instance, teachers normally offer special guidance to those underachievers. Our finding suggests their attention should also be paid equally to

the intermediate and advanced proficiency learners, since they might also experience tremendous anxiety when writing in a foreign language (Horwitz, 1996). Third, the interdependence among the four skill-specific anxieties suggests that foreign language researchers and practitioners should pay equally balanced attention to anxieties arising from each of the four language skills (Pae, 2013).

## Conclusion

Motivated by the earlier attempts, this study aims to understand the correlates of foreign language writing anxiety. Results showed that foreign language writing anxiety has moderate correlations with writing performance and writing self-efficacy. Besides, as compared target language, types of anxiety and language distance, significant moderating effects of learners' age and language proficiency have been obtained. The three low-evidence correlates have moderate effect sizes, with speaking anxiety, reading anxiety and listening anxiety being the moderate and positive correlate.

One potential limitation should be addressed though. Considering the needs of sufficient information from the primary studies, the current study only included limited correlates (writing performance, writing self-efficacy, listening anxiety, speaking anxiety, and reading anxiety) and moderators (learners' age, language proficiency, target language, types of anxiety and language distance). To gain a fuller understanding, future study should consider other equally important correlates (e.g., motivation, strategy and attitude) and moderators (e.g., gender and learning style) of foreign language writing anxiety.

## Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: Data will be publically available upon reasonable request. Requests to access these datasets should be directed to RL, liruidianzi@hotmail.com.

## Author contributions

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

## Funding

This study was funded by the Social Science Foundation of Hunan Province (grant number 20ZDB005).

## Acknowledgments

The author would like to thank the reviewers and editors for their constructive and insightful comments.

## 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

Articles marked with an asterisk (*) were analyzed in the meta-analysis.

*Abdel Latif, M. M. (2015). Sources of L2 writing apprehension: a study of Egyptian university students. *J. Res. Read.*, 38, 194–212, doi: 10.1111/j.1467-9817.2012.01549.x.

Abdel Latif, M. M. M. (2019). Unresolved issues in defining and assessing writing motivational constructs: a review of conceptualization and measurement perspectives. *Assess. Writ.* 42, 1–11. doi: 10.1016/j.asw.2019.100417

Alico, J. C. (2016). Writing anxiety and language learning motivation: examining causes, indicators, and relationship. *Communication and Linguistics Studies* 2, 6–12. doi: 10.11648/j.cls.20160201.12

*Asmari, A. A. (2013). Investigation of writing strategies, writing apprehension, and writing achievement among Saudi EFL-major students. *Int. Educ. Stud.*, 6, 130–143. doi: 10.5539/ies.v6n11p130

Bekleyen, N. (2009). Helping teachers become better English students: causes, effects, and coping strategies for foreign language listening anxiety. *System* 37, 664–675. doi: 10.1016/j.system.2009.09.010

Biernacki, P., and Waldorf, D. (1981). Snowball sampling: problems and techniques of chain referral sampling. *Sociol. Methods Res.* 10, 141–163. doi: 10.1177/004912418101000205

Botes, E., Dewaele, J.-M., and Greiff, S. (2020). The foreign language classroom anxiety scale and academic achievement: an overview of the prevailing literature and a meta-analysis. *Journal for the Psychology of Language Learning* 2, 26–56. doi: 10.52598/jpll/2/1/3

Borenstein, M., Hedges, L. V., Higgins, J. P. T., and Rothstein, H. R. (2009). *Introduction to meta-analysis*. Cornwall: John Wiley & Sons, Ltd.

Brown, A. V., Plonsky, L., and Teimouri, Y. (2018). The use of course grades as metrics in L2 research: a systematic review. *Foreign Lang. Ann.* 51, 763–778. doi: 10.1111/flan.12370

Capan, S. A., and Karaca, M. (2013). A comparative study of listening anxiety and reading anxiety. *Procedia Soc. Behav. Sci.* 70, 1360–1373. doi: 10.1016/j.sbspro.2013.01.198

*Chen, L. Y. (2019). 高中生英语写作焦虑现状及对写作教学的启示 [investigating the current status of junior high school students' writing anxiety and its pedagogical implications]. *Mental Health Education in Primary and Secondary School*, 35, 8–11.

Cheng, Y. S. (1998). Examination of two anxiety constructs: Second language class anxiety and second language writing anxiety [Unpublished doctoral dissertation]. University of Texas at Austin.

Cheng, Y. S. (2002). Factors associated with foreign language writing anxiety. *Foreign Lang. Ann.* 35, 647–656. doi: 10.1111/j.1944-9720.2002.tb01903.x

*Cheng, Y. S. (2004). A measure of second language writing anxiety: scale development and preliminary validation. *J. Second. Lang. Writ.*, 13, 313–335, doi: 10.1016/j.jslw.2004.07.001.

*Cheng, Y. S. (2017). Development and preliminary validation of four brief measures of L2 language-skill-specific anxiety. *System*, 68, 15–25. doi: 10.1016/j.system.2017.06.009

Cheng, Y. S., Horwitz, E. K., and Schallert, D. L. (1999). Language anxiety: differentiating writing and speaking components. *Lang. Learn.* 49, 417–446.

Daly, J. A., and Miller, M. D. (1975). The empirical development of an instrument of writing apprehension. *Res. Teach. Engl.* 9, 242–249.

*Erkan, D. Y., and Saban, A. İ. (2011). Writing performance relative to writing apprehension, self-efficacy in writing, and attitudes towards writing: a correlational study in Turkish tertiary-level EFL. *Asian EFL J.* 13, 164–192.

*Gkonou, C. (2011). Anxiety over EFL speaking and writing: a view from language classrooms. *Studies in Second Language Learning and Teaching*, 2, 267–281. doi: 10.14746/ssllt.2011.1.2.6

*Guo, J. D. (2018). 写作自我概念和自我效能对大学生英语写作成绩的影响: 写作焦虑的中介作用 [effect of EFL writing self-concept and self-efficacy on writing performance: mediating role of writing anxiety]. *Foreign Language Research* 201, 69–74. doi: 10.16263/j.cnki.23-1071/h.2018.02.011

*Guo, Y. (2019). 高中生英语写作焦虑与写作策略, 写作成绩的相关研究 [A correlational study of senior high school students' English writing anxiety, writing strategies and writing achievement] [Unpublished master's thesis]. Yanbian University.

*Guo, Y., and Fan, W. W. (2009). 大学英语分层次教学背景下的写作焦虑实证研究 [an empirical study on foreign language writing anxiety in the context of college English teaching at three levels]. *Journal of Beijing International Studies University*, 174, 79–84.

*Guo, Y., and Qin, X. Q. (2010). 中国非英语专业大学生的外语写作焦虑测试报告及其对写作教学的启示 [an empirical study of Chinese non-English major college students' writing anxiety and the pedagogical implications]. *Foreign Language World*, 137, 54–62.

He, D. Y. (2018). *Foreign language learning anxiety in China: Theories and applications in English language teaching.* Switzerland: Springer.

Horwitz, K. K. (1996). Even teachers get the blues: Recognizing and alleviating language teachers' feelings of foreign language anxiety. *Foreign Lang. Ann.* 29, 365–372. doi: 10.1111/j.1944-9720.1996.tb01248.x

Horwitz, E. K. (2017). "On the misreading of Horwitz, Horwitz, and cope (1986) and the need to balance anxiety research and the experiences of anxious language learners," in *New insights into language anxiety: Theory, research and educational implications*. eds. C. Gkonou, M. Daubney and J. M. Dewaele (Bristol, UK: Multilingual Matters), 31–50.

Jee, M. J. (2016). Exploring Korean heritage language learners' anxiety: 'we are not afraid of Korean!'. *J. Multiling. Multicult. Dev.* 37, 56–74. doi: 10.1080/01434632.2015.1029933

Kim, J. (2000). Foreign language listening anxiety: A study of Korean students learning English [Unpublished doctoral dissertation]. University of Texas at Austin.

Kim, M. K., and Kim, S. M. (2020). Dynamic learner engagement in a wiki-enhanced writing course. *J. Comput. High. Educ.* 32, 582–606. doi: 10.1007/s12528-019-09248-5

*Kırmızı, Ö., and Kırmızı, G. D. (2015). An investigation of L2 learners' writing self-efficacy, writing anxiety and its causes at higher education in Turkey. *International Journal of Higher Education* 4, 57–66. doi: 10.5430/ijhe.v4n2p57

Lervåg, M. M., and Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: a meta-analysis of the correlational evidence. *J. Res. Read.* 34, 114–135. doi: 10.1111/j.1467-9817.2010.01477.x

Levine, G. S. (2003). Student and instructor beliefs and attitudes about target language use, first language use, and anxiety: report of a questionnaire study. *Mod. Lang. J.* 87, 343–364. doi: 10.1111/1540-4781.00194

*Li, H. (2015). 大学生英语写作焦虑和写作成绩的准因果关系: 来自追踪研究的证据 [exploring the potential causal relationship between EFL writing anxiety and writing performance: a longitudinal study]. *Foreign Language World* 168, 68–75.

*Li, H., and Liu, D. R. (2013). 大学生外语写作焦虑与写作自我效能感的关系及其对写作成绩的预测 [a study of Chinese EFL learners' writing anxiety and self-efficacy and their prediction on their writing performance]. *Foreign Language Research*, 138, 48–54. doi: 10.13978/j.cnki.wyyj.2013.02.016

*Li, H., Liu, D. R., and Liu, Y. (2013). 大学生外语写作自我效能感在写作焦虑对写作成绩影响中的中介效应 [the mediating effects of EFL writing self-efficacy on the relationship between EFL writing anxiety and writing performance among Chinese college students]. *Psychol. Dev. Educ.*, 29, 385–390.

Li, R. (2022a). Foreign language reading anxiety and its correlates: a meta-analysis. *Read. Writ.* 35, 995–1018. doi: 10.1007/s11145-021-10213-x

Li, R. (2022b). Examining foreign language listening anxiety and its correlates: a meta-analysis. *Porta Linguarum* 38, 99–117. doi: 10.30827/portalin.vi38.22173

Li, R. (2022c). Research trends of blended language learning: a bibliometric synthesis of SSCI-indexed journal articles during 2000–2019. *ReCALL* 34, 309–326. doi: 10.1017/S0958344021000343

Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., and Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Comput. Assist. Lang. Learn.* 32, 784–804. doi: 10.1080/09588221.2018.1540433

*Liu, M., and Ni, H. (2015). Chinese university EFL learners' foreign language writing anxiety: pattern, effect and causes. *Engl. Lang. Teach.* 8, 46–58. doi: 10.5539/elt.v8n3p46

Ma, J., and Dong, P. (2018). 大数据背景下国内外语写作焦虑研究 [a review of second language writing anxiety: from big data perspective]. *Technology Enhanced Foreign Language Education* 180, 32–37.

MacIntyre, P. D. (2017). "An overview of language anxiety research and trends in its development," in *New insights into language anxiety: Theory, research and educational implications*. eds. C. Gkonou, M. Daubney and J. M. Dewaele (Bristol, UK: Multilingual Matters)

Öztürk, G., and Saydam, D. (2014). Anxiety and self-efficacy in foreign language writing: the case in Turkey. *Başkent Üniversitesi Eğitim Dergisi* 1, 10–21.

Pae, T. (2013). Skill-based L2 anxieties revisited: their intra-relations and the inter-relations with general foreign language anxiety. *Appl. Linguis.* 34, 232–252. doi: 10.1093/applin/ams041

Plonsky, L., and Oswald, F. L. (2012). "How to do a meta-analysis," in *Research methods in second language acquisition: A practical guide. 1st Edn.* eds. A. Mackey and S. M. Gass (Switzerland: Blackwell Publishing Ltd.), 275–295.

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

Rakedzon, T., and Baram-Tsabari, A. (2017). To make a long story short: a rubric for assessing graduate students' academic and popular science writing skills. *Assess. Writ.* 32, 28–42. doi: 10.1016/j.asw.2016.12.004

Russell-Pinson, L., and Harris, M. L. (2019). Anguish and anxiety, stress and strain: attending to writers' stress in the dissertation process. *J. Second. Lang. Writ.* 43, 63–71. doi: 10.1016/j.jslw.2017.11.005

Saito, Y., Garza, T. J., and Horwitz, E. K. (1999). Foreign language reading anxiety. *Mod. Lang. J.* 83, 202–218. doi: 10.1111/0026-7902.00016

*Sarkhoush, H. (2013). Relationship among Iranian EFL learners' self-efficacy in writing, attitude towards writing, writing apprehension and writing performance. *Journal of Language Teaching and Research*, 4, 1126–1132. doi: 10.4304/jltr.4.5.1126-1132

Serraj, S., and Noordin, N. B. (2013). Relationship among Iranian EFL students' foreign language anxiety, foreign language listening anxiety and their listening comprehension. *Engl. Lang. Teach.* 6, 1–12. doi: 10.5539/elt.v6n5p1

*Singh, T. K. R., and Rajalingam, S. K. (2012). The relationship of writing apprehension level and self-efficacy beliefs on writing proficiency level among pre-university students. *Engl. Lang. Teach.* 5, 42–52. doi: 10.5539/elt.v5n7p42

Teimouri, Y., Goetze, J., and Plonsky, L. (2019). Second language anxiety and achievement: a meta-analysis. *Stud. Second. Lang. Acquis.* 41, 363–387. doi: 10.1017/S0272263118000311

*Tola, T., and Sree, P. (2016). Students' writing self-efficacy and writing apprehension relating to their writing performance: reflection on Ethiopian first year university students. *International Research Journal of Humanities, Language and Literature*, 3, 22–36.

Tsao, J. J., Tseng, W. T., and Wang, C. C. (2017). The effects of writing anxiety and motivation on EFL college students' self-evaluative judgments of corrective feedback. *Psychol. Rep.* 120, 219–241. doi: 10.1177/0033294116687123

Tsiriotakis, I. K., Vassilaki, E., Spantidakis, I., and Stavrou, N. A. M. (2017). The examination of the effects of writing strategy-based procedural facilitative environments on students' English foreign language writing anxiety levels. *Front. Psychol.* 7:2074. doi: 10.3389/fpsyg.2016.02074

Valentine, J. C. (2019). "Incorporating judgments about study quality into research syntheses," in *The handbook of research synthesis and meta-analysis*. eds. H. Cooper, L. V. Hedges and J. C. Valentine. *3rd* ed (New York: Russell Sage Foundation), 129–140.

*Wang, C., and Sun, T. (2020). Relationship between self-efficacy and language proficiency: a meta-analysis. *System* 95, 1–11. doi: 10.1016/j.system.2020.102366

Wilson, D. B. (2019). "Systematic coding for research synthesis," in *The handbook of research synthesis and meta-analysis*. eds. H. Cooper, L. V. Hedges and J. C. Valentine. *3rd* ed (New York: Russell Sage Foundation), 153–172.

Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC J.* 37, 308–328. doi: 10.1177/0033688206071315

*Woodrow, L. (2011). College English writing affect: self-efficacy and anxiety. *System*, 39, 510–522, doi: 10.1016/j.system.2011.10.017.

Wu, C., and Lin, H. (2016). Learning strategies in alleviating English writing anxiety for English language learners (ELLs) with limited English proficiency (LEP). *Engl. Lang. Teach.* 9, 52–63. doi: 10.5539/elt.v9n9p52

*Xiao, Y., and Wong, K. F. (2014). Exploring heritage language anxiety: a study of Chinese heritage language learners. *Mod. Lang. J.*, 98, 589–611, doi: 10.1111/modl.12085.

Xiu, Y. J., and Wang, H. (2012). Second language writing anxiety and translation: performance in a Hong Kong tertiary translation class. *The Interpreter and Translator Trainer* 6, 171–194. doi: 10.1080/13556509.2012.10798835

*Yang, Y. (2017). "An empirical study on correlation between second language writing anxiety and CET-4 writing score," in *2017 3rd International Conference on Social Science and Higher Education, Advances in Social Science, Education and Humanities Research*, 99, 387–390. doi: 10.2991/icsshe-17.2017.97

*Zhang, H. X. (2011). A study on ESL writing anxiety among Chinese English majors: Causes, effects and coping strategies for ESL. Unpublished master's dissertation. Kristianstad University.

Zhang, X. (2019). Foreign language anxiety and foreign language performance: a meta-analysis. *Mod. Lang. J.* 103, 763–781. doi: 10.1111/modl.12590

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 28 November 2022
DOI 10.3389/fpsyg.2022.1046340



OPEN ACCESS

EDITED BY
Hui Zhang,
Nanjing Normal University,
China

REVIEWED BY
Susana Silva,
University of Porto,
Portugal
Anita Habók,
University of Szeged,
Hungary

*CORRESPONDENCE
Rui Li
liruidianzi@hotmail.com

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

RECEIVED 16 September 2022
ACCEPTED 10 November 2022
PUBLISHED 28 November 2022

CITATION
Guo N and Li R (2022) Measuring Chinese English-as-a-foreign-language learners' resilience: Development and validation of the foreign language learning resilience scale.
*Front. Psychol.* 13:1046340.
doi: 10.3389/fpsyg.2022.1046340



# Measuring Chinese English-as-a-foreign-language learners' resilience: Development and validation of the foreign language learning resilience scale

Nianyu Guo and Rui Li*

School of Foreign Languages, Hunan University, Changsha, China

Despite the growing body of research on the factors of resilience in diverse fields, there is still a dearth of particular attention on foreign language learning resilience. To fill the gap, this study seeks to develop the foreign language learning resilience scale (*FLLRS*) to measure its psychometric scale reliability and validity in Chinese English-as-a-foreign-language contexts. Valid data were collected from 313 Chinese English-as-a-foreign-language college students who voluntarily participated in the survey. The *FLLRS* was validated based on a series of reliability (e.g., item analysis, split-half reliability and internal consistency) and validity (e.g., construct validity, convergent validity and discriminant validity) tests. Results suggested that the 19-item *FLLRS* presented three factors: *ego resilience*, *metacognitive resilience* and *social resilience*. Besides, all the three factors contributed high effects to foreign language learning resilience. Among the three factors, *metacognitive resilience* was found to have the highest path coefficient, followed by *social resilience*, with *ego resilience* having the lowest. The validated scale could advance knowledge in the field of second language acquisition regarding how learners' individual differences, emotional factors and the contextual antecedents may affect foreign language learning resilience.

KEYWORDS

**foreign language learning resilience scale, English as a foreign language, ego resilience, metacognitive resilience, social resilience**

## Introduction

Resilience is about individuals' capability of making positive adaptation to the stressful and challenging situations (Lazarus and Folkman, 1984; Ungar, 2008; Mansfield et al., 2016). Individuals may encounter a plethora of difficulties brought by high-demanding assignments, negative relationship with peers and teachers and the imposition coming from

their families (Yun et al., 2018; Brewer et al., 2019; Trigueros et al., 2020). Resilient individuals are more capable of dealing with these adversities (Axford et al., 2014; Broadbent and Poon, 2015; Chan et al., 2021), while others are less inclined to absenteeism (Seçer and Ulas, 2020), psychological disorders (Zhang et al., 2020), and even self-handicap (Hunsu et al., 2021).

In the past decade, an increasing number of studies have been designed to validate measures for resilience in clinical psychology (Wagnild and Young, 1993; Connor and Davidson, 2003), education psychology (Block and Kreman, 1996; Cassidy, 2016), mathematics (Martin and Marsh, 2008a, 2008b), engineering learning (Hunsu et al., 2021) and other contexts (van der Meer et al., 2018), which might provide some valuable insights into understanding the factors of resilience. Nevertheless, very few studies to date have sought to dig deeper into the phenomenon of resilience among students in foreign language (FL) contexts (Nguyen et al., 2015; Yun et al., 2018; Sudina and Plonsky, 2021), as it is crucially important to facilitate FL learning process when they can make positive adaptation to such adversities as FL anxieties (Li, 2022a,b), FL guilt and shame (Teimouri, 2018), and untimely feedback from teachers due to high teacher-student ratio (Li, 2021). To our knowledge, most of these studies aimed to explore predictors or correlates of resilience among FL learners in South Korea (Kim et al., 2018, 2019; Kim and Kim, 2021), Western Canada (Lou and Noels, 2020a,b) and Southwestern America (Sudina and Plonsky, 2021) without a detailed scrutiny of its factors. Thus, it is necessary to develop and validate a FL resilience scale by expanding the dimensions under analysis.

To this end, this study measures resilience among Chinese English as a FL (EFL) learners, so large a population that should not be ignored (Li, 2021), and validates a Chinese version of the foreign language learning resilience scale (*FLLRS*) to provide a more up-to-date vision on this issue. In doing so, drawing on existing studies (e.g., Zhang et al., 2021), the instrument scale development should be validated by both the explanatory factor analysis (EFA) and confirmatory factor analysis (CFA) methods. More specifically, while the EFA lacks the goodness-of-fit indexes that CFA offers, it considers the cross-loadings of all items and avoids zero cross-loadings of the related construct, which enables us to gain a clear understanding of the possible dimensions. On the other hand, while the CFA ignores cross-loadings of the measurement model, it can not only provide sufficient goodness-of-fit indexes, but also further test the predictive power of the constructs. To this end, it aims to (a) understand the factors underlying the *FLLRS*; while (b) further validating the scale with CFA to understand the different effects of the factors in Chinese EFL contexts.

# Literature review

## Theoretical framework of resilience

In regard to resilience, this study adopts Lazarus and Folkman (1984) stress and coping theory (SCT) as the framework, as it highlights that resilience tends to occur when "judgment of person-human relationship is stressful hinges on personal, cognitive and situational appraisals" (Lazarus and Folkman, 1984, p. 21). Motivated by the SCT, three components – personal, cognitive and situational appraisals – are elaborated in the remainder of this section, respectively.

Personal appraisal is related with *commitments* and *beliefs*. *Commitments* refer to what is important or meaningful to an individual. For instance, when resilient EFL learners perceive English learning as an important activity, they will appraise or evaluate it as something meaningful and "maintain valued ideals to achieve desired goals" (*ibid.*, p.56). *Beliefs* are defined as "personally formed configurations" (*ibid.*, p.65) that usually operate at a tacit level to determine the understanding of a fact. For instance, perseverant EFL learners might hold the belief that English cannot be mastered for a short period of time.

Cognitive appraisal "can be readily understood as the process of categorizing an encounter, and its various facets, with respect to its significance for well-being" (*ibid.*, p.31). Lazarus and Folkman (1984) further classified it into *primary appraisal* ("Am I in trouble or being benefited, now or in the future, and in what way?") and *secondary appraisal* ("What if anything can be done about it?"). The *primary appraisal* refers to the process in which an individual evaluates the relationship with the situation he/she locates. The *secondary appraisal*, on the other hand, is about an individual's capacities to deal with the stressful and challenging situations. During the process, resilient individuals may adopt a series of cognitive and metacognitive resources, such as perceiving, monitoring, judging and discriminating information, to deliberately focus on the positive aspects of what is happening or solving the underlying distress with positive emotions (Li et al., 2021).

Situational appraisal is defined as the identification of situational properties that "may potentially be harmful, dangerous and threatening" (Lazarus and Folkman, 1984, p. 82). In other words, resilient individuals are able to make the positive adaption to the stressful situations by taking advantage of the limited resources. For instance, resilient EFL learners who have fully evaluated the situational properties would seek help from classmates or teachers in times of English learning difficulties.

Taken together, it should be noted here that, despite the three appraisals being separately classified, they are closely interconnected to shape the capability of resilience. The SCT contributes to advancing our understanding of personal, cognitive and situational appraisals with respect to the domain-general resilience, and paves the way for understanding the factors of FL learning resilience in particular.

## Types of resilience

Drawing on the theoretical insights of SCT, the understanding of resilience's factors has long been a focus of interest in relation to personal, cognitive and situational factors. Accordingly, the

most typical types of resilience are *ego resilience* (Block and Kreman, 1996; Chen et al., 2020; Chen and Padilla, 2022), *cognitive resilience* (Cassidy, 2016; Ang et al., 2021; Jahedizadeh et al., 2021) and *social resilience* (Ungar, 2008; Ungar et al., 2021), among others.

### Ego resilience

*Ego resilience* refers to how individuals respond to adversities they are experiencing, and consider their capacity to recover from the adversities (Maltby et al., 2015). In the social psychology contexts, researchers (Block and Kreman, 1996) have explored the constructs of *ego resilience* and proposed one of the most influential scales – 14-item *Ego-Resiliency Scale (ER89)* – having two factors: *openness to life experience* and *optimal regulation*. Wagnild and Young (1993) also explored the factors of *ego resilience* from a 25-item scale and obtained a two-factor structure, *viz. personal competence* and *acceptance of self and life*. Similarly, Connor and Davidson (2003) developed a 25-item *Connor–Davidson Resilience Scale (CD–RISC)* in a clinical context, and extracted five factors – Factor 1 (*personal competence, high standards and tenacity*), Factor 2 (*trust in one's instincts, tolerance of negative effect, strengthening effects of stress*), Factor 3 (*positive acceptance of change, and secure relationships*), Factor 4 (*control*), and Factor 5 (*spiritual influence*).

### Cognitive resilience

*Cognitive resilience* refers to individuals who seek some positive cognitive strategies (e.g., goal setting, goal planning, help-seeking and control) and activate their cognitive mechanisms (e.g., growth mindset, inhibition control, working memory, cognitive flexibility and cognitive emotion regulation) to deal with the challenging and stressful situations (Cassidy, 2016; Lou and Noels, 2020a,b). In the educational contexts, for instance, Cassidy (2016) developed an *Academic Resilience Scale (ARS-30)* to measure learners' cognitive, strategic and successful adaptation to academic challenges. The *ARS-30* has three factors, namely Factor 1 (*perseverance*), Factor 2 (*reflecting and adaptive help-seeking*), and Factor 3 (*negative affect and emotional response*). In a recent attempt, van der Meer et al. (2018) developed a two-factor *Resilience Evaluation Scale (RES)* – *self-confidence* and *self-efficacy* – to understand how individuals handle academic difficulties in stressful and challenging situations.

### Social resilience

*Social resilience* is the process of "navigating the necessary resources" by positively connecting with parents, schools and communities (Ungar, 2019, p.2). In psychiatric contexts, Friborg et al. (2003) developed a five-factor *Resilience Scale for Adults (RSA)* – *personal competence*, *social competence*, *family coherence*, *social support* and *personal structure* – to investigate how intrapersonal and interpersonal factors help individuals make positive adaptation to stressful situations. Using mixed methods, Liebenberg et al. (2011) validated a three-factor *Child and Youth Resilience Measure (CYRM-28)* in a cross-cultural context – *individual factors*, *caregiving or relational factors*, and *contextual components*. Insights gained from the *CYRM-28*, the exploration of *social resilience* has been conducted in other equally important contexts, such as psychotherapeutics (Chan et al., 2021), education psychology (Lavy and Ayuob, 2019), and social psychology (Ungar et al., 2021).

While the aforementioned studies provide valuable insights into the types of resilience, there are some limitations of the existing literature. On the one hand, it is evident to note that most of these studies examine only one type of resilience but fail to integrate *ego resilience*, *cognitive resilience* and *social resilience* in a study, and "most effectively differentiate the factors that are (and are not) components, causes, and correlates" (Martin and Marsh, 2009, p. 353). On the other hand, while three types of resilience driven by the SCT in the domain-general contexts have been thoroughly investigated, understanding its factors may be context-specific and cannot be easily generalized to EFL learning contexts. In other words, it remains largely unknown whether these studies could be extended to FL learning and there is still substantial room for further research on resilience in FL learning contexts.

## Resilience in FL learning contexts

In FL learning contexts, resilience is defined as the ability to "overcome stress and maintain high mental stamina" in FL learning adversities (Kim et al., 2018, p. 56). To date, an emergent body of research (Nguyen et al., 2015; Kim et al., 2019; Lou and Noels, 2020a,b; Sudina and Plonsky, 2021; Chen and Padilla, 2022) seeks to examine the relationship between resilience and other variables in the FL learning contexts, including stress and coping (Gregersen et al., 2021), (de-)motivation and language proficiency (Nguyen et al., 2015; Kim et al., 2018, 2019; Kim and Kim, 2021), emotions, creativity, or growth mindsets (Sudina and Plonsky, 2021; Chen and Padilla, 2022), and buoyancy, grit and academic perseverance (Lou and Noels, 2020a,b). For instance, Kim et al. (2018) collected qualitative data from 23 EFL learners and nine teachers, and identified four components of FL resilience: *social support*, *emotional regulation*, *a clear learning goal* and *tenacity in EFL learning*. Shortly afterwards, they (Kim et al., 2019) further collected quantitative data from 367 South Korean elementary school students to explore the impact of FL resilience on (de) motivation and language proficiency. Results indicated that FL resilience consists of *metacognitive adaptation*, *sociability*, *optimism*, *perseverance* and *communicative efficacy*, and it was reported to have a direct impact on motivation. In a recent study, informed by academic resilience (Martin and Marsh, 2008a,b; Sudina and Plonsky, 2021) collected data from 360 Fl learners based on an 8-item *Foreign Language Buoyancy Scale (FLBS)* to investigate the related correlates of FL learning perseverance. Using EFA, they obtained two components of FL resilience: *coping with poor grades and criticism* and *dealing with study stress*.

## The present study

What emerges from the above review is that despite the growing diversity of studies providing some valuable insights into understanding the correlates of FL resilience, research gaps on its factors remain open for debate. First, on a scrutiny of the studies involved, some of these lack a solid theoretical framework for FL resilience. For instance, as the first scale measuring FL resilience, eight items of *FLBS* that were directly adopted based on the domain-general academic resilience (Martin and Marsh, 2008a, 2008b) may discount the context-specificity and theoretical foundations. Similarly, Kim et al. (2018, 2019) preliminary attempts for FL resilience also failed to consider the theoretical underpinnings of FL learning resilience. Second, given the increased scrutiny for relationship between resilience and other variables in FL learning contexts, there is a desperate lack of research that should validate and measure the factors of FL resilience. Third and importantly, while FL resilience of the existing studies was measured for very specific population, namely primary, secondary school, and college-level students in South Korea (Kim et al., 2018, 2019; Kim and Kim, 2021), college students in Western Canada (Lou and Noels, 2020a,b) and Southwestern America (Sudina and Plonsky, 2021), its applicability to Chinese EFL learners remains largely underexplored. As Sudina and Plonsky (2021, p.13) put it, "future studies need to cross-validate the factors with a new sample of FL learners."

To fill a void in this line, the purposes of the study are two-fold: Motivated by the theoretical framework of SCT, the first aim is to develop and validate the *FLLRS*, so as to profile the factors underlying the *FLLRS* in Chinese EFL contexts. A second aim is to understand the extent to which different factors may contribute to the overall FL resilience. Consequently, two research questions are to be addressed as follows.

Research question 1 (RQ1): What are the factors underlying the *FLLRS* in Chinese EFL contexts?

Research question 2 (RQ2): How do the factors contribute to FL learning resilience?

# Materials and methods

## Participants

A total of 420 EFL undergraduate students recruited from four Double First-Class (*viz.* world class universities and disciplines) universities in central China volunteered and consented to participate in the online survey[1] through the convenient sampling method in the classroom. It takes roughly 30 min for the participants to complete the questionnaire. The study was approved by the Ethics Committee of Hunan University. Students of the Double First-Class universities were chosen for the following considerations. On the one hand, these four prestigious universities evenly distributed across similar levels of higher education institutions in central China, warranting the homogeneity of the data collected. On the other hand, the emphasis in the Double First-Class universities on an international outlook in general, and on a quality education in English in particular, enables students to acquire a good mastery of English. Thus, students of these universities are more likely to achieve a high level of resilience if they encounter difficulties with the English language. There was neither incentive for completing the survey, nor was there any penalty for not completing the questionnaire. The data of 107 students were removed due to their failure in trap questions, resulting in 313 valid data for analysis. Among the 313 participants, only those in Year 1 (93.6%, $N = 293$) and Year 2 (6.4%, $N = 20$) were investigated because non-English major students of Year 3 and 4 did not attend English class in China. There were 119 males (aged: $18.43 \pm 1.17$) and 195 females (aged: $18.40 \pm 0.72$). The average ages of the participants were 18.41 ($SD = 0.91$) years old. These EFL learners are of intermediate proficiency level based on their national-scale college English entrance test scores of $117.09 \pm 16.83$ (full score: 150). According to Boateng et al. (2018), the minimum number of participants needed for the analysis to be valid should follow the rule of thumb, which requires at least 10 respondents for each scale item. As such, the total of 313 valid participants which was higher than 240 ($24 \times 10$) met the criteria.

## Item generation procedures

Before initial item generation, item development for the *FLLRS* was based on the theoretical framework of SCT (Lazarus and Folkman, 1984), and the synthesis of related existing studies regarding resilience research in educational psychology (Wagnild and Young, 1993; Block and Kreman, 1996; Connor and Davidson, 2003; Friborg et al., 2003; Campbell-Sills and Stein, 2007; Cassidy, 2016; van der Meer et al., 2018; Chen et al., 2020; Ungar et al., 2021), and insights provided by skilled researchers and learners of similar background. We generated an initial pool of 24 items for three proposed factors: *ego resilience* (8 items, e.g., "I am curious about the new knowledge when I study a FL."), *metacognitive resilience* (8 items, e.g., "I would use the feedback to improve my FL.") and *social resilience* (8 items, "When I am encountered with difficulties in FL learning, I would seek help from my teachers.").

After the initial item generation, detailed questionnaire development procedures were observed as follows. First, questionnaire items of the *FLLRS* were first translated into simplified Chinese. Second, Chinese version of the items was translated back to English by a teacher of English translation using a forward-backward translation (Li, 2021; Li et al., 2021). The high similarity between two versions confirmed its accuracy. Minor adjustments to wording and formatting were made accordingly. Third, face validity of the items was reviewed and

1 https://www.wjx.cn/

confirmed by another five researchers, including two researchers in educational psychology and three in second language acquisition. Fourth, to ensure that the questionnaire items caused no misinterpretations and were fully understood, wording of the items was reviewed and discussed in a pilot study of 32 EFL learners with similar educational background. Minor adjustments to wording were further resolved by consensus through discussions. The initial 24 items (Appendix 1) had a 7-point Likert scale survey anchored on "1 = strongly disagree" and "7 = strongly agree."

## Data analysis

A series of explanatory and confirmatory factor analyses was performed in an attempt to solve the two research questions. To gain a better understanding of the factors (RQ1), results of psychometric validity and reliability of the *FLLRS* were reported first. In doing so, item analysis, reliability analyses (internal consistency and split-half reliability) and EFA were conducted. For item analysis, statistical comparison of 27% upper and lower items should be made to ensure the discrimination of each item. For reliability analyses, the cut-off values of Cronbach's α and split-half reliability should be over 0.70 (Li et al., 2019). Second, to answer RQ2 regarding the contribution of the factors, CFA of the *FLLRS* reporting the measurement and structural model was carried out accordingly.

# Results

In what follows, results corresponding to research questions were presented in the remainder of this section.

## Psychometric validity and reliability of the FLLRS

### Item analysis

Item analysis was performed with independent samples *t*-test to compare the statistical difference of responses between 27% upper items (*viz.* the highest 27% ratings) and 27% lower items (*viz* the lowest 27% ratings) based on participants' rating scores of the 7-point Likert scale (Li, 2021). The results indicated that significant between-group difference was obtained for each of the 24 items (all $ps < 0.001$), suggesting the high discrimination of each item appropriate for further analysis.

### EFA

Drawing on Kaiser (1970), the EFA was adopted with a principal components analysis (PCA) and Varimax rotation (e.g., Dewaele and MacIntyre, 2016) to determine which of 24 items clustered together to form general factors. Those factors that had more than one item with an eigenvalue ≥1.00 and the factor loadings greater than 0.4 on the intended factor but less than 0.4 on any other factor were retained.

Bartlett's test of sphericity ($\chi^2 = 5088.039$, $df = 276$, $p = 0.000$) and the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO = 0.926) exceeded the recommended value of 0.6 (Kaiser, 1970). The first factor analysis yielded five factors, which accounted for 68.918% of the total variance. However, the cross-loading problems suggested further iterative deletion and analysis. After the iterative deletion of five cross-loading items (Item 4, 5, 7, 8 and 16, see Appendix 1 for more), factor analysis of the remaining 19 items that did not have the cross-loading problems met the criteria with satisfied Bartlett's test of sphericity ($\chi^2 = 4077.672$, $df = 171$, $p = 0.000$) and the KMO of 0.925. Table 1 demonstrated the results of EFA regarding factors, items, item means and standard deviations, Cronbach's α and factor loadings, respectively.

In Table 1, three factors explained 64.986% of the total variance with robust factor loadings (>0.50) on the intended factor. Factor 1 (eigenvalue = 1.184) was labeled *ego resilience* (four items) and explained 14.936% of the variance, which refers to EFL learners' personal attributes, such as perseverance, curiosity and energy, to recover from the FL learning adversities. Factor 2

TABLE 1 Results of explanatory factor analysis: Varimax rotated factor loadings.

| Factor | Item | *M* ± *SD* | Cronbach's α | Factor loadings | | |
|---|---|---|---|---|---|---|
| | | | | Factor 1 | Factor 2 | Factor 3 |
| ER | – | 5.332 ± 1.083 | 0.806 | – | – | – |
| | 1 | 4.86 ± 1.499 | | 0.804 | | |
| | 2 | 4.57 ± 1.479 | | 0.756 | | |
| | 3 | 5.60 ± 1.386 | | 0.734 | | |
| | 6 | 6.29 ± 1.030 | | 0.544 | | |
| MR | – | 4.861 ± 1.234 | 0.931 | – | – | – |
| | 9 | 4.92 ± 1.475 | | | 0.781 | |
| | 10 | 4.71 ± 1.446 | | | 0.722 | |
| | 11 | 4.80 ± 1.479 | | | 0.806 | |
| | 12 | 5.04 ± 1.489 | | | 0.717 | |
| | 13 | 4.98 ± 1.428 | | | 0.779 | |
| | 14 | 4.73 ± 1.480 | | | 0.786 | |
| | 15 | 4.84 ± 1.483 | | | 0.654 | |
| SR | – | 4.748 ± 1.177 | 0.900 | – | – | – |
| | 17 | 4.73 ± 1.445 | | | | 0.735 |
| | 18 | 4.42 ± 1.487 | | | | 0.741 |
| | 19 | 4.82 ± 1.554 | | | | 0.687 |
| | 20 | 5.07 ± 1.458 | | | | 0.715 |
| | 21 | 4.87 ± 1.460 | | | | 0.715 |
| | 22 | 4.84 ± 1.546 | | | | 0.624 |
| | 23 | 4.64 ± 1.732 | | | | 0.596 |
| | 24 | 4.58 ± 1.563 | | | | 0.744 |

M, mean; SD, standard deviation; factor loadings more than 0.50 are presented; ER, ego resilience; MR, metacognitive resilience; SR, social resilience; factor loadings lower than 0.500 were not presented.

TABLE 2 Overall reliability and validity analysis of the measurement model.

| | Reliability | Convergent validity | Discriminant validity | | |
|---|---|---|---|---|---|
| | | | Latent variable correlations | | |
| **Factor** | **CR** | **AVE** | **ER** | **MR** | **SR** |
| ER | 0.810 | 0.542 | **0.736** | | |
| MR | 0.930 | 0.656 | 0.651** | **0.807** | |
| SR | 0.899 | 0.528 | 0.575** | 0.689** | **0.727** |

ER, ego resilience; MR, metacognitive resilience; SR, social resilience; CR, composite reliability; AVE, average variance extracted. *$p < 0.05$, **$p < 0.01$. Square roots of AVEs are shown as diagonal elements in bold type.

TABLE 3 Model fit indices of confirmatory factor analysis.

| **Model fit indices** | $\chi^2/df$ | **GFI** | **AGFI** | **CFI** | **TLI** | **NFI** | **RMSEA** |
|---|---|---|---|---|---|---|---|
| Result | 2.532 | 0.897 | 0.862 | 0.946 | 0.935 | 0.914 | 0.070 |
| Suggested | <3 | >0.90 | >0.80 | >0.90 | >0.90 | >0.90 | <0.10 |
| Evaluated | Good | Close | Good | Good | Good | Good | Good |

GFI, goodness-of-fit index; AGFI, adjusted goodness-of-fit index; CFI, comparative fit index; TLI, Tucker-Lewis index; NFI, normed fit index; RMSEA, root-mean-square error of approximation.

(eigenvalue = 9.539) was labeled *metacognitive resilience* (seven items) and explained 25.601% of the variance, which means EFL learners may seek for metacognitive strategies, such as goal setting, goal planning, help-seeking and control, to deal with FL learning difficulties. Factor 3 (eigenvalue = 1.624) was labeled *social resilience* (four items) and explained 24.499% of the variance, which means EFL learners may establish positive social connection with parents, schools and communities to solve FL learning problems. The means (*M*) and standard deviations (*SD*) of each factor were presented in Table 1. All scored above 4, with *ego resilience* (5.332 ± 1.083) being the highest, followed by *metacognitive resilience* (4.861 ± 1.234), and *social resilience* (4.748 ± 1.177).

## Reliability

### Internal consistency

Internal consistency was assessed using Cronbach's $\alpha$ for each structure (Table 2), which showed high reliability results as reflected in the Cronbach's α for the overall scale ($\alpha = 0.940$), *ego resilience* ($\alpha = 0.806$), *metacognitive resilience* ($\alpha = 0.931$) and *social resilience* ($\alpha = 0.900$), respectively.

### Split-half reliability

Split-half reliability was used to evaluate the internal reliability of the *FLLRS*. The *rhh* correlation between the two halves (First half: Item 1, 2, 3, 6, 9, 10, 11, 12, 13 and 14; Second half: Item 14, 15, 17, 18, 19, 20, 21, 22, 23 and 24) was 0.745, and the Spearman-Brown *rtt* for the overall scale was 0.854, indicating the high internal reliability of the scale for further analysis.

## CFA of the FLLRS

The confirmation of factors obtained from the EFA was further tested with CFA, *viz.* a technique used to understand the extent to which each factor affects the overall *FLLRS*.

### Measurement model

The reliability, convergent validity and discriminant validity were reported in Table 2. The reliability of the measurement model was confirmed, as composite reliability was over the minimum of 0.60 (Hair et al., 2006). The validity of the measurement model was also confirmed, since values of average variance extracted (AVE) for each factor exceeded the threshold value of 0.05 and discriminant validity was higher than the corresponding latent variable correlations (Fornell and Larcker, 1981). As such, overall results of reliability and validity of the measurement model were confirmed, since all the values met the required criteria.

### Structural model

Using the maximum likelihood method, the structural model was evaluated with six indices involved: normed chi-square, goodness-of-fit index (GFI), adjusted goodness-of-fit index (AGFI), comparative fit index (CFI), Tucker-Lewis index (TLI), and root-mean-square error of approximation (RMSEA), respectively. Results of these indices that were summarized in Table 3 met the suggested values (Li et al., 2021), indicating the appropriateness of the structural model.

The structural model of *FLLRS* was validated and presented in Figure 1. It was found that all the three factors can positively predict EFL learners' FL learning resilience with large effect sizes (0.25, 0.40, and 0.60 for small, moderate, and large, see Plonsky and Oswald, 2014), while the path coefficient of *metacognitive resilience* ($\beta = 0.92$, $p < 0.001$) was higher than that of *social resilience* ($\beta = 0.83$, $p < 0.001$) and *ego resilience* ($\beta = 0.73$, $p < 0.001$). Among the three factors, *ego resilience* has the lowest coefficient.

# Discussion

This study contributes to the field of second language acquisition (SLA) and adds to the emerging body of SLA literature by constructing and validating the factors of *FLLRS*. The deeper understanding of a new cognitive and conative factor in FL learning and teaching—FL learning resilience—is pedagogically crucial to the stakeholders (e.g., teachers, policy-makers, institutional leaders, etc.) who should pay particular attention to FL learners' resilience in times of difficulties or adversities. Considering an increased focus on the language-specific correlates of perseverance in FL learning in general (Sudina and Plonsky, 2021), and FL resilience in particular (Nguyen et al., 2015; Kim

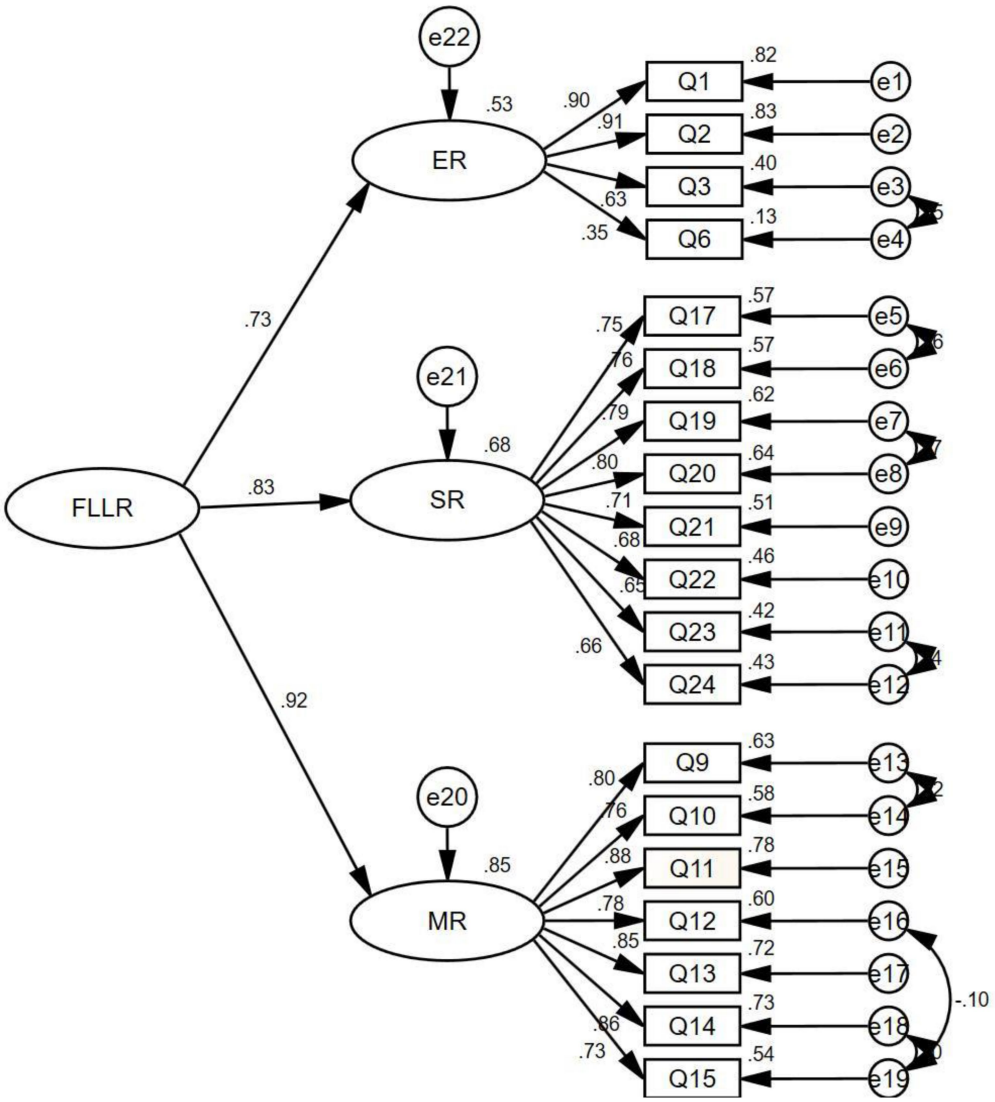



FIGURE 1
Confirmatory factor analysis results for the three-factor structure. $N$=313, $\chi^2$=359.597, $df$=142, $p$<0.001. FLLR, foreign language learning resilience; ER, ego resilience; SR, social resilience; MR, metacognitive resilience.

et al., 2018, 2019; Lou and Noels, 2020a,b), the current study, drawing on the theoretical framework of SCT (Lazarus and Folkman, 1984), provides an FL-specific scale measuring EFL learners' resilience, and explores the different impacts of factors on FL resilience. More specifically, this study first examined the factors of *FLLRS* with a series of reliability (e.g., item analysis, split-half reliability, and internal consistency) and validity (e.g., construct validity, convergent validity and discriminant validity) tests, which showed that the 19-item FLLRS has satisfactory psychometrical properties to be used as a validated scale in the FL learning contexts for future research. Additionally, the 19-item *FLLRS* was further validated from the CFA. Among the three factors, *metacognitive resilience* was found to have the highest path coefficient, followed by *social resilience*, with *ego resilience* having the lowest.

In response to the first research question, FL resilience is found to be a three-factor structure in the Chinese EFL learning contexts, including *ego resilience*, *metacognitive resilience* and *social resilience*, resonating Lazarus and Folkman (1984, p. 21) "personal, cognitive and situational appraisals" with regard to the theoretical assumptions of SCT. Like participants of the other contexts, FL learners who encounter diverse psychological issues and stressful events in FL learning are likely to trigger such personal attributes as perseverance, curiosity and energy (*ego resilience*), or seek for metacognitive strategies, such as goal setting, goal planning, help-seeking and control (*metacognitive resilience*), and establish positive social connections with other parties (*social resilience*). The three factors identified may inform pedagogy for EFL stakeholders on how to build resilient environments and develop learners' resilience in times of FL learning adversities. A more informative interpretation of the

three-factor structure could be achieved by comparing the structures with results of similar studies on FL resilience (Kim et al., 2019; Sudina and Plonsky, 2021). For instance, in a recent study, Sudina and Plonsky (2021) obtained two components of FL resilience: *coping with poor grades and criticism* and *dealing with study stress*, which is similar to *metacognitive resilience* of our study that highlights the use of metacognitive strategies to plan, monitor, evaluate and manage their FL learning difficulties or adversities (Li, 2022a). Likewise, to examine the impact of FL resilience on (de) motivation and language proficiency, Kim et al. (2019) directly used the domain-general academic resilience scale (Shin et al., 2009) to measure South Korean elementary school students' FL resilience. The five-factor structure (*viz. metacognitive adaptation*, *sociability*, *optimism*, *perseverance*, and *communicative efficacy*) of their study could be further simplified as personal-related (*optimism* and *perseverance*), cognitive-related (*metacognitive adaptation* and *communicative efficacy*) and social related (*sociability*) resilience under the frame of Lazarus and Folkman (1984) SCT, suggesting that the three-factor structure of *FLLRS* in this study is more psychometrically elegant and simple.

The second question concerns the extent to which different factors predict the overall FL resilience. The CFA supports results obtained from the EFA, suggesting the psychometric validation of the 19-item *FLLRS* developed in this study. All the path coefficients of *metacognitive resilience* ($\beta = 0.92$, $p < 0.001$), *social resilience* ($\beta = 0.83$, $p < 0.001$) and *ego resilience* ($\beta = 0.73$, $p < 0.001$) have large effect sizes based on Plonsky and Oswald (2014) interpretations of the magnitude: 0.25, 0.40, and 0.60 for small, moderate, and large effects, respectively. However, when comparison of path coefficients was made among the three factors, *metacognitive resilience* was found to be the largest, followed by *social resilience*, with *ego resilience* having the lowest, indicating that EFL learners' metacognitive resilience should be highlighted. In other words, to efficiently overcome adversities in FL learning, EFL learners should be sensitive to adopt various metacognitive strategies to plan, monitor, evaluate and manage their FL learning activities (Li, 2022a). Intriguingly, this result does not lend support to the descriptive statistic results with *ego resilience* ($5.332 \pm 1.083$) being the largest, followed by *metacognitive resilience* ($4.861 \pm 1.234$), and *social resilience* ($4.748 \pm 1.177$) being the lowest. A plausible explanation for the discrepancy might be attributed to EFL learners themselves who intuitively tend to focus on personal appraisals first (Lazarus and Folkman, 1984), hence the highest self-report score of *ego resilience*. In other words, resilient EFL learners tend to first trust in their ability to recover from the adversities (Maltby et al., 2015), then begin to adopt positive cognitive strategies and confront with the challenging and stressful EFL learning situations (Cassidy, 2016).

## Practical implications, limitations and future directions

Some pedagogical implications could be inferred as follows. First, since the *FLLRS* has been validated, future study should examine the relationship between FL learners' individual differences (e.g., age, gender, resilience levels and other demographic variables), FL contextual antecedents (e.g., learning environments, learning protocols and other contextual variables), and FL emotional factors in positive psychology (e.g., motivation, anxiety, boredom, enjoyment, well-being, engagement and flow experience, etc.) as has been done with academic resilience in the domain-general contexts. Second, the largest path coefficient of *metacognitive resilience* warrants the need to explore the predictive effects of *ego resilience*, *metacognitive resilience* and *social resilience* in general, and *metacognitive resilience* in particular on EFL learners' learning aspects and FL performance. Such investigations are especially needed in contexts like China where FL learning is time-consuming with low efficiency (Li, 2021). To make positive adaptation to FL adversities, learners themselves should not only be perseverant in language learning, but also seek help from others and adopt some metacognitive strategies to monitor, evaluate and manage their FL learning behaviors and activities.

Despite the meaningful findings, limitations and future directions should be addressed though. First, this study only adopts a cross-sectional research design to understand the factors of FL resilience at one point in time, future research can adopt sophisticated research designs (e.g., longitudinal research design with mixed methods) to gain a better understanding of the diachronic changes of FL learners' resilience over time. Second, this study is only based on samples of tertiary education level in the EFL learning contexts, its feasibility for primary and secondary educational level in other FL learning contexts remains open for future investigations. Third, while validity of the *FLLRS* is based on the homogenous data collected from Chinese EFL learners in Double First-Class universities, it remains largely unclear whether the *FLLRS* can be generalizable to other institutional contexts, e.g., non-Double First-Class or vocational universities, etc. Future studies should adopt a more comprehensive examination regarding the generalizability of the *FLLRS* across different levels of higher education institutions. Last and importantly, future research should adopt the state-of-the-art explanatory structural equation modeling (ESEM, see Alamer, 2022a,b for excellent methodological synergies) technique that combines both the EFA and CFA into one measurement model, which is a powerful technique in testing the construct validity of the second language acquisition scales in this regard.

## 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 ethics committee of Hunan University.

The patients/participants provided their written informed consent to participate in this study.

## Author contributions

NG is responsible for the design, data collection, and drafting of the manuscript. RL is responsible for design and writing of the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

This study was funded by the Project of Hunan Social Science Achievements Appraisal Committee (grant number XSP22YBC473) and the 14th Five Year Project for Education Science in Hunan Province (grant number XJK22CGD007).

## Acknowledgments

The authors would like to thank the reviewers and editors for their constructive and insightful comments.

## 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



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1046340/full#supplementary-material

## References

Alamer, A. (2022a). Explanatory structural equation modeling in second language research: an applied example using the dualistic model of passion. *Stud. Second Lang. Acqus.*, 1–24. doi: 10.1017/S0272263121000863

Alamer, A. (2022b). Exploratory structural equation modeling (ESEM) and bifactor ESEM for construct validation purposes: guidelines and applied example. *Res. Methods Appl. Linguist.* 1, 100005–100013. doi: 10.1016/j.rmal.2022.100005

Ang, W. H. D., Shorey, S., Lopez, V., Chew, H. S. J., and Lau, Y. (2021). Generation Z undergraduate students' resilience during the COVID-19 pandemic: a qualitative study. *Curr. Psychol. Forthcoming* 41, 8132–8146. doi: 10.1007/s12144-021-01830-4

Axford, N., Jodrell, D., and Hobbs, T. (2014). "Objective or subjective well-being?" in *Handbook of Child Well-Being: Theories, Methods and Policies in Global Perspective*. eds. A. Ben-Arieh, F. Casas, I. Frønes and J. E. Korbin (New York/London: Springer), 2699–2738.

Block, J., and Kreman, A. M. (1996). IQ and ego-resiliency: conceptual and empirical connections and separateness. *J. Pers. Soc. Psychol.* 70, 349–361. doi: 10.1037/0022-3514.70.2.349

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., and Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Front. Public Health* 6, 1–18. doi: 10.3389/fpubh.2018.00149

Brewer, M. L., van Kessel, G., Sanderson, B., Naumann, F., Lane, M., Reubenson, A., et al. (2019). Resilience in higher education students: a scoping review. *High. Educ. Res. Dev.* 38, 1105–1120. doi: 10.1080/07294360.2019.1626810

Broadbent, J., and Poon, W. L. (2015). Self-regulated learning strategies & academic achievement in online higher education learning environments: a systematic review. *Internet High. Educ.* 27, 1–13. doi: 10.1016/j.iheduc.2015.04.007

Campbell-Sills, L., and Stein, M. B. (2007). Psychometric analysis and refinement of the Connor–Davidson resilience scale (CD-RISC): validation of a 10-item measure of resilience. *J. Trauma. Stress.* 20, 1019–1028. doi: 10.1002/jts.20271

Cassidy, S. (2016). The academic resilience scale (ARS-30): a new multidimensional construct measure. *Front. Psychol.* 7, 1–11. doi: 10.3389/fpsyg.2016.01787

Chan, C. W. H., Tang, F. W. K., Chow, K. M., and Wong, C. L. (2021). Enhancing generic capabilities and metacognitive awareness of first-year nursing students using active learning strategy. *BMC Nurs.* 20, 1–8. doi: 10.1186/s12912-021-00601-7

Chen, X., He, J., and Fan, X. (2020). Applicability of the ego-resilience scale (ER89) in the Chinese cultural context: a validation study. *J. Psychoeduc. Assess.* 38, 675–691. doi: 10.1177/0734282919889242

Chen, X., and Padilla, A. M. (2022). Emotions and creativity as predictors of resilience among L3 learners in the Chinese educational context. *Curr. Psychol.* 41, 406–416. doi: 10.1007/s12144-019-00581-7

Connor, K. M., and Davidson, J. R. T. (2003). Development of a new resilience scale: the Connor–Davidson resilience scale (CDRISC). *Depress. Anxiety* 18, 76–82. doi: 10.1002/da.10113

Dewaele, J. M., and MacIntyre, P. D. (2016). "Foreign language enjoyment and foreign language classroom anxiety: the right and left feet of the language learning," in *Positive Psychology in SLA*. eds. P. D. MacIntyre, T. Gregersen and S. Mercer (Bristol: Multilingual Matters), 215–236.

Fornell, C., and Larcker, D. F. (1981). Evaluating structural equation models with unobservable and measurement error. *J. Mark. Res.* 18, 39–50. doi: 10.1177/002224378101800104

Friborg, O., Hjemdal, O., Rosenvinge, J. H., and Martinussen, M. (2003). A new rating scale for adult resilience: what are the central protective resources behind healthy adjustment? *Int. J. Methods Psychiatr. Res.* 12, 65–76. doi: 10.1002/mpr.143

Gregersen, T., Mercer, S., and MacIntyre, P. D. (2021). Language teacher perspectives on stress and coping. *Foreign Lang. Ann.* 54, 1145–1163. doi: 10.1111/flan.12544

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2006). *Multivariate Data Analysis* (*6th Edn.*). New Jersey: Prentice-Hall.

Hunsu, N., Oje, A., Carnell, P., and Sochacka, N. (2021). Examining factorial validity evidence for the academic resilience scale in an engineering learning context. *Int. J. Eng. Educ.* 37, 537–549.

Jahedizadeh, S., Ghonsooly, B., and Ghanizadeh, A. (2021). A model of language students' sustained flow, personal best, buoyancy, evaluation apprehension, and academic achievement. *Porta Linguarum* 35, 257–275. doi: 10.30827/portalin.v0i35.15755

Kaiser, H. F. (1970). A second-generation little jiffy. *Psychometrika* 35, 401–415. doi: 10.1007/BF02291817

Kim, T. Y., and Kim, Y. (2021). Structural relationship between L2 learning motivation and resilience and their impact on motivated behavior and L2 proficiency. *J. Psycholinguist. Res.* 50, 417–436. doi: 10.1007/s10936-020-09721-8

Kim, T. Y., Kim, Y., and Kim, J. Y. (2018). A qualitative inquiry on EFL learning demotivation and resilience: a study of primary and secondary EFL students in South Korea. *Asia Pac. Educ. Res.* 27, 55–64. doi: 10.1007/s40299-017-0365-y

Kim, T. Y., Kim, Y., and Kim, J. Y. (2019). Role of resilience in (de)motivation and second language proficiency: cases of Korean elementary school students. *J. Psycholinguist. Res.* 48, 371–389. doi: 10.1007/s10936-018-9609-0

Lavy, S., and Ayuob, W. (2019). Teachers' sense of meaning associations with teacher performance and Ggraduates' resilience: a study of schools serving students of low socio-economic status. *Front. Psychol.* 10, 1–10. doi: 10.3389/fpsyg.2019.00823

Lazarus, R. S., and Folkman, S. (1984). *Stress, Appraisal, and Coping*. New York/London: Springer Publish Company.

Li, R. (2021). Does game-based vocabulary learning APP influence Chinese EFL learners' vocabulary achievement, motivation and self-confidence? *SAGE Open* 11, 215824402110030–215824402110012. doi: 10.1177/21582440211003092

Li, R. (2022a). Foreign language reading anxiety and its correlates: a meta-analysis. *Read. Writ.* 35, 995–1018. doi: 10.1007/s11145-021-10213-x

Li, R. (2022b). Examining foreign language listening anxiety and its correlates: a meta-analysis. *Porta Linguarum* 38, 99–117. doi: 10.30827/portalin.vi38.22173

Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., and Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Comput. Assist. Lang. Learn.* 32, 784–804. doi: 10.1080/09588221.2018.1540433

Li, R., Meng, Z., Tian, M., Zhang, Z., and Xiao, W. (2021). Modelling Chinese EFL learners' flow experiences in digital game-based vocabulary learning: the roles of learner and contextual factors. *Comput. Assist. Lang. Learn.* 34, 483–505. doi: 10.1080/09588221.2019.1619585

Liebenberg, L., Ungar, M., and Vijver, F. V. D. (2011). Validation of the child and youth resilience Measure-28 (CYRM-28) among Canadian youth. *Res. Soc. Work. Pract.* 22, 219–226. doi: 10.1177/1049731511428619

Lou, N. M., and Noels, K. A. (2020a). Mindsets matter for linguistic minority students: growth mindsets foster greater perceived proficiency, especially for newcomers. *Mod. Lang. J.* 104, 739–756. doi: 10.1111/modl.12669

Lou, N. M., and Noels, K. A. (2020b). Breaking the vicious cycle of language anxiety: growth language mindsets improve lower-competence ESL students' intercultural interactions. *Contemp. Educ. Psychol.* 61, 1–17. doi: 10.1016/j.cedpsych.2020.101847

Maltby, J., Day, L., and Hall, S. (2015). Refining trait resilience: identifying engineering, ecological, and adaptive facets from extant measures of resilience. *PLoS One* 10, 1–27. doi: 10.1371/journal.pone.0131826

Mansfield, C. F., Beltman, S., Broadley, T., and Weatherby-Fell, N. (2016). Building resilience in teacher education: an evidenced informed framework. *Teach. Teach. Educ.* 54, 77–87. doi: 10.1016/j.tate.2015.11.016

Martin, A. J., and Marsh, H. W. (2008a). Academic buoyancy: towards an understanding of students' everyday academic resilience. *J. Sch. Psychol.* 46, 53–83. doi: 10.1016/j.jsp.2007.01.002

Martin, A. J., and Marsh, H. W. (2008b). Workplace and academic buoyancy. *J. Psychoeduc. Assess.* 26, 168–184. doi: 10.1177/0734282907313767

Martin, A. J., and Marsh, H. W. (2009). Academic resilience and academic buoyancy: multidimensional and hierarchical conceptual framing of causes, correlates and cognate constructs. *Oxf. Rev. Educ.* 35, 353–370. doi: 10.1080/03054980902934639

Nguyen, K., Stanley, N., Stanley, L., Wang, Y., and Gritter, K. (2015). Resilience in language learners and the relationship to storytelling. *Cogent Educ.* 2, 1–16. doi: 10.1080/2331186X.2014.991160

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

Seçer, S., and Ulas, S. (2020). The mediator role of academic resilience in the relationship of anxiety sensitivity, social and adaptive functioning, and school refusal with school attachment in high school students. *Front. Psychol.* 11, 1–12. doi: 10.3389/fpsyg.2020.00557

Shin, W. Y., Kim, M. G., and Kim, J. H. (2009). Developing measures of resilience for Korean adolescents and testing cross, convergent, and discriminant validity. *Stud. Korean Youth* 20, 105–131.

Sudina, E., and Plonsky, L. (2021). Academic perseverance in foreign language learning: an investigation of language-specific grit and its conceptual correlates. *Mod. Lang. J.* 105, 829–857. doi: 10.1111/modl.12738

Teimouri, Y. (2018). Differential roles of shame and guilt in L2 learning: how bad is bad? *Mod. Lang. J.* 102, 632–652. doi: 10.1111/modl.12511

Trigueros, R., Padilla, A., Aguilar-Parra, J. M., Mercader, I., Lopez-Liria, R., and Rocamora, P. (2020). The influence of transformational teacher leadership on academic motivation and resilience, burnout and academic performance. *Int. J. Environ. Res. Public Health* 17, 1–12. doi: 10.3390/ijerph17207687

Ungar, M. (2008). Resilience across cultures. *Br. J. Soc. Work.* 38, 218–235. doi: 10.1093/bjsw/bcl343

Ungar, M. (2019). Designing resilience research: using multiple methods to investigate risk exposure, promotive and protective processes, and contextually relevant outcomes for children and youth. *Child Abuse Negl.* 96:104098. doi: 10.1016/j.chiabu.2019.104098

Ungar, M., Theron, L., Murphy, K., and Jefferies, P. (2021). Researching multisystemic resilience: a sample methodology. *Front. Psychol.* 11, 1–18. doi: 10.3389/fpsyg.2020.607994

van der Meer, C. A. I., Brake, H. T., van der Aa, N., Dashtgard, P., Bakker, A., and Olff, M. (2018). Assessing psychological resilience: development and psychometric properties of the English and Dutch version of the resilience evaluation scale (RES). *Front. Psychol.* 9, 1–11. doi: 10.3389/fpsyt.2018.00169

Wagnild, G. M., and Young, H. M. (1993). Development and psychometric evaluation of the resilience scale. *J. Nurs. Meas.* 1, 165–178. PMID: 7850498

Yun, S., Hiver, P., and Al-Hoorie, A. H. (2018). Academic buoyancy: exploring learners' everyday resilience in the language classroom. *Stud. Second. Lang. Acquis.* 40, 805–830. doi: 10.1017/S0272263118000037

Zhang, Y., Aramburo, C., and Tsang, K. K. (2021). Measuring teacher recognition through Honneth's recognition theory: development and validation of a teacher recognition scale. *Asia Pac. Educ. Res.* doi: 10.1007/s40299-021-00637-y

Zhang, M. X., Yang, H. M., Tong, K. K., and Wu, A. M. S. (2020). The prospective effect of purpose in life on gambling disorder and psychological flourishing among university students. *J. Behav. Addictions* 9, 756–765. doi: 10.1556/2006.2020.00046

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 21 December 2022
DOI 10.3389/fpsyg.2022.1071907



OPEN ACCESS

EDITED BY
Jinfen Xu,
Huazhong University of Science and Technology, China

REVIEWED BY
Barry Bai,
The Chinese University of Hong Kong, China
Raquel Fernández-Cézar,
University of Castilla-La Mancha, Spain

*CORRESPONDENCE
Yanling Xiao
991835@hainanu.edu.cn

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

RECEIVED 17 October 2022
ACCEPTED 17 November 2022
PUBLISHED 21 December 2022

CITATION
Qin C, Zhang R and Xiao Y (2022) A questionnaire-based validation of metacognitive strategies in writing and their predictive effects on the writing performance of English as foreign language student writers.
*Front. Psychol.* 13:1071907.
doi: 10.3389/fpsyg.2022.1071907



# A questionnaire-based validation of metacognitive strategies in writing and their predictive effects on the writing performance of English as foreign language student writers

Chenghai Qin, Ruru Zhang and Yanling Xiao*

College of Foreign Languages, Hainan University, Haikou, China

**Introduction:** This study—drawing upon data from a questionnaire—examined 503 Chinese university students' metacognitive strategies in writing (MSW). The focus was on Chinese student writers who are learning English as a foreign language (EFL).

**Methods:** The examination was conducted through a survey on MSW and a writing test administered at the end of the semester. We employed exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) for data analysis. Multiple regression analysis was also adopted for understanding the predictive effects of strategies on writing performance.

**Results:** The findings provided validity to MSW, including person, task, strategies, planning, monitoring, and evaluating. The different components of MSW were reported to significantly affect the participants' writing performance. The findings highlight that EFL student writers were aware of metacognitive writing strategies. The MSW survey could be used to assess EFL students' metacognitive writing strategies and develop curricula in writing strategy training.

**Conclusion:** Writing instruction can direct learners' ability to acquire metacognitive writing strategies, particularly those of planning, monitoring, and evaluating, to build their awareness as agents in EFL writing. Relevant pedagogical implications are discussed.

KEYWORDS

**metacognitive knowledge, writing, metacognitive awareness, metacognitive regulation, self-regulation**

## Introduction

Metacognitive strategies are essential to the process of learning to write when learning English as a foreign language (EFL; Nguyen and Gu, 2013; Teng, 2016, 2019; Teng and Yue,2022). However, in the Chinese EFL context, for which English writing instruction typically emphasizes grammatical correctness rather than idea development, learners may find it difficult to build an awareness of using metacognitive writing strategies (Ruan, 2014).

Through a mixed-methods study, Amani (2014) found that explicit metacognitive strategy instruction had a positive impact on the writing competence of L2 writing students. However, in terms of EFL writing, university EFL students may find it challenging because of their lack of awareness of metacognitive writing strategies (Teng, 2019). In addition, EFL learners in the Chinese context receive limited English language input, making it more challenging to learn to write. Student writers are expected to have repertoires of strategies when learning to write (Raimes, 1987). In particular, they need to build an advanced level of "self-initiated thoughts, feelings, and actions" for them to "attain various literary goals" (Zimmerman and Risemberg, 1997, p.76). Hence, metacognitive writing strategies are essential to possible improvements in EFL writing.

Nevertheless, even though students are taught how to plan, monitor, and evaluate their own writing, students may know little about themselves as writers (Leung and Hicks, 2014). They may also not recognize their own writing strengths or weaknesses, tending to overemphasize the latter and overlook any progress they have made or can make in their writing (Teng, 2016). Wenden (1998) argued that metacognitive knowledge is a prerequisite for self-regulation, and metacognitive knowledge is essential to learner autonomy because it "informs planning decisions taken at the outset of learning and the monitoring processes that regulate the completion of a learning task and decisions to remediate; it also provides the criteria for evaluation made once a learning task is completed" (p. 528). Teng and Zhang (2021) argued that there is a dynamic and longitudinal relationship between metacognitive knowledge and reading and writing in a foreign language context. However, teachers may not recognize the importance of metacognitive knowledge in Chinese EFL writing contexts, wherein teaching academic writing is product oriented (Teng and Zhang, 2016). The student writers were passive and found it difficult to keep positive beliefs in writing (Bruning and Horn, 2000). This may be related to learners' lack of awareness of self-regulation in writing. They may exert more effort learning vocabulary knowledge and grammar for writing, rather than being an agent for writing (Graham and Harris, 2000). Student writers need self-awareness, motivation, and positive behavioral skills for writing (Zimmerman, 2002, p.65–66). Metacognitive writing strategies are thus essential to EFL students' writing performance.

Self-regulation principles, measurements, and practices have a solid ground for enriching second and foreign language learning and teaching (Teng and Zhang, 2022). Through a socio-cognitive approach to writing, Nishino and Atkinson (2015) argued that writing is primarily a cognitive activity and that cognition plays a vital role in writing and its development. To help students become competent English writers and autonomous learners, instructors need to support their development of metacognitive strategies. However, scarce attention was paid to writing strategies from the perspective of metacognition, particularly for low-achieving students in the EFL context. The present study examined Chinese university EFL students' metacognitive strategies in EFL writing. We aim for the following purposes: (a) to assess the reliability of a new scale, which we named it as metacognitive strategies in writing (MSW) and (b) to explore how different components of MSW predict EFL students' writing performance. The findings are insightful in helping researchers and classroom practitioners to diagnose the needs of metacognitive strategies in writing and develop guidelines for instructing writing courses for university EFL students. The findings shed lights on how to teach EFL writing and deliver more effective program for writing teacher preparation.

# Literature review

## Language learning strategies

Oxford (1990) classified a list of language learning strategies based on cognitive learning theory. These strategies include memory, cognitive, compensatory, affective, social, and metacognitive strategies. Past studies have documented differences in strategy use between more and less successful learners. For example, successful learners use these strategies in larger numbers and at higher frequencies (Magogwe and Oliver, 2007). Most importantly, cognitive and metacognitive strategies are associated with a higher level of language proficiency (Peacock and Ho, 2003). However, contradictory findings were also reported, showing that less successful learners used more strategies than more successful learners did because the former automatized their language learning process (Oxford and Cohen, 1992). Another point worth noting is that unsuccessful learners may adopt a large number of strategies frequently, but it does not necessarily mean that they are able to identify appropriate strategy use. In fact, it was reported that successful learners were able to identify appropriate strategies depending on the task requirements, but unsuccessful learners failed to choose the most appropriate and efficient strategies during the task (Chamot and El-Dinary, 1999).

Although ample research has been reported relating to learners' proficiency level and strategy use, learner variables, such as cultural background and national origin, could have a strong influence on learners' strategy use (Oxford and Nyikos, 1989). Therefore, their findings might not be generalizable to learners with completely different cultural backgrounds. In light of this, Lai (2009) conducted a questionnaire survey that investigated the relationships between the language learning strategies used by 418 EFL learners in Taiwan based on learners' language proficiency and their use of strategies. While the more proficient learners used metacognitive strategies and cognitive strategies most frequently and memory strategies least frequently, the less proficient learners preferred social and memory strategies to cognitive and metacognitive strategies. This finding partially echoes Wu (2008), who reported that higher-proficiency EFL students in Taiwan used learning strategies more often than lower-proficiency EFL students did, especially the cognitive, metacognitive and social strategies.

Although research documented in the literature examines general language learning strategy use, it is possible that these

summarized findings could serve as a reference for the specific examination of metacognitive strategy use during English writing.

## Understanding metacognition

Metacognition is multidimensional and domain-general. When we talk about metacognition, we may need to mention the theory of mind (Flavell, 1979). Such theory is the foundation of understanding metacognition. Generally, metacognition is related to self-regulatory capacity because metacognition provides individuals with domain knowledge and regulatory skills that are essential to become an agentive learner in relevant domains (Schraw, 2001, p. 7). Metacognition refers to how learners build an awareness of their own thinking processes and executive processes (Flavell, 1979). Metacognition is essential to helping learners regulate their cognitive processes, and finally, becoming an independent thinker and learner. Zhang and Zhang (2019) applied metacognition in second and foreign language learning, and posited that EFL learners need to plan, monitor, and evaluate their cognitive processes for better language learning performance.

Metacognition includes metacognitive knowledge and metacognitive regulation. Flavell (1985) suggested that person, task, and strategy knowledge are three key elements of metacognitive knowledge. Wenden (1998) explained the three elements. For example, person knowledge is the knowledge for the learners to control their cognitive processes. Task knowledge is the knowledge that can be helpful for the learners to understand the purpose, nature, and demands of different task conditions. Strategy knowledge is the knowledge of different important strategies that are helpful for realizing the pre-determined goals. Metacognitive regulation entails three skills: planning, monitoring, and evaluating (Schraw, 1998). Planning refers to the ability to appropriately select the strategies and adequately allocate the resources for completing tasks. Monitoring refers to learners' capacity to observe their task performance. Evaluating means learners' capacity to reflect on their learning outcome and the use of different strategies for self-regulation.

Teng et al. (2022) summarized the procedures of understanding metacognition. First, monitoring function and control of cognition are two important functions of metacognition. In order to realize the functions, individuals need to process three major stages, i.e., acquisition, retention, and retrieval. Second, learners need metacognitive knowledge and metacognitive experiences to process the monitoring function. In contrast, they need metacognitive strategies or metacognitive skills to fulfill the needs of control of cognition. Third, metacognitive knowledge, metacognitive experiences, and metacognitive skills are interconnected with each other. Metacognitive knowledge includes person, task, and strategies. Metacognitive experiences include feelings and judgments. Metacognitive skills are important for their metacognitive regulation, which needs learners to plan, monitor, and evaluate their learning process. Finally, reflection is the outcome of the interconnected process of planning, monitoring, and evaluating (Figure 1).

## Metacognitive strategies in EFL writing

Macaro (2010) maintains that strategic behavior plays a vital role in second language learning success and proposes that strategic behavior should be essential to linguistic knowledge resources. Dornyei (2010) emphasizes that students need a repertoire of appropriate task-related plans, scripts, and self-regulatory strategies that are activated by their ideal L2 selves; that is, learners' aptitude, motivation, goals, and self-regulatory strategies all interact and affect one another in the SLA process. Writing strategies include rhetorical strategies, metacognitive strategies, cognitive strategies, and social/affective strategies (Wenden, 1991; Riazi, 1997). Writers explore rhetorical strategies to organize and present their ideas based on the writing conventions of the target language. Metacognitive strategies are used to monitor the writing process consciously and evaluate the effectiveness of writing actions. Cognitive strategies are used to implement actual writing actions. Social/affective strategies are employed to interact with others and to regulate emotions, motivation, and attitudes in writing.

Wenden (1991) classifies writing strategies based on metacognitive and cognitive frameworks. She distinguishes general executive metacognitive strategies of planning, self-monitoring, and self-evaluating from more specific cognitive strategies, such as clarification, retrieval, resourcing, avoidance, and verification. Each of these metacognitive strategies is discussed below.

Planning for writing involves thinking and self-questioning strategies such as identifying one's purpose, activating background knowledge, and organizing ideas. Planning is not limited to a specific stage of writing but rather appears recursively throughout the writing process. Flower and Hayes (1981) identified three different types of planning strategies based on the focus of the goal: (1) generating ideas; (2) setting procedural goals; and (3) organizing. Generating ideas includes retrieving information from long-term memory, revising old ideas to incorporate new information, drawing inferences, making connections, and looking for examples, contradictions, and objections. Setting procedural goals includes content goals (e.g., plans for content, text structure and audience, and criteria for evaluation) and process goals (how to proceed, generated by the writer, done at any time during the composing process, followed or preceded by generating ideas, revising strategies, etc.). The third strategy (organizing) includes selecting the most useful materials produced during the generating process and organizing them in the writing plan. Organizing strategies include grouping and sequencing ideas, deciding on the presentation of the text, planning the introduction and conclusions, and structuring the text based on a particular genre. Furthermore, in using these strategies, it is essential to consider the audience, topic, and rhetorical knowledge. Planning in EFL writing determines how writers write in subsequent stages. It engages them in metacognitive activities that allow them to consider the purpose and goals for writing, identify

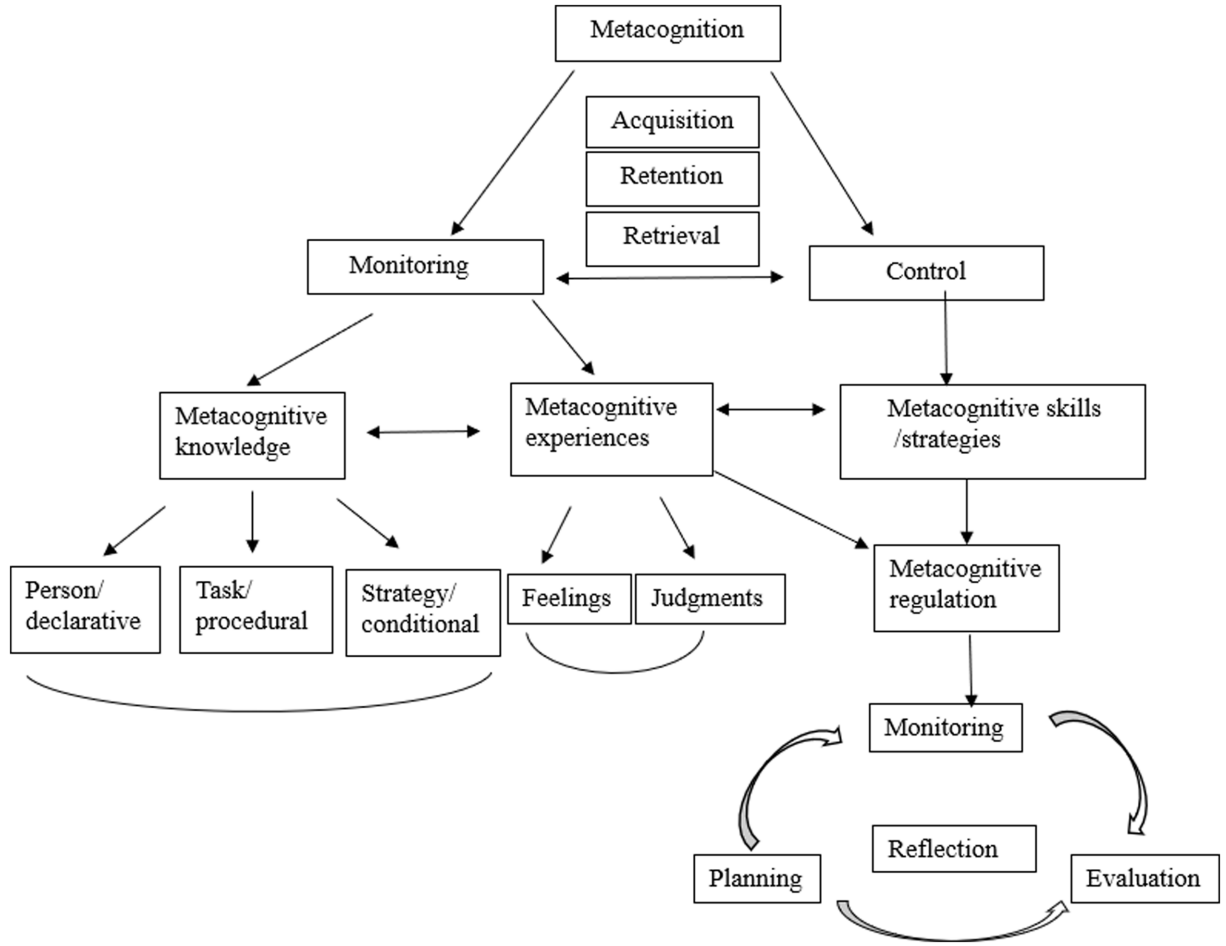



FIGURE 1
The multifaceted elements of metacognition (Teng et al., 2022, p. 171).

their audience, decide upon voice, and generate a framework for their essays.

Monitoring involves conscious control and regulation of the writing process. Hayes and Flower (1980) include self-monitoring in their model of the cognitive processes of writing, noting that the ability to self-monitor the composing process is an important part of writing strategies. Charles (1990) claims that self-monitoring makes it easier for L2 students to avoid uncertainty about any part of their text, to find direct answers to their queries and to encourage them "to look critically and analytically at their writing and to place themselves in the position of readers" (p. 289). The more important functions of self-monitoring are controlling, directing, and sequencing the composing processes and one's progress in the task. Monitoring allows the writer to decide whether something needs to be retrieved, whether new ideas need to be further generated, or whether a given subprocess has ended. Monitoring allows L2 writers to evaluate the effectiveness of writing strategies and how and when to check the outcomes of problem-solving processes and strategically regulate the processes according to cognitive goals (Mayer, 1999).

Self-evaluating—experiencing the quality of one's writing in relation to one's goals—is crucial for developing an individual's perception of writing. In self-evaluation, students can recognize weaknesses, identify needs, and make changes (Zimmerman, 2002). In cognitive research, evaluation has been characterized as a strategy for considering the outcome of the undertaken task, an essential metacognitive strategy that successful learners need to execute and control.

## Empirical studies on the use of metacognitive writing strategies

Various studies have been conducted on EFL students' use of metacognitive writing strategies. Employing think-aloud protocols and immediate retrospective interviews, Chien (2012) investigated the differences in writing strategies and English writing achievements of 20 low-achieving and 20 high-achieving student writers in Taiwan. Chien found that high-achieving student writers were more aware of and focused more on, formulating

their position statements when planning, generating, revising, and editing their essays and focused more on correcting grammatical and spelling errors. Teng and Zhang (2016) validated questionnaire-based self-regulated strategies in EFL writing and highlighted planning, monitoring, and evaluating in EFL writing. Teng and Huang (2019) also suggested that learners' self-regulated strategies in writing, as well as their English proficiency and language learning experiences, and significantly influenced their EFL writing. In a recent publication (Teng et al., 2022), two experimental studies were reported. Study 1 adopted a factorial design using exploratory and confirmatory factor analysis to validate a self-regulatory writing strategy questionnaire. Study 2 assessed the predictive effects of the different components of the scale on students' writing performance. The results supported the construct validity for the six strategy factors, i.e., writing planning, goal-oriented monitoring, goal-oriented evaluation, emotional control, memorization, and metacognitive judgment. The factors also predicted writing performance. Zhang and Qin (2018) also validated the newly developed scale on metacognitive strategies in a multimedia writing context. The results provided evidence for the validation of planning, monitoring, and evaluating strategies. In an early empirical study on the importance of planning in EFL writing, Graham et al. (1995) examined differences between expert and less-skilled L2 writers. They found that expert L2 writers spent considerable time planning and appeared to have higher-level plans and self-conscious control of their planning. In contrast, less-skilled EFL writers were less likely to use knowledge of textual structure in planning, to use heuristic strategies in searching their memory for content, or to establish goals to direct the writing process and were more likely to engage in "knowledge telling" (i.e., writing everything they knew about a topic and stopping when they felt that they had written down everything they knew). Less-skilled writers did not write with goals or plans in mind; rather, they tended to generate ideas through free writing and usually did not organize those ideas. As shown in a longitudinal study (Teng and Zhang, 2021), learners' L2 writing development was dependent on their initial level of metacognitive knowledge. This is evidence for the strong correlation between metacognitive knowledge and writing.

Nguyen and Gu (2013) explored the impact of strategy-based instruction on promoting learner autonomy (operationally defined as learner self-initiation and learner self-regulation) of students at a Vietnamese university; 37 students were in an experimental group, and 54 students were in two control groups. After an 8-week metacognition training intervention, students in the experimental group were found to have improved their planning, monitoring, and evaluating of a writing task more than those in the two control groups. The findings suggest that strategy-based instruction on task-specific metacognitive self-regulation improves learner autonomy and writing performance. Teng (2020) also incorporated training of metacognitive strategies for EFL learners. There were two groups of learners, i.e., those with group feedback guidance and those with self-explanation guidance. The results supported the positive effects of group metacognitive support on EFL students' writing. EFL students need to build a certain level of metacognitive awareness to manage themselves as writers.

Bai et al. (2014) conducted a questionnaire survey to explore the relationship between 1,618 Singapore primary school pupils' reported use of strategies in learning to write and the correlation with their English language proficiency. They found that participants used a wide range of writing strategies at medium frequency. They also reported a significant correlation between the participants' English language proficiency and the use of writing strategies such as planning, text-generating, revising, monitoring and evaluating, and resourcing. Similar results were also found in Bai and Guo (2021), wherein high achievers reported higher levels of motivation (i.e., growth mindset, self-efficacy, and interest) and self-regulated learning strategy use than the average achievers, and average achievers reported more strategy use than the low achievers, Ma and Teng (2021) collected qualitative data from two undergraduate university students learning English as L2 in Hong Kong to explore their use of writing strategies. They reported that both students realized the importance of self-evaluation and revision. It seems that the students perceived affordances in the kind of writing that enabled them to play an active role in seeking, interpreting, and using teacher feedback to perform the evaluation and modification of their own work. However, variations in engagement in the process of learning to write and their metacognitive knowledge development were also detected. For example, students' varying degrees of engagement may result in various degrees of developing metacognitive awareness. Teng et al. (2022) validated a new instrument, i.e., the Metacognitive Academic Writing Strategies Questionnaire (MAWSQ). Analyses were conducted through a series of Confirmatory factor analyses (CFA). Results supported two hypothesized models, i.e., an eight-factor correlated model and a one-factor second-order model. Model comparisons supported the role of metacognition as a higher-order construct. Metacognition also explains the eight metacognitive strategies, including declarative knowledge, procedural knowledge, conditional knowledge, planning, monitoring, evaluating, information management, and debugging strategies. Those strategies also significantly influenced EFL writing performance.

Overall, the studies on metacognition development reviewed in this section highlight the importance of the high-level cognitive processes involved in composing, the development of the autonomous and self-regulated use of effective writing strategies, and the formation of positive attitudes about writing. Metacognitively oriented learners are aware of both their own learner characteristics and the writing task and are able to select, employ, monitor, and evaluate their use of metacognitive strategies.

## The present study

Metacognition functions as an important predictor in EFL writing performance. We aim for two purposes in the present

study. First, we attempted to validate a questionnaire on metacognitive strategies in writing. Second, we assessed the predictive effects of different metacognitive strategies in the outcome EFL writing. The present study sheds light on learners' awareness and use of metacognitive writing strategies. The present study includes two questions:

1. What is the evidence to support the validity and reliability of metacognitive strategies in writing?
2. What is the evidence for the predictive effects of metacognitive strategies on EFL writing proficiency?

# Materials and methods

## Participants

The present study included 503 participants. They were undergraduate students at a university in China. They were first-year students with Chinese as their first language and English as a foreign language. They had received at least 6 years of formal English instruction. Writing is a subject to be taught in college English and a compulsory course for all the participants. We selected the participants because they were all enrolled in a university English course. The first author was teaching the participants, and the sample of participants was a convenient sample. Among the 503 students, 351 were men and 152 were women. An unequal gender balance may be because most of the students were from science and engineering majors. Originally, there were 700 students who responded to the questionnaire. We finally selected data from 503 students for data analysis. Some participants' data were excluded because of missing values or because some were unable to take the writing test. They attended the study voluntarily by signing the consent form.

## Questionnaire development

The questionnaire, which was named Metacognitive Strategies in Writing (MSW), was developed through item generation, reference consultation, initial piloting, psychometric evaluation, and exploratory factor analysis (EFA) in a pilot study. We first invited 10 students to reflect on their writing practices and strategies. The students were mainly interviewed about the strategies they adopted for writing. We generated approximately 50 items based on analyzing the transcriptions of learners' interviews. In the next stage, we consulted relevant literature on metacognition, self-regulation, and language learning strategies (Schraw and Dennison, 1994; Oxford, 2013; Teng et al., 2022). We selected the items that fit with metacognition theories. In the third stage, we invited the 10 students to check the items. In the fourth stage, which was psychometric evaluation, we invited two researchers in L2 writing to assess the items. Based on the comments, we finally removed 10 items. In the final stage, we ran an EFA with a sample of 360 students with similar backgrounds. We deleted 10 items with unsatisfactory factor loading values. The final questionnaire includes 30 items, which are in the Appendix.

This questionnaire was a novel one as it was based on metacognition theory, through which the focus was on understanding metacognitive knowledge and regulation in learning to write. We adopted a seven-point Likert scale (i.e., from 1, Strongly disagree to 7, Strongly agree). MSW focuses on metacognitive knowledge and metacognitive regulation. Metacognitive knowledge includes three factors, i.e., person, task, and strategies. Metacognitive regulation includes three factors: planning, monitoring, and evaluating. Cronbach's alpha, which ranged from 0.81 to 0.90 for the six factors, ensured the internal consistency of responses to the items. The questionnaires were administered to the participants in Chinese. The author translated into Chinese while a research assistant was invited to check the translated items through back translation.

## Writing test

A writing test from IELTS (writing task 2) was adopted to measure learners' writing proficiency. Students were required to write at least 250 words within 1 h. Students were asked to respond to the topic provided by giving and justifying an opinion, discussing the topic, summarizing details, outlining problems, identifying possible solutions and supporting what they wrote with reasons, arguments and relevant examples. The topic proposed the possible influence of social media sites on personal relationships.

The marking scheme was consistent with the writing rubrics in IELTS. However, we adjusted it to fit with our school assessment needs. Each learner was awarded with six marks for task response, coherence and cohesion, lexical resource, and grammatical range and accuracy. The maximum possible score was 24 points. A total of 40 English teachers were paid to rate the writing. The teachers did not know the participants' identities. They also joined a training session on the marking scheme. Disagreements on marking were subject to further discussion. The Cronbach's alpha for the test was.85, indicating acceptable reliability.

## Procedures

We invited 20 EFL teachers to help us distribute a QR code to the students through WeChat group. The students spent an average of 6 min completing the questionnaire. The writing test was administered as an exercise for all students during class. They needed to complete it within 1 h. The format for the writing test was a paper-and-pencil format. All participants received the same format for the questionnaire and the writing test.

10.3389/fpsyg.2022.1071907

TABLE 1 Means, standard deviations. and normality test.

| Strategy category | Mean | S. D. | Skewness | Kurtosis |
|---|---|---|---|---|
| Person | 3.346 | 1.014 | 0.914 | 0.182 |
| Task | 3.620 | 0.782 | 0.225 | −0.666 |
| Strategy | 3.762 | 0.786 | 0.513 | −0.033 |
| Planning | 3.976 | 0.931 | 0.367 | −0.822 |
| Monitoring | 4.075 | 0.841 | 0.359 | −0.633 |
| Evaluating | 4.079 | 0.840 | 0.330 | −0.516 |

TABLE 2 Extraction results for the six factors.

| Factors | Eigen value (Rotated) | % of Variance (Rotated) | Cumulative % of Variance (Rotated) |
|---|---|---|---|
| 1 | 5.023 | 11.232 | 13.202 |
| 2 | 5.002 | 11.342 | 21.217 |
| 3 | 4.532 | 10.543 | 32.832 |
| 4 | 3.732 | 9.643 | 41.322 |
| 5 | 3.122 | 7.243 | 47.655 |
| 6 | 2.933 | 6.821 | 57.411 |

## Data analysis

The final dataset was run through a series of confirmatory factor analyses (CFAs). STATA was used for data analysis. CFA is used to test a theoretical model by confirming factors, correlations, covariance patterns, and residual or error values within a data matrix (Byrne, 2016). We used the maximum likelihood (ML) estimation method. The model fit was evaluated through the following statistics: a chi-square statistic, the degrees of freedom (df), $p$ value, the ratio of chi-square $\chi^2$ divided by the df, the root mean square error of approximation (RMSEA), the standardized root mean square residual (SRMR), the comparative fit index (CFI), and the Tucker–Lewis Index (TLI; DiStefano and Hess, 2005). The following criteria are a relatively good fit between the hypothesized model and the observed data: the value of RMSEA should be close to 0.06, the value of SRMR should be close to 0.08, and the values for CFI and TLI should be close to 0.95 (Hu and Bentler, 1999). Finally, multiple regression analysis was adopted to evaluate the predictive effects of MSW on students' writing proficiency.

# Results

## Descriptive statistics

The kurtosis and skewness values for the metacognitive strategies in writing, as well as the mean and standard deviation, are shown in Table 1. The means of the six factors ranged from 3.346 to 4.079, with the two factors, monitoring and evaluating, greater than 4. There were no noticeable variations based on the standard deviation values.

## Exploratory factor analysis in the pilot study

Exploratory factor analysis was conducted on a sample of 360 learners from similar background in the pilot study. We examined the adequacy of the sample. The Kaiser-Meyer-Olkin value was 0.914, which appropriate for EFA (Tabachnick and Fidell, 2001). Bartlett's test of sphericity was significant, $p < 0.001$; thus, the matrix was adequate for factor analysis. We adopted principal component analysis as a factor extraction method. We finally extracted six factors that explained 57.411% of the variance (Table 2). The scree plot showed a considerable drop after the sixth factor, for which we excluded other possible factors. Based on key theories in metacognition, we named the six factors as following: person, task, strategies, planning, monitoring, and evaluating.

The six factors' eigenvalues exceeded 1. The next step was to examine the factor loadings. We deleted 10 items with factor loadings lower than 0.4. The final version included 30 items across six factors (Table 3). Items' factor loadings ranged from 0.534 to 0.772, while communality ranged from 0.531 to 0.754. The items hence fit their respective factors well.

## Construct validity of metacognitive strategies in writing through CFA

The data fitness metrics for metacognitive strategies in writing are displayed in Table 4. Table 4 shows that the RMSEA was 0.073, less than 0.08, indicating a good fit; CFI, TLI, CNFI, IFI, and GFI all exceeded 0.9, which was ideal for adaptability. Although the $\chi^2$/df was 7.916, larger than 3, the scale on metacognitive strategies in writing still showed reliability when taken as a whole.

According to Figure 2 and Table 5, the factor loadings for Person, Task, Strategy, Planning and Evaluating were all greater than 0.5, while Monitoring was 0.41. Additionally, the average variance extracted (AVE) for each variable was 0.47, and the model's convergent validity was good, as evidenced by the composite reliability (CR) being 0.84, indicating that the model had satisfactory convergent validity.

## Predictive effect of metacognitive strategies in writing on EFL writing

Figure 3 presents the correlations between metacognitive strategies in writing and L2 learners' writing proficiency in English. The findings indicated that each of the six metacognitive strategies was significantly correlated with learners' English writing performance. Writing performance (WP) was correlated with Person ($r = 0.264$), Task ($r = 0.500$),

TABLE 3 Results on factor loadings and the communality.

| Items | Person | Task | Strategies | Planning | Monitoring | Evaluating | Communality |
|---|---|---|---|---|---|---|---|
| 18 | 0.622 | | | | | | 0.571 |
| 4 | 0.632 | | | | | | 0.692 |
| 5 | 0.622 | | | | | | 0.621 |
| 19 | 0.511 | | | | | | 0.632 |
| 1 | 0.522 | | | | | | 0.522 |
| 15 | 0.513 | | | | | | 0.592 |
| 6 | | 0.523 | | | | | 0.632 |
| 11 | | 0.532 | | | | | 0.692 |
| 9 | | 0.523 | | | | | 0.536 |
| 24 | | 0.623 | | | | | 0.665 |
| 22 | | 0.542 | | | | | 0.534 |
| 27 | | 0.643 | | | | | 0.634 |
| 38 | | | 0.523 | | | | 0.534 |
| 36 | | | 0.611 | | | | 0.534 |
| 28 | | | 0.612 | | | | 0.643 |
| 29 | | | 0.645 | | | | 0.794 |
| 21 | | | 0.564 | | | | 0.743 |
| 8 | | | 0.711 | | | | 0.734 |
| 35 | | | | 0.611 | | | 0.634 |
| 37 | | | | 0.622 | | | 0.663 |
| 33 | | | | 0.564 | | | 0.525 |
| 31 | | | | 0.543 | | | 0.531 |
| 32 | | | | | 0.634 | | 0.623 |
| 7 | | | | | 0.632 | | 0.623 |
| 3 | | | | | 0.503 | | 0.623 |
| 34 | | | | | 0.532 | | 0.636 |
| 2 | | | | | | 0.612 | 0.662 |
| 26 | | | | | | 0.732 | 0.742 |
| 13 | | | | | | 0.602 | 0.623 |
| 30 | | | | | | 0.732 | 0.772 |

TABLE 4 Model fit indices for metacognitive writing strategies.

| $\chi^2$/df | RMSEA | CFI | TLI | NFI | IFI | GFI |
|---|---|---|---|---|---|---|
| 7.916 | 0.073 | 0.946 | 0.909 | 0.939 | 0.946 | 0.957 |

Planning ($r = 0.584$), and Monitoring ($r = 0.408$). Strategy ($r = 0.470$) and Evaluating ($r = 0.470$) were significantly correlated with WP.

Moreover, we adopted a structural equation model to investigate the degree to which metacognitive strategies in writing predicted learners' L2 writing proficiency. Table 6 presents the model fitness indices. For our model, seven indices (i.e., $\chi^2$/df, RMSEA, CFI, TLI, NFI, WIFI, and GFI) indicated acceptable model fit (Table 6). Figure 4 shows a structural equation model of the relationship between metacognitive strategies in writing and writing proficiency. The six variables on the left side of the model represent the six factors of metacognitive strategies in writing. The only rectangular variable on the right side of the model was EFL learners' writing proficiency. The findings demonstrated that metacognitive strategies in writing had a predictive power of 0.65 for L2 learners' writing proficiency, indicating that it could account for 65% of the variances in writing performance.

Regression analysis was employed in the study to show the extent to which each factor impacts writing performance. The results presented in Table 7 demonstrate that all factors significantly predicted writing competence ($p < 0.001$), with the exception of Strategy ($p = 0.344$). Planning had the greatest effect on writing abilities, and Task had the least effect. Notably, monitoring and evaluating also had a great effect on EFL learners' writing proficiency. According to the findings, there was no multicollinearity among the strategies, as indicated by the variance inflation factor (VIF), which was less than 3. In addition, the residuals adhered to a normal distribution, as shown in Figure 5. This offered a trustworthy foundation for the regression analysis results.

## Discussion and conclusion

Overall, the present study aims to answer two research questions. The first research question entails the validation of a

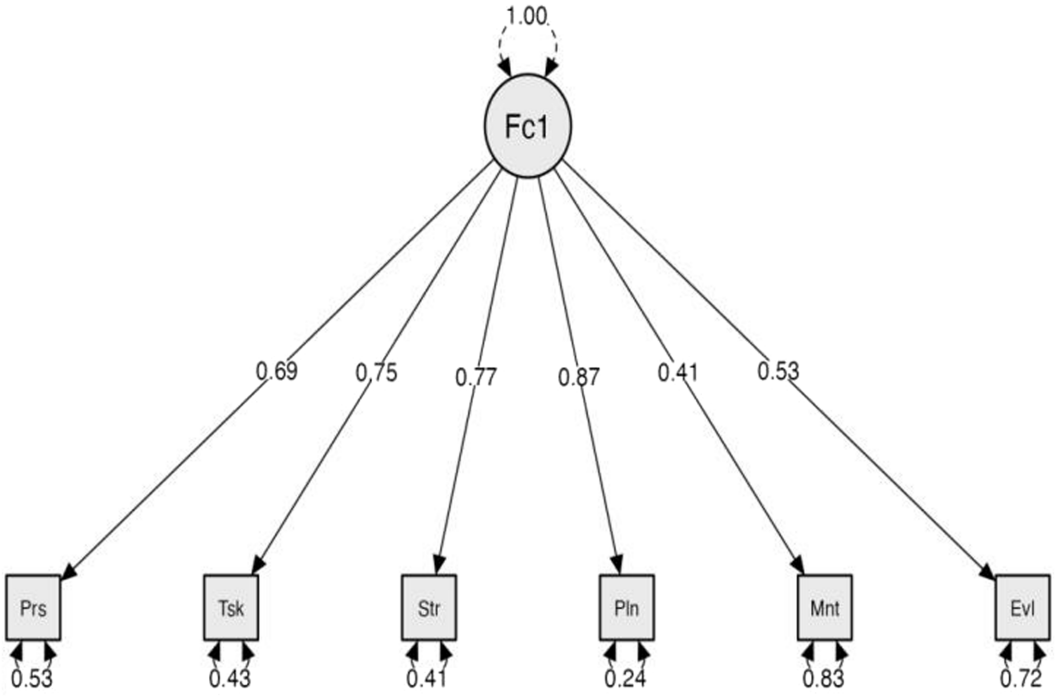



FIGURE 2
A first-order model of metacognitive strategies in writing. Prs, Person; Tsk, Task; Str, Strategy; Pln, Planning; Mnt, Monitoring; and Evl, Evaluating.

TABLE 5 Convergent validity of the model.

| Path | | | Estimate | AVE | CR |
|---|---|---|---|---|---|
| Person | <--- | F1 | 0.53 | 0.47 | 0.84 |
| Task | <--- | F1 | 0.41 | | |
| Strategy | <--- | F1 | 0.87 | | |
| Planning | <--- | F1 | 0.77 | | |
| Monitoring | <--- | F1 | 0.75 | | |
| Evaluating | <--- | F1 | 0.69 | | |

newly developed scale, which we named Metacognitive Strategies in Writing (MSW). The scale was developed based on metacognition theory. The findings supported the factorial structure of the scale. The second research question aims to answer the predictive effects of different factors of MSW in writing performance. Overall, the findings provided evidence for the factorial structure of MSW. The findings also suggested the predictive effects of different factors on writing performance.

## Validation of MSW

First, MSW is with satisfactory psychometric properties. The six factors were reliable in terms of conceptual and empirical evidence. The six factors were distinct but correlated with each other. Consistent with previous studies (Teng et al., 2022), metacognition is an important construct that can explain the significant correlations of different lower-order metacognitive dimensions in writing. In line with Schraw and Moshman (1995), metacognition is a domain that can explain self-regulatory capacity. The present study thus provides insights into metacognition theory, which can entail person, task, strategies, planning, monitoring, and evaluating (Schraw and Dennison, 1994). These strategies are interconnected and reflect the metacognitive process in writing. To build metacognitive awareness, learners need to be engaged in self-reflection and controlling of cognition (Paris and Winograd, 1990). In terms of writing, student writers need to assess their knowledge states and executive abilities to orchestrate different dimensions of metacognitive awareness. Overall, the sum of the six strategies in writing indicates EFL student writers' overall level of metacognitive awareness in writing.

The six factors were interpreted through metacognitive knowledge and regulation. The two paradigms were also conceptualized in early studies (Flavell, 1979; Schraw, 1998; Wenden, 1998). In the present study, the two paradigms can represent key elements of metacognition. Person, task, and strategies represent learners' beliefs and knowledge about themselves. Planning, monitoring, and evaluating reflect the process of cultivating one' self-regulatory capacity for learning to write (Teng and Zhang, 2016; Teng et al., 2022). The findings showed a positive and significant relationship between metacognitive knowledge and regulation (Pugalee, 2001; Teng, 2016). We may need to reconsider the strong connection between metacognitive knowledge and regulation. The positive correlation

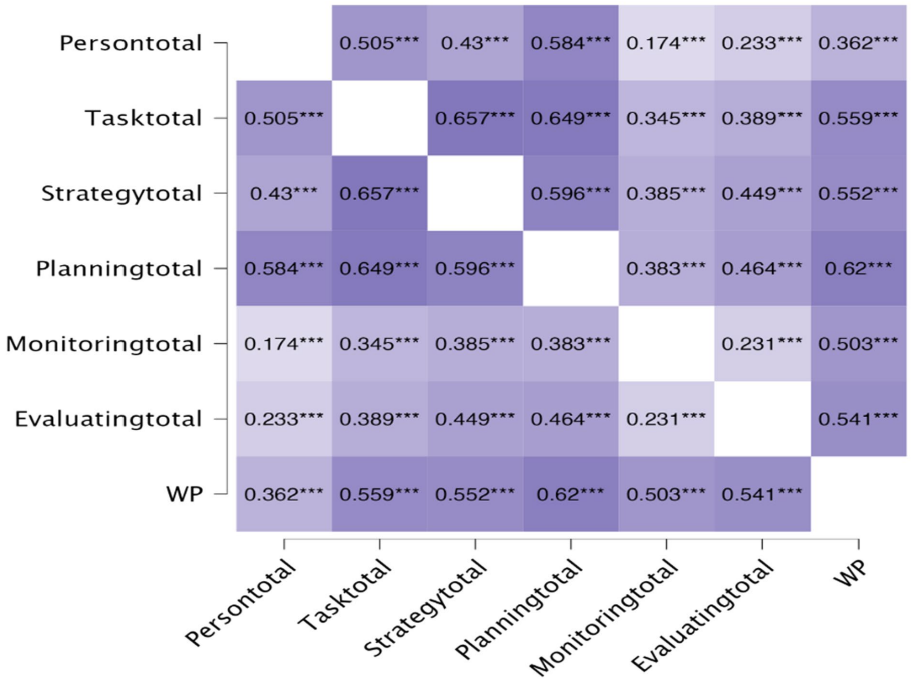



FIGURE 3
Spearman correlation for metacognitive writing strategies and L2 learners' proficiency in English. Persontotal, Person; Tasktotal, Task; Strategytotal, Strategy; Planningtotal, Planning; Monitoringtotal, Monitoring; and Evaluatingtotal, Evaluating.

TABLE 6 Model fit indices for metacognitive writing strategies on writing performance.

| $\chi^2$/df | RMSEA | CFI | TLI | NFI | IFI | GFI |
|---|---|---|---|---|---|---|
| 4.154 | 0.065 | 0.951 | 0.923 | 0.931 | 0.945 | 0.953 |

may reflect the need of both knowledge and regulation in learning to write. For example, EFL students may need cognitive, metacognitive, and regulatory skills and strategies for writing (Teng, 2020). The importance of metacognitive knowledge and regulation may reflect the argument by Wolters (1999) that learners' engagement, effort, and achievement are influenced by their metacognitive knowledge and regulation. Hence, metacognition is essential to the development of self-regulated capacity (Efklides, 2008), build identity as a student writer (Zimmerman and Risemberg, 1997, p.76), and develop self-awareness in processing their second and foreign language learning (Zhang and Zhang, 2019).

Overall, the MSW data suggest that the student writers adopted metacognitive knowledge, i.e., person, task, and strategies, to understand their strengths and weakness in writing, demands in writing, and solutions for solving problems in writing. The data also suggest that the planning strategy should be used. In the planning stage, the student writers directed their attention to fulfilling the goal of the task, planning thoroughly, evaluating the relevance and effectiveness of ideas, and eliminating inappropriate examples. Data regarding the second subscale (monitoring) reflected that students tended to use some metacognitive monitoring strategies. During the monitoring stage, the student writers focused on the overall essay development, concentrating on expanding and developing their initial ideas, evaluating their essay for clear development and focus/unity, and ignoring interruptions posed by language constraints, such as grammar and vocabulary. For the third subscale (self-evaluating), student writers tended to use certain metacognitive strategies. Student writers prioritized their attention to evaluating the unity and effectiveness of their writing before editing local errors, such as grammar, vocabulary, mechanics, and sentence variation.

## Predictive effects of metacognitive strategies in writing

The findings suggest the predictive effects of metacognitive strategies in writing. The results confirmed that the metacognitive strategies significantly predicted learners' writing performance, which was consistent with previous studies (Teng and Huang, 2019; Teng et al., 2022). One reason is that student writers' meager metacognitive knowledge base could result in unsatisfactory cognitive monitoring of production and progress toward the writing task goal, which, in turn, may also affect their writing performance (Teng et al., 2022). For example, lower-level writers tended to be bound to the local areas of writing, focusing on language correctness, while higher-level writers tended to focus on developing ideas and revising at the discourse level, saving editing until later (Teng and Huang, 2019). As supported in

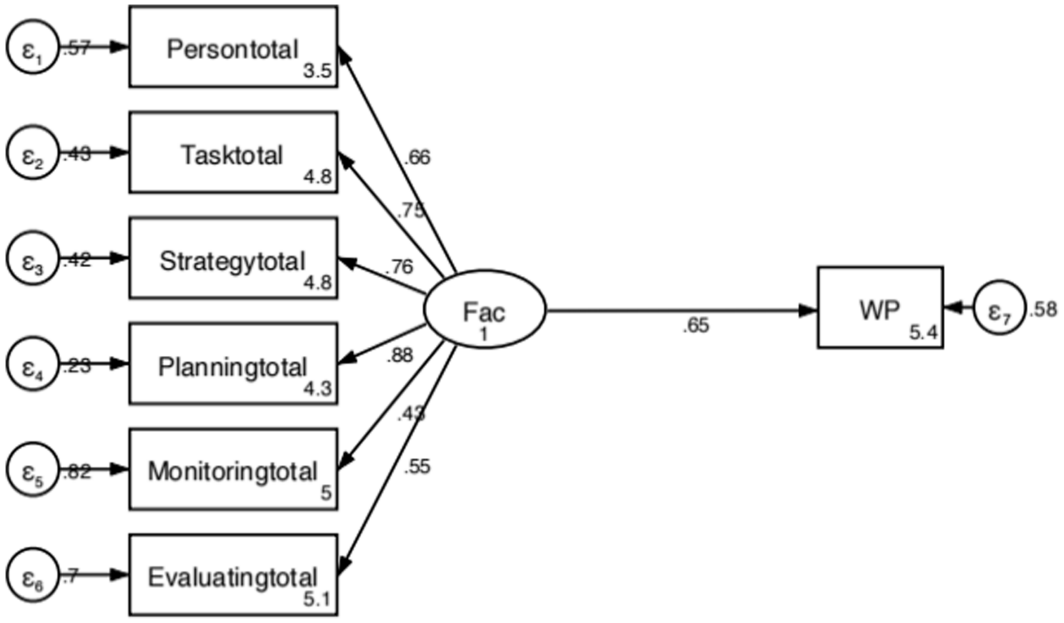



FIGURE 4
The structural equation model of metacognitive strategies in writing proficiency.

TABLE 7 Linear regression results.

| | Unstandardized coefficients | | Standardized coefficients | | | | | Collinearity statistics | |
|---|---|---|---|---|---|---|---|---|---|
| | *B* | SE | Beta | *t* | *p* | $R^2$ | Adjusted $R^2$ | Tolerance | VIF |
| (Intercept) | 3.203 | 0.636 | | 5.038 | < 0.001 | 0.473 | 0.467 | | |
| Person | −0.112 | 0.024 | −0.204 | −4.663 | < 0.001 | | | 0.555 | 1.803 |
| Task | 0.114 | 0.033 | 0.161 | 3.472 | < 0.001 | | | 0.493 | 2.028 |
| Strategy | 0.031 | 0.033 | 0.045 | 0.947 | 0.344 | | | 0.471 | 2.124 |
| Planning | 0.225 | 0.032 | 0.387 | 7.095 | < 0.001 | | | 0.356 | 2.807 |
| Monitoring | 0.136 | 0.024 | 0.203 | 5.588 | < 0.001 | | | 0.802 | 1.247 |
| Evaluating | 0.167 | 0.026 | 0.247 | 6.434 | < 0.001 | | | 0.721 | 1.387 |

previous studies (Chien, 2012; Bai et al., 2014), higher level student writers were more aware of metacognitive strategies and used them more frequently in writing.

The argument revealed, at least for this particular sample and the chosen test, a strong and significant link between the writing abilities of EFL students and the factors of person, task, strategy, planning, monitoring, and evaluation. The EFL learners' writing performance variations were accounted for by the six metacognitive components. The findings complement cognitive writing model of Flower and Hayes (1981), which recognizes the abilities in process writing such as planning, monitoring, and reviewing. Writing necessitates the adaptive use of emotional strategies, performance strategies, and cognitive strategies (Teng et al., 2022). The effectiveness of the strategies highlights the personal, behavioral, and environmental impacts on the regulatory capacity in learning to write (Zimmerman and Risemberg, 1997).

In our study, person and task significantly predicted writing performance with a large effect size. According to earlier research (Brown, 1987; Schraw, 2001), learners who have declarative, procedural, and conditional knowledge are more likely to become strategic learners. These results provide evidence for the idea that to master writing, EFL learners need to be able to distinguish among the various strategies, employ the appropriate strategies, and apply these strategies in their writing. The results also support earlier research that metacognitive knowledge is crucial for encouraging active involvement in applying their understanding of the writing process, recognizing the kinds of strategies useful in the growth of writing, and improving students' writing outputs (Ruan, 2014).

In terms of metacognitive regulation, planning, monitoring, and evaluating are also important for writing performance. The effect size was quite large in the current study, for which we can

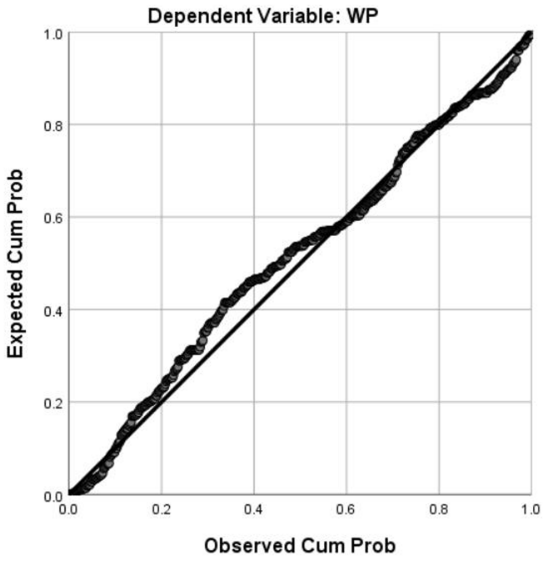



FIGURE 5
Normal P–P plot of regression standardized residual.

detect similar results in previous studies (Teng, 2019; Teng et al., 2022). The writing abilities of students who were more self-controlled in their writing were higher in terms of goal setting, time management, and planning for writing resources (Teng and Zhang, 2016). We argue that Chinese EFL students need an awareness of planning ahead and monitoring and evaluating their planning tactics to produce successful written essays. The success of EFL academic writing depends heavily on this method. Academic writing development may be seen as a complex process for student writers because it depends on how strategically they seek information and modify their planning techniques. Students who have prepared well for academic writing are typically those who have a high level of metacognitive awareness of their writing-related objectives (Zhang and Qin, 2018). When composing their essays, lower-level writers often experienced difficulty in transferring ideas to paper during the planning, monitoring, and self-evaluating stages. The constraints in the lower-level writers' knowledge system, including their limited linguistic competence (grammar and vocabulary), their confusion about their role as writers, their lack of knowledge strategies for overcoming writing difficulties, and their lack of knowledge of how and when to apply those strategies, impeded their composition of a meaningful essay. Consequently, many students tended to simultaneously engage in a few different stages of writing—planning, composing, revising, and editing—without any extra attention resources to monitor the overall unity and coherence of the essay, thus making the essay messy and confusing.

## Limitations and implications

Despite the positive findings, we still need to acknowledge some limitations of this study. First, the strategies described in the questionnaire were still scarce, although we showed excellent content validity. Due to the limited amount of time the learners could invest in data collection, we did not assess metacognitive experiences, another crucial component of metacognition. Interview data with students were not conducted to yield adequate methods connected to metacognitive experiences. Second, a self-report questionnaire served as the foundation for this study. Because they are dependent on the use of self-reported information, surveys may not fully reflect learners' actual metacognitive awareness and activities. The quantitative data in future studies should be triangulated with interview data. Third, the writing test should include additional activity categories that can gauge various writing abilities. We only used one writing performance indicator. The performance of student writers may also be impacted by individual characteristics, including their language learning experiences and English proficiency level (Teng and Huang, 2019). Future studies might look at learners' individual differences and their use of different metacognitive strategies.

However, there are also some implications based on the findings. Our findings suggest directions for pedagogy as well as future research. Considerations include issues of focus on form, development of metacognitive awareness to support metacognitive knowledge and strategies, and appreciation of the many aspects of metacognitive awareness that good L2 writing entails.

Data collected from the surveys suggest a strong connection between EFL student writers' metacognitive knowledge and the regulation strategies they employ. Helping students become more aware of themselves as writers and the metacognitive resources upon which they can draw during the writing process may help them develop their writing competence. Language teachers and instructors should clearly instruct the importance of metacognitive strategies for EFL student writers. Related to this, metacognitive training should help students develop such awareness in learning to write. However, an important step in developing productive pedagogy for metacognitive training is assessing learners' needs and understandings of their metacognitive strategies. The MSW might potentially contribute to EFL writing assessment in China. The MSW monitoring subscale identified the important first step in writing—planning—as a potential problem. So far as these Chinese EFL non-English major student writers were concerned, regardless of their level of English class or their majors, it seems that many of them may need to faster a metacognitive awareness. As a result, it might be helpful to provide these students with additional lessons on metacognitive strategies to address their concerns and the problems evident in their English writing. While dealing with grammatical errors is essential to writing instruction, the students should focus not only on identifying the errors and fixing them but also on finding out why they make those mistakes and how to avoid making them again. In other words, instead of correcting the errors, they should also develop their awareness of

metacognitive strategies to improve their overall language competence. The instructors may also explicitly teach and demonstrate effective strategies to enhance vocabulary acquisition, such as making learners aware of lexical morphology (including word roots and suffixes), synonyms, antonyms, word categories, and similar spellings.

Clearly, it should not be assumed that learners who do not score high on norm-referenced assessments of their L2 writing need to focus exclusively on their metacognitive strategies, even though that is where they may think they need to work. Rather, these learners need to consider not only metacognitive strategies but also discourse organization and considerations of audience, voice, and genre (Hyland, 2007). It is only through an approach raising their awareness of the various aspects that contribute to good writing and through work on writing and revision strategies that they will progress optimally. Additionally, to implement these recommendations for pedagogy, teachers themselves must have substantial knowledge, professional development, and practice regarding approaches to support L2 writing. In the Chinese context, knowledge must be processed and understood in light of the metacognition and experiences of students, colleagues, and the community.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Ethics statement

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

## Author contributions

CQ: Coordinated the study, drafted, and revised the manuscript. RZ: Data collection, drafted literature review. YX: Participated in the design of the study, revised the manuscript and performed the statistical analysis and data interpretation. All authors proofread and approved the final manuscript.

## Funding

This article is supported by the Project from the Education Department of Hainan Province, Project number: Hnky2020ZD-9.

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.1071907/full#supplementary-material

## References

Amani, S. (2014). Metacognitive strategy instruction and pre-task planning: Impact on L2 argumentative writing ability. Unpublished doctoral thesis. University of Auckland, Auckland, New Zealand. Available at: https://researchspace.auckland.ac.nz/handle/2292/21750

Bai, B., and Guo, W. (2021). Motivation and self-regulated strategy use: relationships to primary school students' English writing in Hong Kong. *Lang. Teach. Res.* 25, 378–399. doi: 10.1177/1362168819859921

Bai, R., Hu, G., and Gu, P. Y. (2014). The relationship between use of writing strategies and English proficiency in Singapore primary schools. *Asia Pac. Educ. Res.* 23, 355–365. doi: 10.1007/s40299-013-0110-0

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

Bruning, R., and Horn, C. (2000). Developing motivation to write. *Educ. Psychol.* 35, 25–37. doi: 10.1207/S15326985EP3501_4

Byrne, B. M. (2016). *Structural Equation Modelling With AMOS: Basic Concepts, Applications, and Programming* (*3rd Edn.*). New York: Routledge

Chamot, A. U., and El-Dinary, P. B. (1999). Children's learning strategies in immersion classrooms. *Mod. Lang. J.* 83, 319–338. doi: 10.1111/0026-7902.00025

Charles, M. (1990). Responding to problems in written English using a student self-monitoring technique. *ELT J.* 44, 286–293. doi: 10.1093/elt/44.4.286

Chien, S.-C. (2012). Students' use of writing strategies and their English writing achievements in Taiwan. *Asia Pacif. J. Educ.* 32, 93–112. doi: 10.1080/02188791.2012.655240

DiStefano, C., and Hess, B. (2005). Using confirmatory factor analysis for construct. Validation: an empirical review. *J. Psychoeduc. Assess.* 23, 225–241. doi: 10.1177/073428290502300303

Dornyei, Z. (2010). "Language aptitude and language learning motivation" in *Continuum Companion to Second Language Acquisition*. ed. E. Macaro (London: Continuum), 268–299.

Efklides, A. (2008). Metacognition: defining its facets and levels of functioning in relation to self-regulation and co-regulation. *Eur. Psychol.* 13, 277–287. doi: 10.1027/1016-9040.13.4.277

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

Flavell, J. H. (1985). *Cognitive Development* (*2nd Edn.*). Englewood Cliffs, N.J: Prentice-Hall

Flower, L., and Hayes, J. (1981). A cognitive process theory of writing. *Coll. Compos. Commun.* 32, 365–387. doi: 10.2307/356600

Graham, S., and Harris, R. K. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educ. Psychol.* 35, 3–12. doi: 10.1207/S15326985EP3501_2

Graham, S., MacArthur, C., and Schwartz, S. (1995). The effects of goal setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *J. Educ. Psychol.* 87, 230–240. doi: 10.1037/0022-0663.87.2.230

Hayes, J., and Flower, L. (1980). The cognition of discovery: defining a rhetorical problem. *Coll. Compos. Commun.* 31, 21–32.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model. Multidiscip. J.* 6, 1–55. doi: 10.1080/10705519909540118

Hyland, K. (2007). Genre pedagogy: language, literacy, and L2 writing instruction. *J. Second. Lang. Writ.* 16, 148–164. doi: 10.1016/j.jslw.2007.07.005

Lai, P. F. (2009). Attitudes of adult Chinese-background learners and an ESL teacher towards L1 use in an Australian AMEP class in Australia. *Univ. Sydney Papers TESOL* 4, 57–84.

Leung, C. B., and Hicks, J. (2014). Writer identity and writing workshop: a future teacher and teacher educator critically reflect. *Writ. Pedagog.* 6, 583–605. doi: 10.1558/wap.v6i3.583

Ma, J., and Teng, F. (2021). "Metacognitive knowledge development of students with differing levels of writing proficiency in a process-oriented course: an action research study" in *Innovative Approaches in Teaching Writing to Chinese Speakers*. eds. B. L. Reynolds and F. Teng (Berlin: DeGruyter Mouton), 92–117.

Macaro, E. (2010). "The relationship between strategic behaviour and language learning success" in *Continuum Companion to Second Language Acquisition*. ed. E. Macaro (London: Continuum), 268–299.

Magogwe, J. M., and Oliver, R. (2007). The relationship between language learning strategies, proficiency, age and self-efficacy beliefs: a study of language learners in Botswana. *System* 35, 338–352. doi: 10.1016/j.system.2007.01.003

Mayer, C. (1999). Shaping at the point of utterance: an investigation of the composing processes of the deaf student writer. *J. Deaf. Stud. Deaf. Educ.* 4, 37–49. doi: 10.1093/deafed/4.1.37

Nguyen, L. T. C., and Gu, Y. (2013). Strategy-based instruction: a learner-focused approach to developing learner autonomy. *Lang. Teach. Res.* 17, 9–30. doi: 10.1177/1362168812457528

Nishino, T., and Atkinson, D. (2015). Second-language writing as sociocognitive alignment. *J. Second. Lang. Writ.* 27, 37–54. doi: 10.1016/j.jslw.2014.11.002

Oxford, R. L. (1990). *Language Learning Strategies: What Every Teacher Should Know*. New York: Newbury House/Harper & Row

Oxford, R. L. (2013). *Teaching and Researching Language Learning Strategies* (*2nd Edn.*). Harlow, UK: Pearson

Oxford, R. L., and Cohen, A. D. (1992). Language learning strategies: crucial issues in concept and classification. *Appl. Lang. Learn.* 3, 1–35.

Oxford, R. L., and Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *Mod. Lang. J.* 73, 291–300. doi: 10.1111/j.1540-4781.1989.tb06367.x

Paris, S., and Winograd, P. (1990). "How metacognition can promote academic learning and instruction" in *Dimensions of Thinking and Cognitive Instruction*. eds. B. F. Jones and L. Idol (Hillsdale, NJ: Erlbaum Associates), 15–51.

Peacock, M., and Ho, B. (2003). Student language learning strategies across eight disciplines. *Int. J. Appl. Linguist.* 13, 179–200. doi: 10.1111/1473-4192.00043

Pugalee, D. K. (2001). Writing, mathematics, and metacognition: looking for connections through students' work in mathematical problem solving. *Sch. Sci. Math.* 101, 236–245. doi: 10.1111/j.1949-8594.2001.tb18026.x

Raimes, A. (1987). Language proficiency, writing ability, and composing strategies: a study of ESL college student writers. *Lang. Learn.* 37, 439–468. doi: 10.1111/j.1467-1770.1987.tb00579.x

Riazi, A. (1997). Acquiring disciplinary literacy: a social-cognitive analysis of text production and learning among Iranian graduate students of education. *J. Second. Lang. Writ.* 6, 105–137. doi: 10.1016/S1060-3743(97)90030-8

Ruan, Z. (2014). Metacognitive awareness of EFL student writers in a Chinese ELT context. *Lang. Aware.* 23, 76–91. doi: 10.1080/09658416.2013.863901

Schraw, G. A. (1998). Promoting general metacognitive awareness. *Instr. Sci.* 26, 113–125. doi: 10.1023/A:1003044231033

Schraw, G. A. (2001). "Promoting general metacognitive awareness" in *Metacognition in Learning and Instruction: Theory, Research and Practice*. ed. H. J. Hartman (London: Springer), 3–16.

Schraw, G., and Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemp. Educ. Psychol.* 19, 460–475. doi: 10.1006/ceps.1994.1033

Schraw, G., and Moshman, D. (1995). Metacognitive theories. *Educ. Psychol. Rev.* 7, 351–371. doi: 10.1007/BF02212307

Tabachnick, B. G., and Fidell, L. S. (2001). *Using Multivariate Statistics*. Needham Heights, MA: Allyn & Bacon

Teng, F. (2016). Immediate and delayed effects of embedded metacognitive instruction on Chinese EFL students' English writing and regulation of cognition. *Think. Skills Creat.* 22, 289–302. doi: 10.1016/j.tsc.2016.06.005

Teng, F. (2019). The role of metacognitive knowledge and regulation in mediating university EFL learners' writing performance. *Innov. Lang. Learn. Teach.* 14, 436–450. doi: 10.1080/17501229.2019.1615493

Teng, F. (2020). Tertiary-level students' English writing performance and metacognitive awareness: a group metacognitive support perspective. *Scand. J. Educ. Res.* 64, 551–568. doi: 10.1080/00313831.2019.1595712

Teng, F., and Huang, J. (2019). Predictive effects of writing strategies for self-regulated learning on secondary school learners' EFL writing proficiency. *TESOL Q.* 53, 232–247. doi: 10.1002/tesq.462

Teng, F., Qin, C., and Wang, C. (2022). 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. (2022). Assessing self-regulatory writing strategies and their predictive effects on young EFL learners' writing performance. *Assess. Writ.* 51:100573. doi: 10.1016/j.asw.2021.100573

Teng, F., and Yue, M. (2022). Metacognitive writing strategies, critical thinking skills, and academic writing performance: a structural equation modeling approach. *Metacogn. Learn.* doi: 10.1007/s11409-022-09328-5

Teng, L. S., and Zhang, L. J. (2016). A questionnaire-based validation of. Multidimensional models of self-regulated learning strategies. *Mod. Lang. J.* 100, 674–701. doi: 10.1111/modl.12339

Teng, F., and Zhang, L. J. (2021). Development of children's metacognitive knowledge, and reading and writing proficiency in English as a foreign language: longitudinal data using multilevel models. *Br. J. Educ. Psychol.* 91, 1202–1230. doi: 10.1111/bjep.12413

Teng, L. S., and Zhang, L. J. (2022). Can self-regulation be transferred to second/foreign language learning and teaching? Current status, controversies, and future directions. *Appl. Linguis.* 43, 587–595. doi: 10.1093/applin/amab032

Wenden, A. L. (1991). *Learner Strategies for Learner Autonomy: Planning and Implementing Learner Training for Language Learners*. Cambridge: Prentice Hall International

Wenden, A. L. (1998). Metacognitive knowledge and language learning. *Appl. Linguist.* 19, 515–537.

Wolters, C. A. (1999). The relation between high school students' motivational regulation and their use of learning strategies, effort, and classroom performance. *Learn. Individ. Differ.* 11, 281–299. doi: 10.1016/S1041-6080(99)80004-1

Wu, Y. L. (2008). Language learning strategies used by students at different proficiency levels. *Asian EFL J.* 10, 75–95.

Zhang, L. J., and Qin, L. (2018). "Validating a questionnaire on EFL writers' metacognitive awareness of writing strategies in multimedia environments" in *Metacognition in Language Learning and Teaching*. eds. Å. Haukås, C. Bjørke and M. Dypedahl (New York: Routledge), 157–178.

Zhang, D., and Zhang, L. J. (2019). "Metacognition and self-regulated learning (SRL) in second/foreign language teaching" in *Second Handbook of English Language Teaching* (Singapore: Springer), 883–897.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: an overview. *Theory Pract.* 41, 64–70. doi: 10.1207/s15430421tip4102_2

Zimmerman, B. J., and Risemberg, R. (1997). Becoming a self-regulated writer: a social cognitive perspective. *Contemp. Educ. Psychol.* 22, 73–101. doi: 10.1006/ceps.1997.0919

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 20 February 2023
DOI 10.3389/fpsyg.2023.1098916



OPEN ACCESS

EDITED BY
Hui Zhang,
Nanjing Normal University,
China

REVIEWED BY
Hung Tan Ha,
University of Economics Ho Chi Minh City,
Vietnam
Sida Zhu,
The Education University of Hong Kong,
Hong Kong SAR,
China
Fei Qiu,
Changzhou University,
China

*CORRESPONDENCE
Yingli Wang
✉ wangyingli@dlmu.edu.cn

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

RECEIVED 15 November 2022
ACCEPTED 31 January 2023
PUBLISHED 20 February 2023

CITATION
Wang H, Wang Y and Li S (2023) Unpacking the relationships between emotions and achievement of EFL learners in China: Engagement as a mediator.
*Front. Psychol.* 14:1098916.
doi: 10.3389/fpsyg.2023.1098916



# Unpacking the relationships between emotions and achievement of EFL learners in China: Engagement as a mediator

Haihua Wang[1], Yingli Wang[1]* and Shaojie Li[2]

[1]School of Foreign Languages, Dalian Maritime University, Dalian, China, [2]Department of English Language and Literature, Korea Maritime and Ocean University, Busan, Republic of Korea

Emotions are attracting growing attention in second language acquisition (SLA), especially with the advent of positive psychology (PP). The fundamental role of emotions in affecting learners' second language (L2) achievement has been well-documented. Evidence also indicates that emotions can significantly influence learners' L2 learning engagement which profoundly impacts their academic performance. However, the links between emotions, engagement, and L2 achievement remain underexplored. To contribute to this research domain, the present study sought to unpack the relationships between learners' emotions, such as foreign language enjoyment (FLE), foreign language classroom anxiety (FLCA), and foreign language learning boredom (FLLB), and engagement as well as their English achievement. A total of 907 learners of English as a foreign language (EFL) from a university in China were recruited to complete an online questionnaire. Structural equation modeling (SEM) was performed to test the hypothesized relations among the variables. Results revealed correlations between learners' FLE, FLCA, and FLLB. Furthermore, learners' engagement was found to mediate the relationships between their emotions (FLE, FLCA, and FLLB) and English achievement. The findings broaden the nomological network of emotions and engagement in the EFL context, and provide evidence for the mechanism underlying the relationships between emotions, engagement, and achievement, thereby shedding light on EFL teaching and learning at the tertiary level in China.

KEYWORDS

**emotions, foreign language enjoyment, foreign language classroom anxiety, foreign language learning boredom, engagement, English achievement, learners of English as a foreign language, structural equation modeling**

## 1. Introduction

Learners experience various emotions when they attend class, participate in activities, interact with teachers and peers, or take exams, showing that emotions are ubiquitous in academic and language learning (Pekrun and Linnenbrink-Garcia, 2012; Plonsky et al., 2022). In the past decades, a sizable amount of research on emotions in the field of Second Language Acquisition (SLA) has been carried out, with the main focus on a negative emotion—foreign language classroom anxiety (FLCA). Scholars have correlated this construct with numerous learning variables, including motivation (Hashimoto, 2002; Fathi and Mohammaddokht, 2021), willingness to communicate (WTC; MacIntyre et al., 2002), learner personality traits (Dewaele,

2002), emotional intelligence (Dewaele et al., 2008), learner-related variables (Dewaele, 2013), various measures of language achievement (Shao et al., 2013; Li et al., 2019; Zhang, 2019; Botes et al., 2020), and other emotions (Dewaele and Proietti Ergün, 2020; Dewaele et al., 2022; Li and Han, 2022). However, on the way to learning a foreign language, foreign language enjoyment (FLE) and FLCA were the metaphorical "right and left feet of language learner" (Dewaele and MacIntyre, 2016, p: 215). Therefore, researchers have juxtaposed FLCA with FLE in research.

Recently, with the emergence and flowering of Positive Psychology (PP) in educational psychology, researchers in SLA started to shift their focus from investigating negative emotions to exploring positive affective variables. The role of positive emotions was highlighted, among which FLE was one of the most studied constructs. Scholarly studies on FLE addressed its conceptualization and measurement (Dewaele and MacIntyre, 2014, 2016; Jin and Zhang, 2019), its sources and effects on foreign language learning (Jin and Zhang, 2021; Botes et al., 2022), as well as its linkages with other learner-related variables (Dewaele et al., 2017, 2019a; Li et al., 2021a). Furthermore, the correlations and combined impacts of FLE and FLCA on foreign language achievement have been examined in studies (Dewaele and Alfawzan, 2018; Li et al., 2019; Dewaele and Proietti Ergün, 2020). Apart from FLE and FLCA, another emotion learners most frequently experience in foreign language classrooms is boredom (Pekrun et al., 2010). Despite its popularity in educational psychology for decades, boredom has not received much attention in SLA until recently. Research on foreign language learning boredom (FLLB) concentrated on its conceptualization and measurement (Kruk and Zawodniak, 2017; Li et al., 2021b), as well as antecedents and effects (Kruk and Zawodniak, 2018; Li, 2021; Derakhshan et al., 2021a,b). In language learning, FLE, FLCA, and FLLB could be simultaneously experienced by learners. A holistic view of the complex emotions of foreign language learners can throw light on teachers' pedagogical practices, and facilitate learners' language learning to achieve better learning outcomes. Nonetheless, few studies have been undertaken to delve into how FLE, FLCA, and FLLB are associated with achievement (Dewaele et al., 2022; Li and Han, 2022; Li and Wei, 2022), leaving their relationships largely underexplored.

Despite a growing number of studies on emotions in foreign language learning, the nomological network of emotions still needs to be further expanded (Botes et al., 2022). In the field of education and educational psychology, one of the variables that have been evidenced to be the result of emotions is engagement (Oga-Baldwin, 2019), a construct conducive to scholarly success (Pekrun and Linnenbrink-Garcia, 2012; Reschly and Christenson, 2012). Previous studies have demonstrated that engagement was a mediator between learners' emotions and their academic learning and achievement (Pekrun, 2006; Linnenbrink, 2007). In language education, although studies showed that emotions were associated with engagement (Dewaele and Li, 2021; Mohammad Hosseini et al., 2022), which was recognized as one of the strongest predictors of language achievement (Masgoret and Gardner, 2003), only a handful of studies have expounded on the relationships between emotions, engagement, and achievement in a single study (Dewaele and Li, 2021; Khajavy, 2021; Feng and Hong, 2022).

To fill the research gaps, the present study aimed to unpack the relationships between emotions, engagement, and achievement in language learning by collecting data from Chinese learners of English as a foreign language (EFL) to test a model that hypothesized different emotional variables (FLE, FLCA, FLLB) as predictors of English achievement, with engagement as the mediator. The findings may broaden the nomological network of emotions and engagement in SLA and offers pedagogical implications for EFL teachers and practitioners at the tertiary level.

## 2. Literature review

### 2.1. Emotions

#### 2.1.1. Foreign language classroom anxiety

Since the 1970s, research on affective variables, especially negative emotions, has garnered SLA researchers' attention (Dewaele et al., 2017). FLCA has been the most studied negative emotion in SLA (MacIntyre, 2017). Horwitz et al. (1986) defined FLCA as "a distinct complex of self-perceptions, beliefs, feelings, and behaviors related to classroom learning arising from the uniqueness of the language learning process" (p: 128), highlighting the multifaceted concept of anxiety. Since the introduction of FLCA in 1986, several scales have been developed to measure it, among which Foreign Language Classroom Anxiety Scale (FLCAS) constructed by Horwitz et al. (1986) has been widely accepted and used by researchers. It is a five-point Likert scale questionnaire comprising 33 items, which has been adapted, shortened, and translated in subsequent studies (Tóth, 2008; Dewaele and MacIntyre, 2014; Dewaele and Al-Saraj, 2015; Li and Wei, 2022).

Adopting various measures of FLCA, researchers have investigated its potential sources, effects, and correlations (Dewaele and MacIntyre, 2014). The existing literature indicated that anxiety of language learners was negatively associated with their motivation (Hashimoto, 2002; Fathi and Mohammaddokht, 2021), their WTC (MacIntyre et al., 2002), and their engagement (Feng and Hong, 2022) in foreign language learning. Also, recent studies have yielded an inverse relationship between FLCA and various language achievement measures (Shao et al., 2013; Li et al., 2019; Zhang, 2019; Botes et al., 2020; Dewaele and Proietti Ergün, 2020; Dewaele et al., 2022; Li and Han, 2022). Shao et al. (2013) explored 510 Chinese EFL students' emotional intelligence and English classroom learning anxiety, and found negative associations between students' FLCA and their self-rated English proficiency as well as English achievement measured by CET-4 scores. Dewaele and Proietti Ergün (2020) investigated the relationship between Turkish pupils' FLCA and their course marks in two foreign languages, Italian and English. The finding exhibited that pupils with high FLCA had lower course marks in both foreign languages. Nevertheless, according to Pekrun (2006) control-value theory, anxiety, as an activating, negative, and achievement-related emotion, can have an ambivalent effect on academic achievement. Therefore, the complex relationship between FLCA and achievement still needs more empirical evidence.

#### 2.1.2. Foreign language enjoyment

With the emergence of PP, researchers in general education gradually shifted their obsession with exploring negative emotions to more positive ones. PP was first introduced in a seminal paper by Seligman and Csikszentmihalyi (2000), who argued the need to focus on the positive aspects of human experience and the reasons for their

initiation (MacIntyre, 2021). PP was defined as "the scientific study of what goes right in life, from birth to death and at all stops in between" (Peterson, 2006, p: 4). The two underpinnings for PP are the broaden-and-build theory of positive emotions (Fredrickson, 2001) and the control-value theory of achievement emotions (Pekrun et al., 2002; Pekrun, 2006). The broaden-and-build theory states that positive emotions, including joy, interest, pride, and love, can "broaden people's momentary thought-action repertoires and build their enduring personal resources" (Fredrickson, 2001, p: 219). On the other hand, based on the control-value theory, achievement emotions can be grouped according to their object focus (activity focus vs. outcome focus), valence (positive vs. negative), and the degree of activation (activating vs. deactivating). However, it was not until recently that PP research "penetrated the mainstream" (Dewaele et al., 2019b) in the field of SLA.

With the popularity of PP in SLA, researchers switched their interest from negative emotions to the positive factors involved in language learning, among which FLE had been one of the most investigated emotions in SLA. Dewaele and MacIntyre (2016) defined FLE as "a complex emotion, capturing interacting dimensions of the challenge and perceived ability that reflect the human drive for success in the face of difficult tasks" (p: 216), which occurs "when people not only meet their needs, but exceed them to accomplish something new or even unexpected" (p: 217). Dewaele and MacIntyre (2014) conducted a pioneering study on the relationship between FLE and FLCA. Based on the FLE scale developed, which comprised 21 items covering various facets of FLE in the foreign language class, a significant and negative correlation was found between FLE and FLCA but with a small amount of shared variance, displaying that they were different emotion dimensions. In addition, levels of FLE were reported to be significantly higher than those of FLCA. This study paved the way for applying PP in SLA (Wang et al., 2021), after which research on FLE in western and eastern contexts flourished.

Studies on FLE have explicated its measurement (Dewaele et al., 2017; Li et al., 2018; Jin and Zhang, 2019; Botes et al., 2021; Jin and Zhang, 2021), and how it is associated with learner-related variables (Dewaele et al., 2019a; Li et al., 2021a). For instance, based on the FLE scale constructed by Dewaele and MacIntyre (2014, 2016), Li et al. (2018) developed the Chinese Version of the FLE Scale (CFLES). By collecting data from 2,078 Chinese students, they conducted a Principal Component Analysis to confirm and validate a new FLE scale containing 11 items with three factors (FLE-Private, FLE-Teacher, and FLE-Atmosphere). The participants reported that their FLE arose through direct teachers' intervention and indirect peer interaction.

Relevant studies have also explored the relationships between FLE and different measures of foreign language performance. Significantly positive relationships between FLE and both perceived and actual language achievement were found in relevant studies (Piechurska-Kuciel, 2017; Li, 2020; Jin and Zhang, 2021; Botes et al., 2022). Jin and Zhang (2021) collected data from 320 Chinese EFL senior high school students and investigated the dimensions of foreign language classroom enjoyment and their effect on foreign language achievement, and found that the participants' enjoyment of foreign language learning directly affected mid-term scores, while both enjoyment of teacher support and students support had indirect influences, revealing that FLE impacted foreign language learning in a complex way. A meta-analysis conducted by Botes et al. (2022) suggested a moderate positive correlation between FLE and academic achievement, as well as self-perceived achievement, confirming the significance of FLE in foreign language learning.

Additionally, FLE was also juxtaposed with FLCA since "the combination of positive and negative emotions together is more powerful for influencing teaching practice than looking at them individually" (MacIntyre, 2021, p: 11). Li et al. (2019) examined the correlation between FLE and FLCA, as well as their combined effects on self-perceived English proficiency and actual English achievement of Chinese EFL students. The results revealed a negative correlation between FLE and FLCA, which echoed the previous studies (Dewaele and MacIntyre, 2014, 2016; Dewaele et al., 2016) and was then confirmed in the later study (Dewaele and Proietti Ergün, 2020). Moreover, FLE and FLCA could co-predict self-perceived English proficiency and actual English achievement, with FLCA being the stronger predictor. This result contradicted the findings of Dewaele and Alfawzan (2018) as well as Li and Wei (2022), which showed the positive effect of FLE on performance outweighed that of FLCA. As such, the complex relationships between these two emotions and foreign language performance still need further exploration.

### 2.1.3. Foreign language learning boredom

Boredom is among the most frequently experienced and potentially devastating academic emotions in the classroom (Pekrun et al., 2010, 2014). It can be defined as "a mild, unpleasant or even painful affective state" that involves "a combination of dissatisfaction, disappointment, annoyance, inattention, lack of motivation to pursue previously set goals and impaired vitality" (Kruk and Zawodniak, 2018, P: 177). Boredom has attracted the interest of researchers in psychology and educational psychology for decades, whereas it was not until relatively recently that boredom received increasing attention in second language (L2) learning and teaching. Chapman (2013) was the first researcher exploring German learners' and their teachers' beliefs about boredom. Later, several studies were carried out in the Polish educational context to investigate this negative emotion in terms of changes in the level of boredom (Kruk, 2016a,b), the relationship between boredom experienced by learners and the boredom exhibited in EFL classes (Kruk and Zawodniak, 2017), as well as the experience of boredom in EFL classes (Kruk and Zawodniak, 2018). Other research into boredom concentrated on its impacts on WTC (Zhang et al., 2022), engagement (Dewaele and Li, 2021; Derakhshan et al., 2022), its causes, effects, and solutions in online classes (Derakhshan et al., 2021a,b; Pawlak et al., 2022), together with its conceptualization and measurement (Kruk and Zawodniak, 2017; Li, 2021; Li et al., 2021b). Kruk and Zawodniak (2017) developed the Boredom in Practical English Language Classes Questionnaire (BPELC) to measure this negative emotion in foreign language learning. However, Li et al. (2021b) in their study pointed out the weaknesses of BPELC and they developed the Foreign Language Learning Boredom Scale (FLLBS), a seven-factor scale containing 32 items and exhibiting good psychometric properties, which was adopted in the present study to evaluate the participants' boredom.

Additionally, some studies juxtaposed FLLB with other emotions, such as FLE and FLCA, to expound their interrelations and influences on foreign language performance. Li and Han (2022) investigated the effects of FLE, FLCA, and FLLB on Chinese EFL learners' self-perceived and actual achievement in an online learning environment. FLLB was found to have a positive relation with FLCA but a negative association with FLE, and independent negative predictive effects on perceived learning achievement and actual achievement. When

entering into the same regression model with FLE and FLCA, FLLB maintained its predictive power on perceived online learning achievement, but failed to directly influence actual test scores. This finding was confirmed in Dewaele et al. (2022) study, showing that FLLB was significantly interrelated with FLE and FLCA, and had no predictive effect on actual achievement when combined with the other two co-predictors. Nevertheless, in the domain of education, the role of boredom as a negative predictor of achievement has been evidenced in quite a few studies (Maroldo, 1986; Pekrun et al., 2009, 2010; Ahmed et al., 2013). The control-value theory also suggested that the effects of boredom can be "detrimental" to academic achievement (Pekrun, 2006). Consequently, FLLB, as an under-investigated emotion in SLA, is in dire need of empirical exploration into its complex relationships with other emotions simultaneously experienced by learners, and its complicated impacts on different measures of foreign language achievement.

## 2.2. Engagement

Engagement, a key contributor to learning and academic success, is about the energy learners spend toward the achievement (Fredricks et al., 2016). As a multifaceted concept describing what and how students think, act, and feel in a classroom setting (Fredricks et al., 2004; Oga-Baldwin, 2019), engagement is conceptualized as being comprised of three dimensions—behavioral, emotional, and cognitive engagement (Fredricks et al., 2004). Behavioral engagement refers to learners' qualitative behavioral choices in learning (Hiver et al., 2021a), such as their participation, effort, attention, and persistence (Fredricks et al., 2016). Emotional engagement includes the positive and negative affective reactions in the classroom toward teachers, classmates, schools, or school activities (Finn, 1989; Fredricks et al., 2004). Cognitive engagement is conceived as using deep learning strategies and putting effort into comprehending complex ideas (Fredricks et al., 2004). More recently, scholars have proposed additional dimensions to the conceptualization of engagement, such as social engagement (Svalberg, 2009; Wang et al., 2016), agentic engagement (Reeve and Tseng, 2011), and volitional engagement (Filsecker and Kerres, 2014). In language learning research, social engagement has been regarded as a critical dimension of engagement (Hiver et al., 2021a) as it is "essentially linked to interaction and to learners' initiation and maintenance of it" (Svalberg, 2009, p: 252). To have a holistic understanding of engagement, the dimensions of engagement should be considered together in research instead of focusing on one or two dimensions separately (Zhou et al., 2021). Accordingly, this study delved into four dimensions of engagement, namely behavioral engagement, emotional engagement, cognitive engagement, and social engagement comprehensively.

Engagement, as a significant factor in PP (Wang et al., 2021), has been receiving great attention in educational psychology, on which much research has been conducted regarding its link to better academic achievement (Wang and Holcombe, 2010), self-efficacy (Schunk and Mullen, 2012), achievement goals (Anderman and Patrick, 2012), and emotions (Pekrun and Linnenbrink-Garcia, 2012). Despite its enormous popularity in the educational field, there remains a paucity of research on engagement in SLA. L2 engagement is defined as the extent to which a language learner is involved in doing a language learning task (Hiver et al., 2021b). Studies have connected engagement to foreign language classroom environment (Sulis and Philp, 2021; Mohammad Hosseini et al., 2022), learning and communication mode (Carver et al., 2021), learner-related variables (Mercer and Dörnyei, 2020; Dewale and Li, 2021; Guo, 2021; Derakhshan et al., 2022; Zhao and Yang, 2022), and foreign language achievement (Masgoret and Gardner, 2003; Eren and Rakıcıoğlu-Söylemez, 2020; Kang and Wu, 2022). Nevertheless, the sources and effects of this important construct in foreign language learning still have not been explicated in depth. Previous research in education has proposed the contextual model in which engagement was influenced by the learning environment, such as interpersonal relationships in the classroom, and personal factors like emotions and beliefs. In turn, it affects learners' future attitudes and achievements (Oga-Baldwin, 2019). Few studies in SLA, specifically in the EFL context, have evidenced the role of L2 engagement as a mediator, especially between emotions and L2 achievement. For example, Khajavy (2021) hypothesized a model in which L2 engagement mediated the relationship between L2 emotions, L2 grit, and L2 reading comprehension in the Iranian EFL context. The finding confirmed the role of L2 engagement as a mediator between perseverance and L2 reading achievement, interest and L2 reading achievement, as well as L2 enjoyment and L2 reading achievement. Feng and Hong (2022) explored the relationship between achievement emotions (FLE and FLCA), behavioral engagement, and self-reported achievement of Chinese EFL learners. The results demonstrated that behavioral engagement mediated the relationship between FLE and self-reported achievement, as well as between FLCA and self-reported achievement. The mediating role of behavioral engagement was also confirmed in Kang and Wu (2022) study, which investigated whether behavioral engagement mediated the academic enjoyment and English achievement of Chinese EFL learners. Therefore, the mediating role of engagement between various emotions and achievement needs further exploration.

Taken together, to our best knowledge, scant research has ever delved into different emotional variables, such as FLE, FLCA, FLLB, L2 engagement, and achievement in a single study, not to mention the mechanism underlying their relationships. To this end, the present study aims to unpack the relationships between emotions and the achievement of EFL learners in China, with engagement as a mediator.

Based on the theoretical and empirical backgrounds of the constructs reviewed above, a structural model of FLE, FLCA, FLLB, L2 engagement, and English achievement was hypothesized. The model and its hypothesized paths are displayed in Figure 1. Given the hypothesized model, the following hypotheses were proposed:

> *Hypothesis 1*: FLE, FLCA, and FLLB are correlated significantly.
>
> *Hypothesis 2*: L2 emotions affect L2 engagement significantly.
>
> *Hypothesis 3*: L2 emotions influence English achievement significantly.

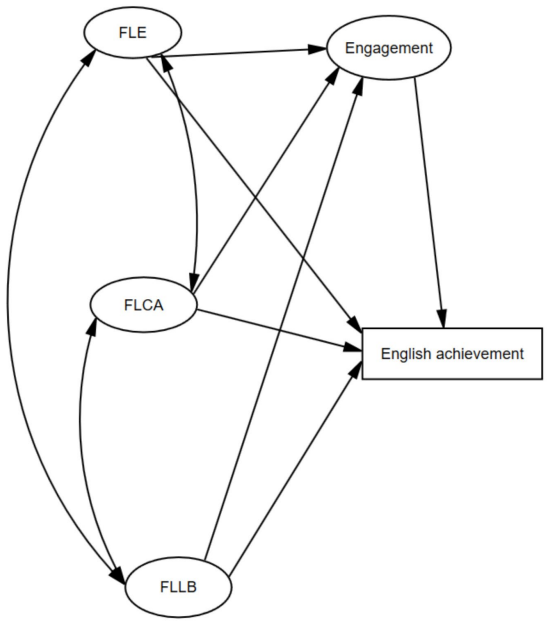



FIGURE 1
The hypothesized model.

*Hypothesis 4*: L2 engagement exerts significant effects on English achievement.

*Hypothesis 5*: L2 engagement mediates L2 emotions and English achievement.

# 3. Methodology

## 3.1. Participants

Convenience sampling was adopted in this study. A total of 921 second-year non-English majors at a university located in northeast China originally participated in the online questionnaire survey. After eliminating 14 questionnaires due to incompleteness, the final sample size was 907. They were from more than 10 majors such as International Economics and Trade, Business Administration, and Logistics Engineering. Among them, 394 were male and 513 were female, with ages ranging from 18 to 23 ($M = 20.55$, $SD = 0.902$). The participants were enrolled in the university after taking the college entrance examination. They were all native Chinese EFL learners with no experience of studying abroad and had learned English for at least 7 years. In China, English teaching at the tertiary level follows the Guidelines for College English Teaching (College Foreign Language Teaching Advisory Committee, 2020). Learners' skills in English listening, speaking, reading, writing, and translation are mainly cultivated in an English course known as "College English," a compulsory subject for the participants at this university. They were supposed to take the English course for an average of 3 hours per week for 2 academic years. At the end of the first academic year, they were required to take the College English Test Band 4 (CET-4), a national and standardized test held by the Ministry of Education of the People's Republic of China to evaluate the undergraduates' English proficiency. At the time of data collection, all of them had taken the CET-4.

## 3.2. Instruments

The research instrument employed in this study was a composite questionnaire, which was comprised of two major parts: the first part consisted of questions about personal background information (e.g., name, gender, age, and CET-4 scores); the second part was composed of items measuring the participants' FLE, FLCA, FLLB, and their L2 engagement. All the items in the questionnaire were designed on a five-point Likert scale (1 = never true of me at all to 5 = very true of me). For complete understanding, the items were in Chinese. Before the main administration, the first two authors designed a pilot study and asked their colleagues to carry it out among 212 Chinese EFL learners who made up the peer group for the participants in the main study. The questionnaire administrators were asked to take notes of the questions raised by the respondents when they were filling out the questionnaire, but no question was asked. Based on the item analysis of this pilot study, some modifications were made before the questionnaire was finalized. For example, the item from engagement "In the English course, I enjoyed spending time learning with peers in the class" was eliminated since there was no significance between high score group and the low score group ($p > 0.05$) in its item analysis. A detailed description of the scales used in the questionnaire is as follows.

### 3.2.1. Foreign language enjoyment

The participants' enjoyment was measured through an adapted version of the CFLES (Li et al., 2018). It consists of nine items extracted from the CFLES, including three subscales, namely FLE-Private (FLE-P, three items, e.g., I enjoyed learning English.), FLE-Teacher (FLE-T, three items, e.g., The teacher is encouraging. It makes me feel good), and FLE-Atmosphere (FLE-A, three items, e.g., There is a good atmosphere, which makes me feel happy in English class.). All the items were positively phrased. In Li et al. (2018) study, the CFLES was tested among a total of 1,718 Chinese EFL students at the secondary level, and the results showed high reliability with its Cronbach's alpha reaching 0.83. In the present study, Cronbach's alpha coefficient was 0.903.

### 3.2.2. Foreign language classroom anxiety

Anxiety was assessed using the eight-item scale applied in Dewaele and MacIntyre (2014) study, which was developed based on the FLCAS (Horwitz et al., 1986). The eight items investigated the symptoms of anxiety, nervousness, and lack of confidence of the learners. In the scale, six items were phrased to indicate high anxiety, while two reverse-coded items indicated low anxiety. An example is, "Even if I am well prepared for a language class, I feel anxious about it." In Dewaele and MacIntyre (2014) study, Cronbach's alpha coefficient was 0.86. In the current study, Cronbach's alpha coefficient was 0.902.

### 3.2.3. Foreign language learning boredom

The learners' boredom was measured by items extracted from the FLLBS (Li et al., 2021b), which contained seven subscales. Three items

were used to measure foreign language class boredom (FLLB-FLC, e.g., It is difficult for me to concentrate in the English class.), three for under-challenging task boredom (FLLB-UCT, e.g., I believe an analysis of long text in English is really dreary.), three for homework boredom (FLLB-H, e.g., I get bored of too much English homework.), three for teacher-dislike boredom (FLLB-TD, e.g., The English teacher is uninteresting, so the English class is dull.), three for general learning trait boredom (FLLB-GLT, e.g., Not only learning English, studying is dull in general.), three for PowerPoint presentation boredom (FLLB-PPTP, e.g., Reading from the script in the PPT slides bores me.), and three for over-challenging or meaningless task boredom (FLLB-OCMT; e.g., If I cannot understand classmates' presentations, I become really bored). As measured by Cronbach's alpha coefficient, internal consistency was high (0.946).

#### 3.2.4. L2 engagement

L2 engagement was measured from four dimensions through 16 items adapted from Li and Li (2022) questionnaire which was constructed based on the scales developed by Hiver et al. (2020) and Wang et al. (2019). Altogether 12 items were used to assess behavioral engagement (BE, four items, e.g., When I cannot understand in my language class, I stay focused until I do.), emotional engagement (EE, four items, e.g., I enjoy learning new things about languages in class.), and cognitive engagement (CE, four items, e.g., In my language class, I think about different ways to solve a problem.). A total of four items were adapted to evaluate social engagement (SE, four items, e.g., In the English course, I was willing to work with other students, and we could learn from each other.) In the present research, Cronbach's alpha coefficient for the scale was 0.878.

#### 3.2.5. English achievement

The participants' English achievement was measured by their scores on the CET-4, a criteria-related, norm-referenced, large-scale English proficiency test administered on behalf of the Ministry of Education of China. The CET-4 corresponds to the standards set in Guidelines for College English Teaching (College Foreign Language Teaching Advisory Committee, 2020), with the aim to assess the English proficiency of the students and provide information for teachers to improve their pedagogical practices. It has strong social impacts and is widely recognized among institutions and employers in Mainland China. The validity and reliability of the CET-4 scores have been established by the previous study (Yang and Weir, 1998), indicating that it is the best fit for measuring the participants' English achievement.

According to the Syllabus for College English Test—Band Four (National College English Testing Committee, 2016), developed by the National College English Testing Committee, CET-4 is composed of four parts, including listening comprehension, reading comprehension, translation, and writing. After a series of transformation processes, including score weighting, score equating, and score normalization, the reported total test score is 710, with listening comprehension accounting for 35%, reading comprehension 35%, translation 15%, and writing 15%.

### 3.3. Procedure for data collection and analysis

Data collection took place in March 2022. The first author contacted her colleagues individually and asked them to recruit their students as the participants of the present study. Information about the purpose and details of the survey was provided to the teachers and the participants. After obtaining consent, we administered an online questionnaire through Questionnaire Star (a tool for online surveys) to the participants at the beginning of their classes. According to the data provided by Questionnaire Star, the participants took about 15 min on average to complete the questionnaire. As suggested by Dörnyei et al. (2006), there are always a small number of participants who do not take the process seriously in large-scale surveys. Accordingly, before starting the quantitative analyses, we eliminated 14 questionnaires due to incompleteness of the information, which was a low proportion (less than 1.6%) and was therefore considered acceptable.

The quantitative data analyses were performed using the SPSS 23.0 and Amos 21.0 software. Correlations and path analysis were carried out to unpack the relationships between variables. For correlations, effect sizes were used as $r=0.25$ (small effect size), $r=0.40$ (medium effect size), and $r=0.60$ (large effect size; Plonsky and Oswald, 2014). Confirmatory factor analysis (CFA) was run for each construct. The evaluation of the model and CFAs was based on some goodness-of-fit indices. In this study, Chi-square divided by degree of freedom ($\chi^2$/df), comparative fit index (CFI), Tucker-Lewis index (TLI), root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR) were applied. To have a fit model, $\chi^2$/df should be less than 3, CFI and TLI should be above 0.90, and RMSEA and SRMR should be less than 0.08 (Tseng and Schmitt, 2008).

## 4. Results

### 4.1. Validity of the scales

CFA was conducted to assess the measurement model of each construct. Table 1 presents the goodness-of-fit indices for all tested models. Results of the Chi-square test and $\chi^2$/df indices of FLCA, FLLB, and L2 engagement indicated less-than-adequate fit. Nevertheless, it has been suggested that the Chi-square test may not be accurate with sample sizes over 200 (e.g., Bagozzi and Yi, 1988; Schumacker and Lomax, 1996). As the sample in this study was 907, we adopted the other four fit indices to test the adequacy of each model. The fit indices CFI, TLI, SRMR, and RMSEA showed that all four models fitted the data adequately.

### 4.2. Descriptive statistics and correlations

Descriptive statistics and correlations were calculated for all variables. Table 2 displays descriptive statistics including means, standard deviation, as well as skewness and kurtosis. Concerning the

TABLE 1 Goodness of fit indices for the measurement models.

| | $\chi^2$/df | CFI | TLI | SRMR | RMSEA |
|---|---|---|---|---|---|
| FLE | 2.385 | 0.994 | 0.991 | 0.019 | 0.039 |
| FLCA | 4.845 | 0.979 | 0.971 | 0.026 | 0.065 |
| FLLB | 4.178 | 0.968 | 0.960 | 0.050 | 0.059 |
| engagement | 3.910 | 0.967 | 0.959 | 0.022 | 0.038 |

TABLE 2 Descriptive statistics and correlations.

| No. | Variables | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| 1 | FLE | 1 | | | | |
| 2 | FLCA | −0.063* | 1 | | | |
| 3 | FLLB | −0.541** | 0.082* | 1 | | |
| 4 | Engagement | 0.589** | −0.332** | −0.508** | 1 | |
| 5 | English achievement | 0.577** | −0.210** | −0.459** | 0.609** | 1 |
| | Mean | 3.380 | 2.588 | 2.728 | 3.487 | 482.700 |
| | Std. Deviation | 0.666 | 0.740 | 0.744 | 0.544 | 89.512 |
| | Minimum | 1.00 | 1.00 | 1.00 | 1.25 | 221 |
| | Maximum | 5.00 | 4.38 | 4.81 | 4.81 | 651 |
| | Skewness | −0.189 | −0.046 | 0.407 | −0.506 | −2.007 |
| | Kurtosis | 0.331 | −0.837 | 0.694 | 1.684 | 5.687 |

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

TABLE 3 Degree of model fit.

| Fitting index | Acceptable range | Measured value |
|---|---|---|
| $\chi^2$/df | <3 | 3.238 |
| CFI | >0.9 | 0.918 |
| TLI | >0.9 | 0.913 |
| SRMR | <0.08 | 0.078 |
| RMSEA | <0.08 | 0.050 |

normality of the scales, based on Kline (1998) threshold values, all the absolute values of skewness were lower than 3, and those of kurtosis were lower than 10, demonstrating that the data were normally distributed.

Results of Pearson's correlation analyses suggested that FLE was negatively related to FLCA and FLLB, with a small effect size ($r = -0.063$, $p < 0.05$) and a large effect size ($r = -0.541$, $p < 0.01$), respectively, whereas FLCA was found to have a small and positive relation with FLLB ($r = 0.082$, $p < 0.05$). In terms of the relationship between emotions and engagement, FLE was positively related to engagement, with a large effect size ($r = 0.589$, $p < 0.01$). In contrast, both FLCA and FLLB were negatively related to engagement, with medium ($r = -0.332$, $p < 0.01$) and large ($r = -0.508$, $p < 0.01$) correlations, respectively. Concerning the relationships between emotions and English achievement, a positive relation was obtained between FLE and English achievement, with a large effect size ($r = 0.577$, $p < 0.01$). Meanwhile, both FLCA and FLLB were found to have negative relationship with English achievement, with a small effect size ($r = -0.210$, $p < 0.01$) and a medium effect size ($r = 0.459$, $p < 0.01$), respectively. Finally, results indicated that engagement had a positive and large relation ($r = 0.609$, $p < 0.01$) with English achievement.

## 4.3. SEM analysis

To unpack the relationships between L2 emotions, L2 engagement, and English achievement, the proposed model was tested with SEM. By using the maximum likelihood method, the overall goodness-of-fit and the standardized coefficient of each path were calculated. The model fit indices in Table 3 showed that the model achieved close fit as CFI (0.918) and TLI (0.913) were both above 0.90. Additionally, both SRMR (0.078) and RMSEA (0.050) were below 0.08. However, $\chi^2$/df (3.238) was above the maximum limit of 3. Since four of the fit indices met the suggested threshold values, and one was close to the threshold, the model, therefore, fitted the data adequately. Figure 2 depicts the final model for the relationships between L2 emotions, L2 engagement, and English achievement.

The standardized estimates for all paths are presented in Table 4. The model accounted for 76.2% of the variance in L2 engagement. Among the three emotions, FLE positively influenced engagement ($\beta = 0.717$, $p < 0.05$), whereas both FLCA ($\beta = -0.136$, $p < 0.001$) and FLLB ($\beta = -0.185$, $p < 0.05$) negatively affected L2 engagement. The role of emotions and engagement in English achievement was also examined. The model accounted for 47.2% of the variance in English achievement. Both FLE ($\beta = 0.293$, $p < 0.05$) and L2 engagement ($\beta = 0.351$, $p < 0.05$) exerted positive influences on English achievement. Conversely, FLCA had a negative impact ($\beta = -0.121$, $p < 0.05$), while FLLB failed to predict English achievement. Bias-corrected bootstrap tests (2,000 times iterations) were performed to test the mediating role of L2 engagement between emotions and English achievement. The 95% confidence interval (CI) showed that L2 engagement mediated the relationships between FLE and English achievement ($\beta = 0.251$, $p < 0.05$, 95% CI [0.042, 0.437]), FLCA and English achievement ($\beta = -0.048$, $p < 0.05$, 95% CI [−0.108, −0.009]), as well as FLLB and English achievement ($\beta = -0.065$, $p < 0.05$, 95% CI [−0.150, −0.009]).

# 5. Discussion

This study was intended to unpack the relationships between L2 emotions and English achievement, and whether L2 engagement mediated among the constructs. Questionnaire data were collected to investigate the complex relationships and the underlying mechanism. The results revealed that Chinese EFL learners' FLE, FLCA, and FLLB predicted their engagement in foreign language learning, which further influenced their English achievement. Despite the corroboration with the findings of the existing literature, the results of the present study displayed some unique features of Chinese EFL

10.3389/fpsyg.2023.1098916

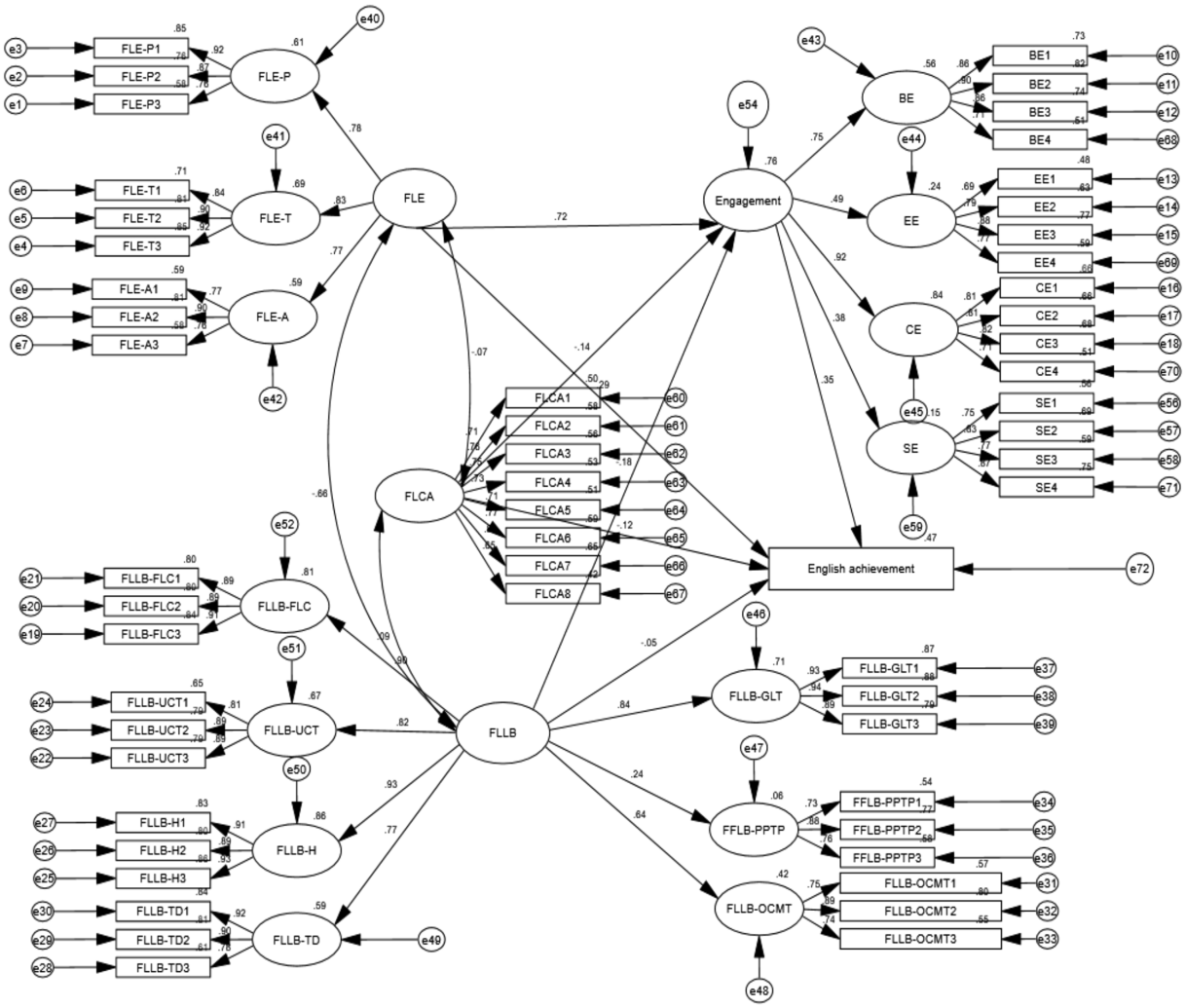



FIGURE 2
The final model of emotions, engagement, and English achievement.

TABLE 4 Unstandardized and standardized path coefficients.

| Path relationship | B | SE | β | Bias-corrected 95%CI | | |
|---|---|---|---|---|---|---|
| | | | | Lower | Upper | *P* |
| FLE→Engagement | 0.687 | 0.059 | 0.717 | 0.587 | 0.822 | 0.002 |
| FLCA→Engagement | −0.106 | 0.030 | −0.136 | −0.200 | −0.083 | 0.000 |
| FLLB→Engagement | −0.119 | 0.050 | −0.185 | −0.282 | −0.085 | 0.002 |
| Engagement→English achievement | 53.448 | 0.147 | 0.351 | 0.042 | 0.610 | 0.030 |
| FLE→English achievement | 42.801 | 0.144 | 0.293 | 0.029 | 0.591 | 0.027 |
| FLCA→English achievement | −14.415 | 0.032 | −0.121 | −0.184 | −0.056 | 0.003 |
| FLLB→English achievement | −5.379 | 0.060 | −0.055 | −0.171 | 0.064 | 0.365 |
| FLE→Engagement→English achievement | 36.742 | 0.103 | 0.251 | 0.042 | 0.437 | 0.024 |
| FLCA→Engagement→English achievement | −5.682 | 0.025 | −0.048 | −0.108 | −0.009 | 0.018 |
| FLLB→Engagement→English achievement | −6.354 | 0.035 | −0.065 | −0.150 | −0.009 | 0.020 |

learners with regard to their emotional variables, engagement, and achievement.

The findings demonstrated that FLE was negatively correlated with FLCA and FLLB, while FLCA and FLLB were positively interrelated, supporting Hypothesis 1. In other words, learners who experienced more enjoyment in their foreign language learning were less likely to be anxious and bored than those who did not enjoy themselves. Furthermore, learners who reported higher levels of anxiety felt more boredom. In line with the findings of Dewaele et al. (2022) and Li and Han (2022), the obtained results ascertained the interconnections between FLE, FLCA, and FLLB, and confirmed and extended the findings about the negative relationship between FLE and FLCA reported in the previous studies (Dewaele and MacIntyre, 2014, 2016; Dewaele et al., 2016; Li et al., 2019; Dewaele and Proietti Ergün, 2020) and the negative correlation between FLE and FLLB (Dewaele and Li, 2021). Based on the finding that the participants who felt more positive emotions like enjoyment tended to experience fewer negative emotions like anxiety and boredom, it can be argued that positive emotions reduce or neutralize the effects of negative emotions (Dewaele et al., 2022), further evidencing the undoing hypothesis put forward by Fredrickson (2003) in EFL context, which proposes that "positive emotions 'undo' the lingering effects of negative emotions" (p: 334).

Results also indicated that all three emotions have significant and direct impacts on L2 engagement, confirming Hypothesis 2. Large, small, and medium correlations were found for the relations between FLE, FLCA, and FLLB with L2 engagement, respectively. Results of path analysis further suggested that FLE was a significant and positive predictor of L2 engagement, whereas FLCA and FLLB negatively predicted L2 engagement. In addition, the R-square ($R^2 = 0.762$) revealed that the three emotions "yielded a significant amount of variance" (Zhang et al., 2022) in L2 engagement. Among the three predictors of L2 engagement, FLE was the strongest one, showing the importance of enjoyment in increasing learners' engagement in language learning. Considering these findings, it can be inferred that when learners enjoyed themselves in language learning, they would be more engaged in foreign language classes. In contrast, they were less likely to participate and involve themselves in foreign language learning when they felt anxious or bored. According to the broaden-and-build theory (Fredrickson, 2001, 2003), such positive emotions as enjoyment can broaden people's momentary thought-action repertoires and build their enduring personal resources, whereas negative ones narrow people's thoughts and actions. Accordingly, we can argue that FLE can improve learners' engagement in learning by enlarging their action repertoire, broadening their momentary mindset, and building their individual resources, while such negative emotions as FLCA and FLLB narrow the same repertoires, thereby decreasing their participation and involvement in language activities. Findings of the present study also provide empirical evidence for the control-value theory (Pekrun, 2006) in the EFL context, which addresses the effects of achievement emotions on student engagement. Regarding the relations found between emotions and engagement, it can be inferred that emotions, specifically enjoyment, play an essential role in engaging learners in the language classroom. The findings, in part, echo those in previous research (Mercer and Dörnyei, 2020; Dewaele and Li, 2021; Guo, 2021; Khajavy, 2021; Derakhshan et al., 2022; Feng and Hong, 2022; Zhao and Yang, 2022).

The third hypothesis that the three emotions significantly influenced English achievement was partly confirmed. Path analysis demonstrated that FLE exerted a positive and significant influence on English achievement. Therefore, learners' enjoyment of L2 learning improved their English achievement, supporting the findings of previous studies (Piechurska-Kuciel, 2017; Li, 2020; Botes et al., 2021; Jin and Zhang, 2021). This finding also evidences the control-value theory (Pekrun, 2006), which assumes that activating positive emotions like enjoyment benefit academic achievement. Unlike the positive link between FLE and English achievement, FLCA was found to be a negative predictor, indicating that learners who felt anxious about their L2 learning were less likely to perform well in tests, which has been well-documented (Shao et al., 2013; Li et al., 2019; Zhang, 2019; Botes et al., 2020; Dewaele and Proietti Ergün, 2020; Dewaele et al., 2022; Li and Han, 2022). The results also indicated that the effect of FLE ($\beta = 0.293$, $p < 0.05$) on English achievement outweighed that of FLCA ($\beta = -0.121$, $p < 0.05$), showing that FLE might be more relevant to learners' performance in tests, which resonates with the findings of Dewaele and Alfawzan (2018) along with Li and Wei (2022). Surprisingly, FLLB failed to have a significant and direct effect on English achievement when combined with enjoyment and anxiety. This finding is inconsistent with most existing literature that reported a negative impact of boredom on achievement (Pekrun et al., 2009, 2010; Ahmed et al., 2013; Pekrun et al. 2014). This might be because prior research did not involve FLE, FLCA, and FLLB in one model. In the present study, the regression weights of boredom were much lower than those of enjoyment and anxiety as predictors of English achievement, which might result in the non-significant predictive power of FLLB on achievement. Nonetheless, this is not without precedence. Dewaele et al. (2022) also reported that FLLB did not predict achievement when combined with FLE and FLCA. The links between emotions and achievement in this study partly support the findings of Li and Han (2022) research on the relationship between three emotions and English achievement, but are partly inconsistent with theirs. The agreement concerns the significant and negative influence of FLCA on learners' actual English achievement and the non-significant role of FLLB in learning outcomes. Still, their finding about FLE as a non-significant predictor of achievement differs from the predictive power of FLE displayed here. One possible explanation might be the differences in the learning environment. Li and Han (2022) explored learners' emotions and their English achievement in the context of online learning during the period of COVID-19 in China. Compared with the traditional learning environment, this change may increase learners' anxiety, influencing their enjoyment of L2 learning. Nevertheless, the present study did not highlight the learning environment, and the participants took their CET-4 offline. This means that the learning environment might be taken into account when conducting the relevant study.

Hypothesis 4, that L2 engagement had significant effects on English achievement, was also confirmed. Results displayed a positive correlation between L2 engagement and English achievement. Path analysis further showed that L2 engagement exerted a moderate effect ($\beta = 0.351$, $p < 0.05$) on learning outcomes, being the strongest predictor among all the significant predictors of achievement in this model. This indicated that L2 engagement had its unique variance in L2 achievement, thus demonstrating the importance of L2 engagement in foreign language learning. Based on the findings, it can be inferred that when learners engage in foreign language classroom activities and put effort into language learning, they are more likely to score high on tests. This outcome supports those results that L2 engagement,

whether it was treated as a multidimensional concept or a single dimension, had a positive and predictive impact on L2 achievement (Masgoret and Gardner, 2003; Eren and Rakıcıoğlu-Söylemez, 2020; Khajavy, 2021; Feng and Hong, 2022; Kang and Wu, 2022). From a broader perspective, the role of engagement as a critical contributor to learning and academic success was further confirmed, corroborating the findings of the existing literature (Klem and Connell, 2004; Reeve and Lee, 2014; Wang and Fredricks, 2014).

Path analysis also indicated that L2 engagement was a mediator between FLE, FLCA, FLLB, and English achievement, supporting Hypothesis 5. The results revealed that learners' FLE exerted a significant and positive influence on their engagement in EFL class, which, in turn, positively impacted their English achievement. On the contrary, their FLCA and FLLB negatively affected their engagement in class and then would incur lower scores on tests. The mediating effect of engagement between FLE and English achievement ($\beta = 0.251$, $p < 0.05$) is larger than those in the other two pathways ($\beta = -0.048$, $p < 0.05$; $\beta = -0.065$, $p < 0.05$). The findings partly correspond to Feng and Hong (2022), which reported the mediating role of behavioral engagement between FLE, FLCA, and self-reported achievement, with a larger mediating effect size of engagement between FLE and SRA. Moreover, these results were partly consistent with Kang and Wu (2022) which revealed the mediating role of behavioral engagement between learners' academic enjoyment and their English achievement. Likewise, the finding that L2 engagement as a mediator between L2 enjoyment and L2 reading comprehension was reported in Khajavy (2021) study. Furthermore, it was noted that FLLB indirectly affected English achievement through L2 engagement, although it failed to affect achievement directly. According to the control-value theory (Pekrun, 2006), for boredom, a negative, deactivating, and activity-related achievement emotion, control, and values refer to the action, but not outcomes. In other words, the attentional focus of a learner experiencing boredom is on the activity of learning, not on their grades on tests. In this regard, we can argue that if learners feel bored in their EFL class, this emotion is less likely to influence their achievement directly, but would decrease their participation and involvement in EFL class activities. The disengagement in learning would, in turn, impact their academic achievement. This is in accord with Pekrun and Linnenbrink-Garcia (2012) emphasis that emotions can influence students' engagement, which then affects their academic learning and achievement, as has been reported in other scholarship as well (Pekrun, 1992, 2006; Linnenbrink, 2007).

## 6. Conclusion and pedagogical implication

The present study was one of the first attempts to unpack the relationships between learners' emotions (FLE, FLCA, FLLB), engagement, and their actual English achievement in the EFL context. Through examining whether English achievement was influenced by FLE, FLCA, and FLLB, with the mediating effect of engagement, we found that learners' FLE and FLCA, respectively, exerted a significantly positive impact and a significantly negative influence on their English achievement both directly and indirectly through the mediation of engagement, whereas FLLB failed to have such an effect. Only through the mediation of engagement could FLLB affect English achievement, indicating that boredom experienced by learners in EFL class did not directly lead to lower achievement. This negative emotion, however, would decrease learners' participation and involvement in language learning, which, in turn, resulted in lower marks on tests. Another important finding was the significant interconnections between FLE, FLCA, and FLLB. These findings contributed to the literature by uncovering the mechanism underlying the relationships between emotions, engagement, and achievement in the Chinese EFL context, broadened the nomological network of emotions and engagement, and provided empirical evidence to Pekrun's control-value theory (2006) and Fredrickson (2003) undoing hypothesis in the field of SLA. The current research also highlighted the significance of emotions and engagement in EFL learning, and underscored the importance of learners' emotions in influencing their engagement.

The findings of the present study had some pedagogical implications for Chinese EFL teachers and practitioners teaching at the tertiary level. The finding that FLE had the largest predictive effects on both engagement and English achievement when combined with FLCA and FLLB suggests that EFL teachers need to take steps to boost learners' enjoyment in class. For example, they can provide learning materials that learners are interested in and feel capable of dealing with (Pekrun, 2006). In addition, pleasant activities that are controllable and valued positively by learners are also recommended to enhance their FLE, thus increasing their engagement and achievement. Moreover, teacher-related factors, such as emotional support, use of humor, and positive mood, will also play an essential role in affecting learners' FLE (Dewaele et al., 2019b). But teachers do not need to be excessively anxious about eliminating FLCA and FLLB experienced by learners since positive and negative emotions are like "the right and left feet of language learner" (Dewaele and MacIntyre, 2016, p.215), meaning that a balance will be found between both. Positive emotions like FLE may neutralize the negative effects of the negative emotions according to the results of the correlation analysis. Additionally, given the strongest predictive impact of L2 engagement on English achievement among all the predictors included in the model, teachers are suggested to plan motivating and enjoyable activities to maintain learners' focus and concentration in EFL class. For instance, teachers can design activities that require a high level of learners' participation and interaction, instead of being teacher-led (Sulis, 2022). The other finding that engagement mediated FLLB to influence English achievement significantly indicates the negative causal effect of learners' boredom experiences on their participation in class, which, in turn, impacts their achievement. By implication, teachers should pay due attention to minimizing learners' boredom in ELF classes. For example, they can design activities or change the learning environment to arouse learners' interest in language learning as interest can protect against feeling bored (Pekrun et al., 2010).

## 7. Limitations and suggestions for further research

This study has evidenced the relationships between three L2 emotions, L2 engagement, and achievement in the EFL context, offering essential implications. However, there are some limitations to be acknowledged. Firstly, the participants of the present study were recruited from one university in China, thus influencing the generalizability of the results in other settings. Hence, future studies can select larger samples of participants at different proficiency levels from

various EFL contexts. Secondly, the present research was cross-sectional, failing to capture dynamic features of emotions and engagement and their impacts on achievement. It is advised that longitudinal studies be carried out to further investigate the dynamic changes of emotions, engagement, and achievement among EFL learners and their reciprocal rather than unidirectional causation across time. Finally, the current study only employed a quantitative method to measure emotions and engagement, which cannot reveal the detailed and dynamic features of the constructs due to individual differences. Further studies are recommended to integrate the quantitative data with qualitative ones, such as data collected from interviews and classroom observation, to achieve a deeper understanding of the in-depth characteristics of the constructs and their relationship.

## 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 were not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

HW and YW conceptualized, designed the study, and drafted the manuscript. HW, YW, and SL collected the data and processed the data. All authors contributed to the article and approved the submitted version.

## Funding

This research was supported by the Social Science Foundation of Liaoning Province (Grant no. L21AYY002).

## Conflict of interest

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

## Publisher's note

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

## References

Ahmed, W., van der Werf, G., Kuyper, H., and Minnaert, A. (2013). Emotions, self-regulated learning, and achievement in mathematics: a growth curve analysis. *J. Educ. Psychol.* 105, 150–161. doi: 10.1037/a0030160

Anderman, E. M., and Patrick, H. (2012). "Achievement goal theory, conceptualization ofability/intelligence, and classroom climate" in *Handbook of research on student engagement*. eds. S. L. Christenson, A. L. Reschly and C. Wylie (Boston, MA: Springer US), 173–191.

Bagozzi, R. P., and Yi, Y. (1988). On the evaluation of structural equation models. *J. Acad. Mark. Sci.* 16, 74–94. doi: 10.1007/bf02723327

Botes, E., Dewaele, J.-M., and Greiff, S. (2020). The foreign language classroom anxiety scale and academic achievement: an overview of the prevailing literature and a meta-analysis. *Journal for the Psychology of Language Learning* 2, 26–56. doi: 10.52598/jpll/2/1/3

Botes, E., Dewaele, J.-M., and Greiff, S. (2021). The development and validation of the short form of the foreign language enjoyment scale. *Mod. Lang. J.* 105, 858–876. doi: 10.1111/modl.12741

Botes, E., Dewaele, J.-M., and Greiff, S. (2022). Taking stock: a meta-analysis of the effects of foreign language enjoyment. *Studies in Second Language Learning and Teaching* 12, 205–232. doi: 10.14746/ssllt.2022.12.2.3

Carver, C., Jung, D., and Gurzynski-Weiss, L. (2021). "Examining learner engagement in relationship to learning and communication mode" in *Student engagement in the language classroom*. eds. P. Hiver, A. H. Al-Hoorie and S. Mercer (Bristol: Multilingual Matters), 120–142.

Chapman, K. E. (2013). *Boredom in the German foreign language classroom*. [doctoral dissertation]. [Madison (WI)]: University of Wisconsin-Madison

College Foreign Language Teaching Advisory Committee. (2020). *Guidelines for college English teaching*. Beijing: Higher Education Press.

Derakhshan, A., Fathi, J., Pawlak, M., and Kruk, M. (2022). Classroom social climate, growth language mindset, and student engagement: the mediating role of boredom in learning English as a foreign language. *J. Multiling. Multicult. Dev.*, 1–19. doi: 10.1080/01434632.2022.2099407

Derakhshan, A., Kruk, M., Mehdizadeh, M., and Pawlak, M. (2021a). Activity-induced boredom in online EFL classes. *ELT J.* 76, 58–68. doi: 10.1093/elt/ccab072

Derakhshan, A., Kruk, M., Mehdizadeh, M., and Pawlak, M. (2021b). Boredom in online classes in the Iranian EFL context: sources and solutions. *System* 101:102556. doi: 10.1016/j.system.2021.102556

Dewaele, J.-M. (2002). Psychological and sociodemographic correlates of communicative anxiety in L2 and L3 production. *Int. J. Biling.* 6, 23–38. doi: 10.1177/13670069020060010201

Dewaele, J.-M. (2013). The link between foreign language classroom anxiety and psychoticism, extraversion, and neuroticism among adult bi- and multilinguals. *Mod. Lang. J.* 97, 670–684. doi: 10.1111/j.1540-4781.2013.12036.x

Dewaele, J.-M., and Alfawzan, M. (2018). Does the effect of enjoyment outweigh that of anxiety in foreign language performance? *Studies in Second Language Learning and Teaching* 8, 21–45. doi: 10.14746/ssllt.2018.8.1.2

Dewaele, J.-M., and Al-Saraj, T. (2015). Foreign language classroom anxiety of Arab learners of English: the effect of personality, linguistic and sociobiographical variables. *Studies in Second Language Learning and Teaching* 5, 205–228. doi: 10.14746/ssllt.2015.5.2.2

Dewaele, J.-M., Botes, E., and Greiff, S. (2022). Sources and effects of foreign language enjoyment, anxiety, and boredom: a structural equation modeling approach. *Stud. Second. Lang. Acquis.* 1–19, 1–19. doi: 10.1017/s0272263122000328

Dewaele, J.-M., Chen, X., Padilla, A. M., and Lake, J. (2019b). The flowering of positive psychology in foreign language teaching and acquisition research. *Front. Psychol.* 10:2128. doi: 10.3389/fpsyg.2019.02128

Dewaele, J.-M., and Li, C. (2021). Teacher enthusiasm and students' social-behavioral learning engagement: the mediating role of student enjoyment and boredom in Chinese EFL classes. *Lang. Teach. Res.* 25, 922–945. doi: 10.1177/136216882110145

Dewaele, J.-M., and MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching* 4, 237–274. doi: 10.14746/ssllt.2014.4.2.5

Dewaele, J.-M., and MacIntyre, P. D. (2016). "Foreign language enjoyment and foreign language classroom anxiety: the right and left feet of the language learner" in *Positive psychology in SLA*. eds. P. MacIntyre, T. Gregersen and S. Mercer (Bristol: Multilingual Matters), 215–236.

Dewaele, J.-M., MacIntyre, P., Boudreau, C., and Dewaele, L. (2016). Do girls have all the fun? Anxiety and enjoyment in the foreign language classroom. *Theory and Practice of Second Language Acquisition* 2, 41–63.

Dewaele, J.-M., Magdalena, A. F., and Saito, K. (2019a). The effect of perception of teacher characteristics on Spanish EFL learners' anxiety and enjoyment. *Mod. Lang. J.* 103, 412–427. doi: 10.1111/modl.12555

Dewaele, J.-M., Petrides, K. V., and Furnham, A. (2008). Effects of trait emotional intelligence and sociobiographical variables on communicative anxiety and foreign language anxiety among adult multilinguals: a review and empirical investigation. *Lang. Learn.* 58, 911–960. doi: 10.1111/j.1467-9922.2008.00482.x

Dewaele, J.-M., and Proietti Ergün, A. L. (2020). How different are the relations between enjoyment, anxiety, attitudes/motivation and course marks in pupils' Italian and English as foreign languages? *Journal of the European Second Language Association* 4, 45–57. doi: 10.22599/jesla.65

Dewaele, J.-M., Witney, J., Saito, K., and Dewaele, L. (2017). Foreign language enjoyment and anxiety: the effect of teacher and learner variables. *Lang. Teach. Res.* 22, 676–697. doi: 10.1177/1362168817692161

Dörnyei, Z., Csizér, K., and Németh, N. (2006). *Motivation, language attitudes and globalisation: a Hungarian perspective*. Bristol: Multilingual Matters.

Eren, A., and Rakıcıoğlu-Söylemez, A. (2020). Language mindsets, perceived instrumentality, engagement and graded performance in English as a foreign language students. *Lang. Teach. Res.* 1-31:136216882095840. doi: 10.1177/1362168820958400

Fathi, J., and Mohammaddokht, F. (2021). Foreign language enjoyment and anxiety as the correlates of the ideal L2 self in the English as a foreign language context. *Front. Psychol.* 12:790648. doi: 10.3389/fpsyg.2021.790648

Feng, E., and Hong, G. (2022). Engagement mediates the relationship between emotion and achievement of Chinese EFL learners. *Front. Psychol.* 13:895594. doi: 10.3389/fpsyg.2022.895594

Filsecker, M., and Kerres, M. (2014). Engagement as a volitional construct. *Simul. Gaming* 45, 450–470. doi: 10.1177/1046878114553569

Finn, J. D. (1989). Withdrawing from school. *Rev. Educ. Res.* 59, 117–142. doi: 10.3102/00346543059002117

Fredricks, J. A., Blumenfeld, P. C., and Paris, A. H. (2004). School engagement: potential of the concept, state of the evidence. *Rev. Educ. Res.* 74, 59–109. doi: 10.3102/00346543074001059

Fredricks, J. A., Filsecker, M., and Lawson, M. A. (2016). Student engagement, context, and adjustment: addressing definitional, measurement, and methodological issues. *Learn. Instr.* 43, 1–4. doi: 10.1016/j.learninstruc.2016.02.002

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: the broaden-and-build theory of positive emotions. *Am. Psychol.* 56, 218–226. doi: 10.1037//0003-066x.56.3.218

Fredrickson, B. L. (2003). The value of positive emotions. *Am. Sci.* 91, 330–335. doi: 10.1511/2003.4.330

Guo, Y. (2021). Exploring the dynamic interplay between foreign language enjoyment and learner engagement with regard to EFL achievement and absenteeism: a sequential mixed methods study. *Front. Psychol.* 12:766058. doi: 10.3389/fpsyg.2021.766058

Hashimoto, Y. (2002). Motivation and willingness to communicate as predictors of reported L2 use: the Japanese context. *Second Language Studies* 20, 29–70.

Hiver, P., Al-Hoorie, A. H., and Mercer, S. (2021a). *Student engagement in the language classroom*. Bristol: Channel View Publications.

Hiver, P., Al-Hoorie, A. H., Vitta, J. P., and Wu, J. (2021b). Engagement in language learning: a systematic review of 20 years of research methods and definitions. *Lang. Teach. Res.* 1–30:136216882110012. doi: 10.1177/13621688211001289

Hiver, P., Zhou, S., Tahmouresi, S., Sang, Y., and Papi, M. (2020). Why stories matter: exploring learner engagement and metacognition through narratives of the L2 learning experience. *System* 91:102260. doi: 10.1016/j.system.2020.102260

Horwitz, E. K., Horwitz, M. B., and Cope, J. (1986). Foreign language classroom anxiety. *Mod. Lang. J.* 70, 125–132. doi: 10.1111/j.1540-4781.1986.tb05256.x

Jin, Y., and Zhang, L. J. (2019). A comparative study of two scales for foreign language classroom enjoyment. *Percept. Mot. Skills* 126, 1024–1041. doi: 10.1177/0031512519864471

Jin, Y., and Zhang, L. J. (2021). The dimensions of foreign language classroom enjoyment and their effect on foreign language achievement. *Int. J. Biling. Educ. Biling.* 24, 948–962. doi: 10.1080/13670050.2018.1526253

Kang, X., and Wu, Y. (2022). Academic enjoyment, behavioral engagement, self-concept, organizational strategy and achievement in EFL setting: a multiple mediation analysis. *PLoS One* 17:e0267405. doi: 10.1371/journal.pone.0267405

Khajavy, G. (2021). "Modeling the relations between foreign language engagement, emotions, grit and reading achievement" in *Student engagement in the language classroom*. eds. P. Hiver, A. Al-Hoorie and S. Mercer (Bristol: Multilingual Matters), 241–259.

Klem, A. M., and Connell, J. P. (2004). Relationships matter: linking teacher support to student engagement and achievement. *J. Sch. Health* 74, 262–273. doi: 10.1111/j.1746-1561.2004.tb08283.x

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Kruk, M. (2016a). "Investigating the changing nature of boredom in the English language classroom: results of a study" in *Nowy wymiarfilologii*. eds. A. Dlutek and D. Pietrzak (Plock: Wydawnictwo NaukowePanstwowej Wyzszej Szkoly Zawodowej w Plocku), 252–263.

Kruk, M. (2016b). Variations in motivation, anxiety and boredom in learning English in second life. *The EUROCALL Review* 24, 25–39. doi: 10.4995/eurocall.2016.5693

Kruk, M., and Zawodniak, J. (2017). Nuda a praktyczna nauka języka angielskiego. *Neofilolog* 49, 115–131. doi: 10.14746/n.2017.49.1.07

Kruk, M., and Zawodniak, J. (2018). "Boredom in practical English language classes: insights from interview data" in *Interdisciplinary views on the English language, literature and culture*. eds. L. Szymanski, J. Zawodniak, A. Lobodziec and M. Smoluk (Poland: Uniwersytet Zielonogórski), 177–191.

Li, C. (2020). A positive psychology perspective on Chinese EFL students' trait emotional intelligence, foreign language enjoyment and EFL learning achievement. *J. Multiling. Multicult. Dev.* 41, 246–263. doi: 10.1080/01434632.2019.1614187

Li, C. (2021). A control–value theory approach to boredom in English classes among university students in China. *Mod. Lang. J.* 105, 317–334. doi: 10.1111/modl.12693

Li, C., Dewaele, J.-M., and Hu, Y. (2021b). Foreign language learning boredom: conceptualization and measurement. *Applied Linguistics Review* 0, 1–27. doi: 10.1515/applirev-2020-0124

Li, C., Dewaele, J.-M., and Jiang, G. (2019). The complex relationship between classroom emotions and EFL achievement in China. *Applied Linguistics Review* 11, 485–510. doi: 10.1515/applirev-2018-0043

Li, C., and Han, Y. (2022). The predictive effects of foreign language anxiety, enjoyment, and boredom on learning outcomes in online English classrooms. *Modern Foreign Languages «*现代外语*»* 45, 207–219.

Li, C., Huang, J., and Li, B. (2021a). The predictive effects of classroom environment and trait emotional intelligence on foreign language enjoyment and anxiety. *System* 96:102393. doi: 10.1016/j.system.2020.102393

Li, C., Jiang, G., and Dewaele, J.-M. (2018). Understanding Chinese high school students' foreign language enjoyment: validation of the Chinese version of the foreign language enjoyment scale. *System* 76, 183–196. doi: 10.1016/j.system.2018.06.004

Li, Z., and Li, J. (2022). Using the flipped classroom to promote learner engagement for the sustainable development of language skills: a mixed-methods study. *Sustainability* 14:5983. doi: 10.3390/su14105983

Li, C., and Wei, L. (2022). Anxiety, enjoyment, and boredom in language learning amongst junior secondary students in rural China: how do they contribute to L2 achievement? *Stud. Second. Lang. Acquis.* 1–16, 1–16. doi: 10.1017/S0272263122000031

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 (Burlington: Academic Press), 107–124.

MacIntyre, P. D. (2017). "An overview of language anxiety research and trends in its development" in *New insights into language anxiety: Theory, research and educational implications*. eds. C. Gkonou, M. Daubney and J. Dewaele (Bristol: Multilingual Matters), 11–30.

MacIntyre, P. D. (2021). "Exploring applications of positive psychology in SLA" in *Positive psychology in second and foreign language education*. eds. K. Budzinska and O. Majchrzak (Cham: Springer), 3–17.

MacIntyre, P. D., Baker, S. C., Clément, R., and Donovan, L. A. (2002). Sex and age effects on willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Lang. Learn.* 52, 537–564. doi: 10.1111/1467-9922.00194

Maroldo, G. K. (1986). Shyness, boredom, and grade point average among college students. *Psychol. Rep.* 59, 395–398. doi: 10.2466/pr0.1986.59.2.395

Masgoret, A. M., and Gardner, R. C. (2003). Attitudes, motivation, and second language learning: a meta-analysis of studies conducted by Gardner and associates. *Lang. Learn.* 53, 167–210. doi: 10.1111/1467-9922.00227

Mercer, S., and Dörnyei, Z. (2020). *Engaging language learners in contemporary classrooms*. Cambridge: Cambridge University Press.

Mohammad Hosseini, H., Fathi, J., Derakhshesh, A., and Mehraein, S. (2022). A model of classroom social climate, foreign language enjoyment, and student engagement among English as a foreign language learners. *Front. Psychol.* 13:933842. doi: 10.3389/fpsyg.2022.933842

National College English Testing Committee. (2016). *Syllabus for college English test–band four*. Shanghai: Shanghai Jiao Tong University Press.

Oga-Baldwin, W. L. Q. (2019). Acting, thinking, feeling, making, collaborating: the engagement process in foreign language learning. *System* 86:102128. doi: 10.1016/j.system.2019.102128

Pawlak, M., Derakhshan, A., Mehdizadeh, M., and Kruk, M. (2022). Boredom in online English language classes: mediating variables and coping strategies. *Lang. Teach. Res.* 1-26:136216882110649. doi: 10.1177/13621688211064944

Pekrun, R. (1992). The impact of emotions on learning and achievement: towards a theory of cognitive/motivational mediators. *Appl. Psychol.* 41, 359–376. doi: 10.1111/j.1464-0597.1992.tb00712.x

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

Pekrun, R., Elliot, A. J., and Maier, M. A. (2009). Achievement goals and achievement emotions: testing a model of their joint relations with academic performance. *J. Educ. Psychol.* 101, 115–135. doi: 10.1037/a0013383

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., and Perry, R. P. (2010). Boredom in achievement settings: exploring control–value antecedents and performance outcomes of a neglected emotion. *J. Educ. Psychol.* 102, 531–549. doi: 10.1037/a0019243

Pekrun, R., Goetz, T., Titz, W., and Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: a program of qualitative and quantitative research. *Educ. Psychol.* 37, 91–105. doi: 10.1207/S15326985EP3702_4

Pekrun, R., Hall, N. C., Goetz, T., and Perry, R. P. (2014). Boredom and academic achievement: testing a model of reciprocal causation. *J. Educ. Psychol.* 106, 696–710. doi: 10.1037/a0036006

Pekrun, R., and Linnenbrink-Garcia, L. (2012). "Academic emotions and student engagement" in *Handbook of research on student engagement*. eds. S. L. Christenson, A. L. Reschly and C. Wylie (Boston, MA: Springer US), 259–282.

Peterson, C. (2006). *A primer in positive psychology*. New York: Oxford University Press.

Piechurska-Kuciel, E. (2017). "L2 or L3? Foreign language enjoyment and proficiency" in *Multiculturalism, multilingualism and the self*. eds. D. Gabrys-Barker, D. Galajda, A. Wojtaszek and P. Zakrajewski (Cham: Springer), 97–111.

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

Plonsky, L., Sudina, E., and Teimouri, Y. (2022). Language learning and emotion. *Lang. Teach.* 55, 346–362. doi: 10.1017/s0261444821000434

Reeve, J., and Lee, W. (2014). Students' classroom engagement produces longitudinal changes in classroom motivation. *J. Educ. Psychol.* 106, 527–540. doi: 10.1037/a0034934

Reeve, J., and Tseng, C.-M. (2011). Agency as a fourth aspect of students' engagement during learning activities. *Contemp. Educ. Psychol.* 36, 257–267. doi: 10.1016/j.cedpsych.2011.05.002

Reschly, A., and Christenson, S. L. (2012). "Jingle, jangle, and conceptual haziness: evolution and future directions of the engagement construct" in *Handbook of research on student engagement*. eds. S. L. Christenson, A. L. Reschly and C. Wylie (Boston, MA: Springer), 3–20.

Schumacker, R. E., and Lomax, R. G.. (1996). *A Beginner's guide to structural equation modelling*. Mahwah, NJ: Lawrence Erlbaum.

Schunk, D. H., and Mullen, C. A. (2012). "Self-efficacy as an engaged learner" in *Handbook of research on student engagement*. eds. S. L. Christenson, A. L. Reschly and C. Wylie (Boston, MA: Springer US), 219–236.

Seligman, M. E. P., and Csikszentmihalyi, M. (2000). Positive psychology: an introduction. *Am. Psychol.* 55, 5–14. doi: 10.1037/0003-066X.55.1.5

Shao, K., Yu, W., and Ji, Z. (2013). An exploration of Chinese EFL students' emotional intelligence and foreign language anxiety. *Mod. Lang. J.* 97, 917–929. doi: 10.1111/j.1540-4781.2013.12042.x

Sulis, G. (2022). Engagement in the foreign language classroom: micro and macro perspectives. *System* 110:102902. doi: 10.1016/j.system.2022.102902

Sulis, G., and Philp, J. (2021). "Exploring connections between classroom environment and engagement in the foreign language classroom" in *Student engagement in the language classroom*. eds. P. Hiver, A. H. Al-Hoorie and S. Mercer (Bristol: Multilingual Matters), 101–119.

Svalberg, A. M.-L. (2009). Engagement with language: interrogating a construct. *Lang. Aware.* 18, 242–258. doi: 10.1080/09658410903197264

Tóth, Z. (2008). A foreign language anxiety scale for Hungarian learners of English. Working papers in language. *Pedagogy* 2, 55–78.

Tseng, W.-T., and Schmitt, N. (2008). Toward a model of motivated vocabulary learning: a structural equation modeling approach. *Language Learning* 58:357–400. doi: 10.1111/j.1467-9922.2008.00444.x

Wang, Y., Derakhshan, A., and Zhang, L. J. (2021). Researching and practicing positive psychology in second/foreign language learning and teaching: the past, current status and future directions. *Front. Psychol.* 12:731721. doi: 10.3389/fpsyg.2021.731721

Wang, M. T., and Fredricks, J. A. (2014). The reciprocal links between school engagement, youth problem behaviors, and school dropout during adolescence. *Child Dev.* 85, 722–737. doi: 10.1111/cdev.12138

Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., and Linn, J. S. (2016). The math and science engagement scales: scale development, validation, and psychometric properties. *Learn. Instr.* 43, 16–26. doi: 10.1016/j.learninstruc.2016.01.008

Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., and Linn, J. S. (2019). Conceptualization and assessment of adolescents' engagement and disengagement in school: a multidimensional school engagement scale. *Eur. J. Psychol. Assess.* 35, 592–606. doi: 10.1027/1015-5759/a000431

Wang, M. T., and Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. *Am. Educ. Res. J.* 47, 633–662. doi: 10.3102/0002831209361209

Yang, H., and Weir, C. J. (1998). *The CET validation study*. Shanghai: Shanghai Foreign Language Education Press.

Zhang, X. (2019). Foreign language anxiety and foreign language performance: a meta-analysis. *Mod. Lang. J.* 103, 763–781. doi: 10.1111/modl.12590

Zhang, L. J., Saeedian, A., and Fathi, J. (2022). Testing a model of growth mindset, ideal L2 self, boredom, and WTC in an EFL context. *J. Multiling. Multicult. Dev.* 1–16, 1–16. doi: 10.1080/01434632.2022.2100893

Zhao, Y., and Yang, L. (2022). Examining the relationship between perceived teacher support and students' academic engagement in foreign language learning: enjoyment and boredom as mediators. *Front. Psychol.* 13:987554. doi: 10.3389/fpsyg.2022.987554

Zhou, S., Hiver, P., and Al-Hoorie, A. H. (2021). "Measuring L2 engagement: a review of issues and applications" in *Student engagement in the language classroom*. eds. P. Hiver, A. H. Al-Hoorie and S. Mercer (Bristol: Multilingual Matters), 75–98.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 27 February 2023
DOI 10.3389/fpsyg.2023.1120701



OPEN ACCESS

EDITED BY
Hui Zhang,
Nanjing Normal University, China

REVIEWED BY
Ramazan Yilmaz,
Bartin University, Türkiye
Xuyan Qiu,
Hong Kong Polytechnic University, Hong Kong SAR, China

*CORRESPONDENCE
Ahmet Güneyli
✉ aguneyli@eul.edu.tr

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

RECEIVED 10 December 2022
ACCEPTED 06 February 2023
PUBLISHED 27 February 2023

CITATION
Karakaş Kurt E and Güneyli A (2023) Teaching the Turkish language to foreigners at higher education level in Northern Cyprus: An evaluation based on self-perceived dominant intelligence types, twenty-first-century skills and learning technologies. *Front. Psychol.* 14:1120701. doi: 10.3389/fpsyg.2023.1120701



# Teaching the Turkish language to foreigners at higher education level in Northern Cyprus: An evaluation based on self-perceived dominant intelligence types, twenty-first-century skills and learning technologies

Esra Karakaş Kurt[1] and Ahmet Güneyli[2]*

[1]Department of Turkish and Social Sciences Education, Education Faculty, Cyprus International University, Nicosia, Cyprus, [2]Department of Turkish and Social Sciences Education, Education Faculty, European University of Lefke, Lefke, Cyprus

**Introduction:** There are many foreign students in higher education in Northern Cyprus. Both the academic and life skills of these students depend on attaching the necessary importance to their Turkish language teaching. The goal of this study is to examine how university students employ learning technology, twenty-first-century abilities, and perceived categories of intelligence in the process of learning a foreign language.

**Methods:** In line with the quantitative research design, this study utilized a descriptive approach. Purposeful and convenience sampling methods were used to create the study sample. As a result, the institution in Northern Cyprus with the largest international student body was chosen. At this university, one of the authors of this study has been employed, and Turkish is the language of teaching. The study sample consisted of 431 university students who took Turkish as a foreign language in the 2021–2022 academic year at the selected university.

**Results:** The results of the study revealed a weak yet statistically significant correlation between twenty-first-century skills and usage of foreign language-learning technologies. Additionally, students' twenty-first-century skill scores differed significantly, whereas their foreign language-learning technology scale scores did not match their self-perceived intelligence types.

**Conclusion:** The research's findings indicate that students in higher education possess twenty-first-century skills. Based on this finding, it is possible to engage students in the courses and accomplish effective foreign language acquisition if foreign language education is carried out in accordance with modern methodologies and based on twenty-first-century abilities. It has been revealed in this study that it is important to include social learning rather than individual and competitive learning in foreign language education classes.

KEYWORDS

**teaching Turkish as a foreign language, higher-education students, twenty-first-century skills, learning technologies, perceived dominant intelligence types**

## 1. Introduction

Considering that the world is digitalized and globalized more with each passing day, it is self-evident that the number of people who speak a foreign language will increase progressively all over the world as it is in EU countries. It has been stated in the Organization for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA) 2018 Report, which is based on data obtained from 75 countries, that more than 95% of students either speak more than one language or are learning at least one foreign language (OECD and PISA, 2018). These data indicate that we need to focus on teaching and learning foreign languages even more. No matter how foreign language teaching or learning is conducted, taking into account the current state of technology, enriched, interactive digital resources and the digital platforms where these materials are presented rank among the most crucial tools in the process. Today, there are rich contents and various tools (z-books, digital games, speech bots, web 2.0 tools, etc.) that can be used by both foreign language teachers and learners for the improvement of reading, listening, writing, and speaking skills. In short, as in all other areas of life, all learning-teaching activities at school are affected by technology and media-oriented lives. In light of this, it is evident that modern individuals require functional skills such as media literacy, information literacy, and computer and information technology (Partnership for 21st Century Learning, 2021).

The pandemic caused by the coronavirus disease (COVID-19) has brought about a compulsory transformation in education. In this context, it has been observed that learning technologies and virtual environments were particularly effective on the learning skills of students who were confined to their homes (Adedoyin and Soykan, 2020). While technology-supported teaching skills or competencies of teachers were of more importance before the pandemic, the issue of how students can learn more effectively by using technology has gained more importance with the pandemic (Daniel, 2020).

Recent studies on foreign language education demonstrated that there remains a strong and growing demand for employees with high language and cultural competencies in both the private and public sectors, notably in healthcare, social services, translation and interpretation services, travel, and tourism sectors (Damari et al., 2018; Looney and Lusin, 2018). However, studies have also shown that the foreign language-learning process is not efficient enough to meet this demand of the business sector (Stein-Smith, 2016; Quicios, 2018). Studies conducted on students revealed that traditional approaches tend to dominate the foreign language-learning process. For example, although there is a tendency toward adopting learner-centered approaches in education in general, traditional, and conceptual approaches continue to be used frequently in language education (Kim, 2019).

It is especially important for foreigners who come to study in Northern Cyprus, which has a multicultural and multilingual structure, to learn Turkish language not only for their academic life but also for them to continue their daily lives without any problems. In fact, the number of international students studying in the universities of Northern Cyprus, which is located in the northeast of an island in the Mediterranean Sea, is even more than the number of domestic students. Statistical data announced by the Ministry of National Education of Northern Cyprus for the 2020–2021 academic year indicated that a total of 103,108 university students have been studying in Northern Cyprus and only 13% (13,427) of these students were Cypriots. Students coming from Turkey constitute the largest group of international students, followed by students coming from the African continent (Ministry of National Education Culture of North Cyprus, 2021). In a study by Osmanli (2018), it was noted that international students make up almost half of the population in some of the cities in Northern Cyprus. For example, in 2021, the population of local residents of Nicosia, the capital of Northern Cyprus, was 61,376, and the number of international students was 41,416. Thus, international students accounted for ∼40% of the city's population. Therefore, as also stated in several studies available in the literature (Gülmez, 2018; Yücel, 2018), the universities in Northern Cyprus have both a multilingual and multicultural structure. The aim of this study is to assess how well university students in Northern Cyprus who have a sizable international student population are learning Turkish as a foreign language.

"Self-perceived intelligence type" is the first factor taken into account when conducting research on this subject. Breakspear (2013, p. 692) explains the new definition of the intelligence in his article as *"Intelligence is a corporate capability to forecast change in time to do something about it. The capability involves foresight and insight, and is intended to identify impending change which may be positive, representing opportunity, or negative, representing threat."* In this study, self-perceived intelligence areas are the types of intelligence defined by Gardner's Multiple Intelligence Theory. Gardner proposed seven different intelligence dimensions in his book "Frames of Mind" published in 1983. Later, in his work titled "Intelligence Reframed" published in 1999, he added the new intelligence dimension and created eight different intelligence dimensions. These are; Verbal-linguistic, Logical-mathematical, Visual-spatial, Musical-rhythmic, Bodily-kinesthetic, Interpersonal, Intrapersonal, and Naturalistic. Karadag and Baştug (2018) reveal that despite the increasing rate of mental assessment, intelligence is still not evaluated more intelligently in Turkey. One of the reasons for this is the problems related to the training and competence of psychologists who apply the intelligence tests. It is observed that families encourage their children to take intelligence tests not for a clinical purpose but because of their personal curiosity. It is also noted that the principle of being beneficial or not harming is not implemented much in Turkey. For example, based on the intelligence test results, it is decided whether a student will receive inclusive education or not. A wrong decision can lead to an education that is not suitable for the level of the student. Furthermore, it is revealed that ethical principles such as responsibility, respect for human rights and non-discrimination are sometimes not taken into consideration. As an alternative to these problems, as Salman et al. (2017) stated, psychologists and educators suggest that Gardner's theory of intelligence can be used in education. The experts state that the theory of multiple intelligences is objective, the level of the students is not graded and the students are not labeled as sufficient or insufficient. Valuing intelligence types other than mathematical and verbal intelligence types is also considered as an important

feature of multiple intelligences. In multiple intelligence theory, it is stated that the tests do not necessarily have to be administered by experts and this situation creates an ease of application for teachers, parents and psychological counselors. It is highlighted that it is an important acquisition for a person who makes a self-evaluation in the fields of multiple intelligences to be aware of their own abilities and skills.

Gardner's point of view on the concept of intelligence, which has been highly criticized and controversial in the scientific world, has led to mobility in the field of education and training. In 1983, Gardner argued in his book Frames of Mind that there is no single intelligence measured by the well-known IQ test, otherwise termed "g." Gardner added that IQ is not exclusively assessed by standardized testing. The many intelligences theory can be used to help someone choose the best learning method for them. According to some psychologists, the multiple intelligence theory is not acknowledged as a valid theory and is not considered to be a universal instrument for explaining human cognitive skills (Waterhouse, 2010; Sternberg, 2015). However, it is very important for a person to be aware of their own abilities and to be able to know themselves. For this reason, Gardner's theory has been especially selected and it aims to contribute to the literature with the findings revealed in the research.

The Theory of Multiple Intelligences asserts that intelligence is multifaceted and aims to improve the existing abilities and potentials of individuals. Traditional educational approaches which are based only on verbal-linguistic and logical-mathematical intelligence fields, have been eliminated increasing the diversity in education. According to the Theory of Multiple Intelligences, students who are successful in intelligence areas other than verbal-linguistic and logical-mathematical intelligence areas can also be described as successful or intelligent. Seven different intelligence types were defined in Gardner's multiple intelligence theory (Gardner, 1983). Gardner later defined another intelligence type, making a total of eight: verbal–linguistic, logical–mathematical, visual–spatial, musical–rhythmic, bodily–kinesthetic, interpersonal, intrapersonal, and naturalistic (Gardner, 1999). According to Gardner, cognitive abilities are independent of each other; therefore, there can be different intelligence types depending on the cognitive domains. Studies on intelligence types have also been undertaken in the field of education. According to the multiple intelligence theory, students have different intelligence types and their learning is affected by the dominant intelligence type they have (Zebari et al., 2018; Lei et al., 2021; Gandasari et al., 2022; Wreede, 2022). The theory of multiple intelligences asserts that intelligence is multifaceted. Traditional educational approaches which are based only on verbal–linguistic and logical–mathematical intelligence fields, have been eliminated, increasing the diversity in education. According to the theory of multiple intelligences, students who are successful in intelligence types other than verbal–linguistic and logical–mathematical intelligence types can also be described as successful or intelligent (Keskin, 2019).

Gardner's theory has also been discussed in the literature in the context of teaching Turkish as a foreign language. In one of these studies, Keskin (2019) reviewed the course material used in teaching Turkish as a foreign language (Yedi Iklim Turkish Teaching Set) in terms of the multiple intelligence theory and determined that the Yedi Iklim Turkish Teaching Set utilized the verbal–linguistic intelligence area the most; however, it did not equally address the remaining seven intelligence types. In another study, Çökmez (2017) determined that Turkish language teaching materials addressed verbal–linguistic and logical–mathematical intelligence areas at a rate of 59.1 and 35.2%, respectively. Creating different activities to develop intelligence types that are little used or not used at all has been suggested. As stated in the literature, it is emphasized that there are some problems due to the use of multiple intelligence theory in teaching Turkish as a foreign language. For this reason, Gardner's theory has been specially selected, and this study aims to contribute to the literature with the findings revealed in the research.

## 2. Conceptual framework

Intelligence in foreign language education is associated with the cognitive dimension when evaluated theoretically. Güneş (2011) suggests that, besides the cognitive dimension, behavioral, and constructivist theories are important in foreign language education. Based on the behavioral language education theory, language-learning technologies have been discussed and researched in the literature. In this study, language learning technologies were defined using Hayta's (2014) study. Accordingly, language learning technologies are computers, internet, media, and mobiles technologies. The tools used in these learning technologies are; movies, short videos, online dictionaries, songs, grammer/exercise websites, podcasts, audio books, short stories and novels on computers, journals and newspaper on the internet, social communication networks (Skype, facebook, twitter, whatsApp, video calling etc.), translation facilities on the internet (Google translations). On the basis of the constructivist language education theory, it can be seen that one of the current issues, that of twenty-first-century skills, has drawn attention. In this study, twenty-first-century skills are defined on the basis of Eker's (2020) study. Ac-cording to this; Communication and Collaboration (Communicate Clearly, Collaborate with Others, Think Interdependently), Creativity and Innovation (Think Creatively, Work Creatively with Others, Apply Past Knowledge to New Situations), Critical Thinking and Problem Solving (Think Critically, Make Judgments and Decisions, Ask Questions, Solve Problems), Reflection and Awareness, (Metacognition/Thinking about Our Thinking, Reflect and Synthesize). Thus, language-learning technologies and twenty-first-century skills are investigated in addition to the intelligence type variable in this research. These two variables are important in teaching Turkish as a foreign language, as pointed out in the literature.

Kalemkuş and Özek (2021) conducted content analysis on 115 studies on twenty-first-century skills carried out between 2000 and 2020 and found that the Turkish language curriculum, Turkish teachers, Turkish teacher candidates and Turkish-language textbooks were evaluated based on twenty-first-century skills, but teaching Turkish as a foreign language was not. Dündar and Polat (2021) investigated teaching Turkish as a foreign language within the scope of twenty-first-century skills of the curriculum,

including the acquisitions of listening, spoken interaction, spoken production, reading, and writing skills from A1 to C1 levels, and concluded that twenty-first-century skills, which are key for students developing their competencies, were not sufficiently included in the curriculum. The other issue is about learning technologies in foreign language teaching. Since the 2000s, the number of scientific studies on teaching Turkish as a foreign language has increased. However, the studies that address the technological aspect of the matter are still not of the desired quality and quantity (Güntaş et al., 2021). It is important to follow current and technological developments in education in order to increase the quality of language teaching and the active participation of students.

## 2.1. Learning technologies in foreign language education

Foreign language learning is the most suitable field of education for use of information technologies (Ahmadi, 2018). Information technologies aid students in a variety of ways. First, information technologies allow a smooth transition from the traditional model of teacher-centered learning to learner-centered learning. In this way, the individual differences between learners can be addressed and their motivation can be increased as a result.

Using a variety of resources, such as short films, online dictionaries, songs, websites with grammar exercises, podcasts, audiobooks, short stories and novels, journals and newspapers, and social media platforms, it is possible to teach and learn foreign languages successfully (Hayta, 2014). Interactive digital materials (z-books, digital games, speech bots, web 2.0 tools, etc.) and the digital platforms where these materials are presented offer enriched solutions to improve reading, listening, writing, and speaking skills, regardless of the method used for teaching/learning a foreign language. In short, as in all other areas, teaching/learning a foreign language is not outside the scope of technology and media.

Parallel to this, there have been further studies on the use of technology in teaching Turkish as a foreign language as well as in global foreign language education. In this context, integration of technology with language learning–teaching (Birinci, 2020; Repetto et al., 2021; Van Lieshout and Cardoso, 2022), the use of web 2.0 tools, social media, blogs and extracurricular learning environments in language teaching (Bozavli, 2017; Taylan, 2018; Ustabulut and Keskin, 2020; Inal and Arslanbaş, 2021; Sarigül, 2021), digital stories (Akdag and Altinay, 2021; Çokyaman and Çelebi, 2021; Kazazoglu and Bilir, 2021), e-portfolios (Erice and Ertaş, 2011), virtual classrooms (Parmaxi, 2020), and robot teachers (Edwards and Cheok, 2018) have been addressed in the literature. Nevertheless, the results of these studies on the competencies of both teachers and students regarding the use of technology in the language learning–teaching process are contradictory. The discrepancies between these studies may be attributed to the differences between the characteristics of the respective samples since it is known that some individuals easily adapt to the use of technology in the language-learning process, while others show resistance. Indeed, investigating the reasons for these differences between individuals in adapting to the use of technology in the language-learning process and raising awareness about the use of learning technologies in foreign language education were the primary motivational factors for this study.

Many researchers in the field of language education support an open transition to technology-enhanced, student-centered instruction that improves language proficiency (Amini and Amini, 2017; Hong et al., 2017). In addition, these researchers promote the use of a holistic approach in language education that combines language, literature, and culture (Mohr and Welker, 2017; Morska et al., 2018). It is very important for students to actively participate in foreign language classes regardless of the grade level. However, challenging curricular content, contextually inappropriate learning tasks and teaching approaches that fail to involve students as active participants in their learning are reasons why students' active participation cannot be achieved at the levels desired (Philp and Duchesne, 2016; Park and Hiver, 2017). Therefore, language education in general and foreign language education in particular should not focus solely on specific contents, themes and concepts. Rather, language education should prepare students for rapidly changing economic, political, and social conditions and develop their twenty-first-century skills in the ever-changing realities of a globalized society (Moeller and Abbott, 2018; Quicios, 2018). Yeni's (2018) found that twenty-first-century skills increased the educational technology and material development competencies of foreign language teachers.

## 2.2. Twenty-first-century skills

The classification of twenty-first-century skills was made within the framework of the Partnership for twenty-first-century skills (P21). Accordingly, skills have been placed in several categories. The first is "Life and Career Skills," which includes flexibility, communication and cooperation skills. These skills focus on critical thinking and adaptability, entrepreneurship and self-management, social and intercultural skills, productivity and accountability, leader-ship, and responsibility. The second category is "Learning and Innovation Skills," which focuses on several dimensions, i.e., critical thinking, communication, collaboration and creativity dimensions. In the critical thinking dimension, the focus is on creativity and innovation, critical thinking and problem-solving, the ability to analyse complex problems, investigate unclarified matters and evaluate different perspectives or sources of in-formation, and arrive at appropriate conclusions based on evidence and reason (Ravitz et al., 2012; Toharudin, 2017; Tuzlukova and Prabhukanth, 2018). In the communication dimension, the focus is on listening skills as well as being able to communicate effectively using various oral, written, and digital tools in the communication dimension (Fullan, 2013). In the collaboration dimension, the focus is on working respectfully and effectively as a team to generate, use and share knowledge and innovating by providing solutions (Trilling and Fadel, 2012). Lastly, in the creativity dimension, the focus is on creative thinking skills in the context of the production of knowledge, including different ideas for social progress. Creativity is emphasized in all classifications of twenty-first-century skills. The third category is "Information, Media and Technology Skills," which focuses on information literacy,

media literacy, and information and communication technology (ICT) skills.

Language proficiency has been closely linked to communication in the modern world, which is the most fundamental building block for learning new knowledge and bringing about change. Accessing information and making use of the obtained information by analyzing it accurately, comprehension and expression skills, and language use are among the most basic components of twenty-first-century skills. The twenty-first-century skills also include reading, writing, interpretation, and synthesis skills (Ananiadou and Claro, 2009; Trilling and Fadel, 2012; Geisinger, 2016).

One of the core aspects of twenty-first-century skills is "language." Language skills are important for all dimensions of twenty-first-century skills. The importance of language skills in the context of twenty-first-century skills was highlighted by the Modern Language Association (MLA). The re-port published by the MLA in 2007 suggested combining language teaching programmes with twenty-first-century skills. The MLA has highlighted the need to prepare a curriculum that will enable the students learning a second language to effectively communicate with native speakers through the effective use of the second language in question. In addition, MLA envisaged the development of a perspective that would enable students to under-stand the world in terms of another language (Geisler et al., 2007). The studies that ad-dressed the current situation in light of the MLA's report 10 years later (Lomicka and Lord, 2018; Cox and Montgomery, 2019) concluded that the curricular changes needed to support the development of twenty-first-century skills were not sufficiently implemented in most of the currently available language programmes, and a significant large-scale reform has yet to be achieved.

Twenty-first-century skills enable the individual in learning what is needed to be competent and qualified in the most efficient way (Louis, 2012; Hamarat, 2019). Learning twenty-first-century skills is not limited to educational environments. As a matter of fact, twenty-first-century skills can be more effectively acquired within the scope of lifelong learning. Individuals with twenty-first-century skills are expected to be productive, efficient, responsible, entrepreneurial, and social individuals with leadership qualifications who can think, communicate, analyze, and synthesize critically and creatively (Kurudayioglu and Taşkin, 2019).

One of the studies investigating the relationship between multiple intelligence types and twenty-first-century skills is by Ipekşen (2019). They found that multiple intelligence types predicted twenty-first-century skills. In addition to this result, it has been revealed that twenty-first-century skills of students can be developed with activities based on multiple intelligences. In studies investigating the relationship between multiple intelligences and twenty-first-century skills, emotional intelligence is addressed in particular. For example, many studies examining the effect of emotional intelligence on problem-solving skills argue that intelligence and problem-solving skills are related (Kim and Han, 2015; Aslan, 2019; Ndawo, 2021). Other intelligence types are also effective in the problem-solving skills of individuals. Intelligence types may change according to social, environmental and economic conditions, and this may affect people's problem-solving skills (Çinkiliç and Soyer, 2013). It is known that intelligence also has a positive effect on cooperation and leadership skills (Zhang et al., 2018). Similarly, in many studies investigating the relationship between creativity and intelligence, a highly significant relationship was found between intelligence types and creativity ability (Xu et al., 2019; Plucker et al., 2020; Frith et al., 2021). In another recent study (Uçar, 2021), the role of intelligence and creativity in the entrepreneurial tendencies of the Z generation was examined. According to the findings of the research, the creativity levels of the Z generation predict their entrepreneurial tendencies positively and significantly. In all the studies mentioned, the positive relationship between intelligence and twenty-first-century skills are emphasized.

To summarize, the concepts of twenty-first-century skills, intelligence types and learning technologies in the context of foreign language learning process were emphasized in this study. The primary goal of this study was to establish how well the concept of twenty-first-century abilities, as was discussed above, might predict the use of foreign language learning technology by higher education students. The second objective of this study was determined as to evaluate whether the perceived intelligence type differentiates higher education students' use of foreign language learning technologies.

## 2.3. Three variables of the research in the context of teaching Turkish as a foreign language

The International Society for Technology in Education (2017) Report drew attention to the relationship between intelligence types, twenty-first-century skills, and the use of technology for learning purposes. The ISTE 2017 report revealed that logical–mathematical intelligence is related to innovative skills, which are directly linked to technology. Innovation is one of the twenty-first-century skills. It is accepted that students who develop and improve their innovative skills can easily adapt to technology and construct knowledge. While adapting to technology, students can produce original ideas, analyse and evaluate their thoughts, and try different ways to solve the problems they encounter (Anagün et al., 2016). Innovative and applicable technological methods were proposed for use by teachers teaching Turkish as a foreign language with a view to making teaching Turkish more interesting, effective and enjoyable (Özkan et al., 2017). In another study, specific course activities were prepared in order to incorporate technology-based materials into teaching Turkish as a foreign language (Ural, 2016). It has been observed in all these studies that the use of low-cost technological materials increased the active participation of the students in the classes and in their willingness to learn. In other related study, Güler and Kalin Sali (2021) determined that the use of Edmodo positively affected university students' learning of Turkish as a foreign language. In short, as stated in studies by Liu and Xin (2018) and Zhao and Tianyuan (2019), it is self-evident that foreign language education must be supported with learning technologies. Mettursun (2018) addressed all types of intelligence in the context of teaching Turkish to foreigners and determined that taking multiple intelligence theory into account in teaching the

language to foreigners positively affected students' Turkish learning and facilitated their acquisition of Turkish skills. In parallel, Tilbe (2006) determined that the students who learned Turkish from course materials prepared based on the multiple intelligence theory (the experimental group) were more successful than other students (the control group). Similarly, in a study where the effects of teaching practices based on multiple intelligence theory on Turkish reading comprehension skills were investigated, Epçaçan (2013) found that teaching Turkish language using different applications based on intelligence types was very effective in improving students' reading comprehension skills.

Eubanks et al. (2018) investigated whether the technology-integrated twenty-first-century writing workshop was effective for students' writing skills and attitudes, and determined as a result that their writing barriers decreased as they used technology within the scope of the technology-integrated twenty-first-century writing workshop. Bican (2021) discussed the opportunities offered by digital learning environments for writing skills in the context of teaching Turkish to foreigners and found that digital environments have contributed to students' writing skills in and outside the classroom. This finding was attributed to the fact that students were able to utilize their writing skills and receive feedback in virtual learning environments. Digital learning environments were also stated to increase students' problem-solving, critical thinking and creativity skills (Yilmaz et al., 2020, 2022; Atasoy, 2021). In a study by Güngör (2021), "Teaching Turkish as a Foreign Language in the Context of Twenty-First-Century Skills," learning and innovative skills in addition to communication, cooperation, creativity and critical thinking skills were analyzed, and it was determined that Turkish lessons do not reflect contemporary approaches. Yilmaz and Babacan (2015) investigated podcast applications aimed at improving listening skills in teaching Turkish as a foreign language and found that they enriched students' listening skills and the process of teaching Turkish as a foreign language. In addition, in a study conducted with a view to increasing the listening comprehension success of students learning Turkish as a foreign language and reducing their listening anxiety, Berk and Açik (2021) concluded that e-audience-based activities increased the success of listening comprehension.

## 3. Objective and research questions

When the studies examining the relationship between research variables were examined, it was noticed that some subjects were not investigated. It is seen that the studies examined the relationship between intelligence types and twenty-first-century skills focus on emotional intelligence. It is worth investigating the nature of the relationship between intelligence types (other than emotional intelligence) and twenty-first-century skills. It is emphasized that a person's usage of technology may have anything to do with their family, their education, or even themselves. The link between technology and individual competence is the main topic of this study. It aims to reveal which intelligence type most affects the use of technology in foreign language education.

Information, media, and technology skills are one aspect of twenty-first-century abilities. It is anticipated that students who are highly motivated and skilled in this subject would use technology extensively. This study focused on the learning and innovation skills category of twenty-first-century skills, whereas the other two categories, namely, life and career skills and information, media and technology skills, were deliberately excluded from the scope of the research. This is because, based on the results of a vast number of studies available in the literature (Ertmer and Ottenbreit-Leftwich, 2010; Young, 2012; Chang and Chen, 2015; Garba et al., 2015; Koh et al., 2015; Eubanks et al., 2018; LaForce, 2018), it is expected that the use of technology in foreign language learning, which is the dependent variable primarily investigated in this study, would be related to information, media, and technology skills. In this study, the focus is on learning and innovation skills, which is another twenty-first-century skill area. Does having learning and innovation skills affect the use of technology in foreign language education? This study sought to answer that question. Twenty-first-century skills in the context of the education, learning and teaching process do not only imply technology competence or technology use. The original aspect of this study that distinguishes it from many other relevant studies available in the literature is that it focuses on learning and innovation skills rather than information, media and technology skills, which were already addressed numerous times in the context of technology use for foreign language learning. The research questions prepared based on this objective are as follows.

i. Is there a difference between self-perceived dominant intelligence types in terms of using learning technologies in Turkish language learning?

ii. Is there a difference between self-perceived dominant intelligence types in terms of students' twenty-first-century skills scores?

iii. Is there a correlation between twenty-first-century skills scale scores and Turkish language learning technologies scale scores?

## 4. Material and methods

In line with quantitative research design, this study utilized a descriptive approach. The basic feature of descriptive research is to study the current situation, in its own conditions and as it is. In this type of research, researchers are observers; they do not interfere or make any changes. The selection of the sample, the quality of the data collection tools and the accuracy of the data analysis are especially important in quantitative descriptive research (Bacon-Shone, 2013). This study was created using the correlational research model in accordance with the quantitative research technique. According to Creswell (2002), correlational design can be used to predict and explain the relationship between variables. Two or more variables are related and how they affect each other is found out in correlational design. The dependent variables of the research were as follows: "Foreign Language Learning Technologies Scale score" and "Twenty-First-Century Skills Scale score." The relationships, if any, between the total Twenty-First-Century Skills Scale scores, scores obtained from the subscales of Twenty-First-Century Skills Scale and the Foreign Language Learning Technologies Scale scores were investigated by correlation and regression analysis. On the other hand, the independent variable of the research was "perceived intelligence type." Accordingly, it was investigated whether the foreign language

learning technologies scale scores and twenty-first-century skills scale scores of university students, one of the dependent variables, differed according to the perceived intelligence type.

## 4.1. Population and sample

Purposeful and convenience sampling methods were used to create the study sample. Accordingly, the university with the highest number of international students in Northern Cyprus, where the Turkish language is the medium of instruction and one of the authors of this study has been working, was selected. In this way, the research data could be easily accessed and the data collection phase could be completed in a fast and economical manner. The study sample consisted of 431 university students who took Turkish as a foreign language in the 2021–2022 academic year at the selected university.

The sociodemographic data of the students included in the study sample are shown in Table 1. 66.8 percent of the students participating in this research are female and 33.2 percent are male. The majority of the students participating in the research are between the ages of 18–21 (54.8%). Most of the students are from the Middle East (32.3%) and African (64.7%) countries. The departments where the students study are dentistry (13.7%), medicine (16.2%), nursing and health sciences (55.7%), and physiotherapy and nutrition/dietetics (14,4%).

As can be seen in Table 1, the university students have been studying at different faculties and have different Turkish proficiency levels. In this way, maximum diversity could be achieved in the study sample. The Turkish proficiency levels shown in Table 1 have been determined in accordance with the Common European Framework of Reference for Languages (CEFR). Accordingly, students were divided into A1, A2, and B1 levels based on officially announced final grades and the results of the Turkish language proficiency exam carried out in the university where this study was conducted. Each language level consisted of two stages. Students who are successful in the exams pass to the next level and continue to learn the language by increasing their level.

TABLE 1 Sociodemographic characteristics of the university students who participated in the study.

| | Number (*n*) | Percentage (%) |
|---|---|---|
| **Gender** | | |
| Female | 288 | 66.8 |
| Male | 143 | 33.2 |
| **Age** | | |
| 18–21 | 236 | 54.8 |
| 22–25 | 140 | 32.5 |
| 26 and older | 55 | 12.8 |
| **Turkish proficiency levels*** | | |
| A1.1 | 143 | 33.2 |
| A1.2 | 192 | 44.5 |
| A2.1 | 40 | 9.3 |
| A2.2 | 49 | 11.4 |
| B1.1 | 7 | 1.6 |
| **Which of the following grade ranges does the final grade you received from the Turkish course fall into?** | | |
| 0–50 | 129 | 29.9 |
| 51–60 | 72 | 16.7 |
| 61–70 | 86 | 20.0 |
| 71–80 | 49 | 11.4 |
| 81–90 | 53 | 12.3 |
| 91–100 | 42 | 9.7 |
| **The geographical region of origin** | | |
| Middle East | 139 | 32.3 |
| Africa | 279 | 64.7 |
| Other | 13 | 3.0 |
| **Have you ever resided in a Turkish-speaking country before?** | | |
| Yes | 108 | 25.1 |
| No | 323 | 74.9 |
| **Department** | | |
| Dentistry | 59 | 13.7 |
| Medicine | 70 | 16.2 |
| Nursing/health science/first and emergency aid | 240 | 55.7 |
| Nutrition and dietetics/physiotherapy & rehabilitation | 62 | 14.4 |

*The Turkish proficiency levels have been determined in accordance with the Common European Framework of Reference for Languages (CEFR).

## 4.2. Data collection tools

The scales of twenty-first-century skills and the foreign language-learning technologies were used to collect the research data. In addition to these scales, a personal information form was used to define the sample and perform the related statistical analyses.

Personal Information Form: This included questions about students' gender, age, nationality, Turkish proficiency level, the faculty and department in which they were enrolled, whether they lived in a Turkish-speaking country before, the final grade they received for the Turkish course, how often they used computers while learning Turkish and their weekly Internet usage time. The form consisted mostly of multiple-choice questions.

Self-Perceived Intelligence Types: In the personal information form, there is a question about the most important independent variable of the study, which is "self-perceived intelligence type." Saban (2010), one of the researchers who has been working on multiple intelligence theory, mentioned many techniques, i.e., observation, anecdote recording and student self-assessment, that can be used to determine intelligence types in addition to scales. Saban asked the students about their perceptions of their intelligence type and to provide information based on

their self-awareness. A self-assessment question in which students evaluate their own intelligence types and find the most dominant intelligence types is included in the personal information form. The questionnaire is taken from Selçuk et al. (2004). In this questionnaire, university students read 32 statements about 8 intelligence types and assign 0, 1, 2, 3, or 4 points to each statement. A high score for the statements means that it is suitable for the respondent, and a low score indicates that it is not appropriate. Then, respondents write their scores in the table and find the total score for each intelligence type. If the score is equal, they read the statements again and score. In the end, respondents find a single intelligence type that is dominant for them. In this study, the self-perceived intelligence type is a categorical variable and will be used as an independent variable. The dominant intelligence type of the respondents will be deter-mined and the analysis will be carried out with it. Intelligence types other than the dominant intelligence type of the respondents will not be used in the analysis.

The Twenty-First-Century Skills Scale: For this, the Survey Questionnaire of the Implementation of 4Cs (Critical Thinking, Communication, Collaboration, Creativity), which was developed by Eker (2020), was used. The scale consists of 40 items. All items were constructed using a positive sentence structure. Answer choices in each item were prepared in accordance with a five-point Likert-type rating. Accordingly, the following answer choices were included in each item: always true of me, usually true of me, some-what true of me, usually not true of me and never true of me. The Turkish validity and reliability studies of the scale were also conducted by Eker. *Validity:* Given that this study focused on learning and that the studies on the relationship between learning and twenty-first-century skills available in the literature employed only the learning and innovation skills-4Cs category of the twenty-first-century skills, only the "Learning and Innovation Skills" category of the twenty-first-century skills scale was considered in this study with reference to Eker's abovementioned work. The twenty-first-century skills scale developed by Eker consists of communication and collaboration (Communicate Clearly, Collaborate with Others, Think Interdependently), creativity and innovation (Think Creatively, Work Creatively with Others, Apply Past Knowledge to New Situations), critical thinking and problem solving (Think Critically, Make Judgements and Decisions, Ask Questions, Solve Problems), and reflection and awareness (Metacognition-Thinking About Our Thinking, Reflect and Synthesize) sub-dimensions. *Reliability:* The Cronbach's Alpha values of the subscales were 0.907 for communication and collaboration, 0.932 for creativity and innovation, 0.898 for critical thinking and problem solving, 0.918 for the reflection and awareness subscales, and 0.970 for the overall twenty-first-century skills scale. In this study, twenty-first-century skills score is a continuous variable and is used as a dependent variable.

Foreign Language-Learning Technologies Scale: The foreign language-learning technologies scale developed by Hayta (2014) was used. All items were constructed using a affirmative sentence structure. Answer choices in each item were prepared in accordance with a five-point Likert-type rating. Accordingly, the following answer choices were included in each item: never, rarely, sometimes, often, and always. The Turkish validity and reliability studies of the scale were also conducted by Hayta. *Validity:* The scale, which has no sub-dimensions, was developed as a single-factor scale consisting of 41 items. The exploratory factor analysis of the scale was repeated for Turkish Cypriots. The variances explained by the factors were reviewed based on the results of the exploratory factor analysis applied by principal component analysis and varimax transformation, and it was found that the foreign language-learning technologies scale had a single-factor structure with an Eigenvalue $>1$. It was observed that the factor load of 41 items on the scale was 0.5 or higher, and thus no item was removed from the scale. It was determined that the scale's single dimension explained 45.60% of the total variance. *Reliability:* The Cronbach's Alpha value of the scale was calculated as 0.984. In this study, foreign language-learning technologies score is a continuous variable and will be used as a dependent variable.

## 4.3. Data collection process

First, the researchers who developed the twenty-first-century skills scale and the foreign language-learning technologies scale, which were intended to be used in this study, were contacted *via* e-mail, and their permission was obtained. The study protocol was submitted to the scientific ethics committee of the university where this study was conducted and the required ethics committee approval was granted. The nine instructors who teach Turkish to international students were informed of the ethics committee's acceptance of the project and provided with the pertinent details. The research questions were constructed into an online scale and each faculty member was asked to share this online scale with their students. The purpose of the study and the consent form were included in the first section of the online scale. Only the students who wanted to participate in the study voluntarily were expected to fill out the online scale. The names of the students were not included in the forms, and both the faculty members and students were informed that the research data would be kept confidential.

## 4.4. Data analysis

SPSS 24.0 (Statistical Package for Social Sciences for Windows, version 24.0, IBM Corp., Armonk, NY, U.S., 2016) software was used in the statistical analyses of the quantitative data collected. The Kolmogorov–Smirnov test, Quantile–Quantile Plots (QQ plots) and Skewness and Kurtosis coefficients were used to determine whether the research data conformed to the normal distribution, and it was determined that the scores obtained from the scales did not conform to the normal distribution. Accordingly, descriptive statistics pertaining to the scores obtained from the scales were expressed using arithmetic mean and standard deviation values and minimum and maximum values.

Since the data set did not show a normal distribution, Spearman's correlation and Kruskal–Wallis $H$-test, which are non-parametric tests, were used. Spearman's correlation analysis was used to determine the relationship between two dependent variables, foreign language-learning technologies scale scores and twenty-first-century skills scale scores. With the Kruskal–Wallis

TABLE 2 Self-perceived intelligence types of university students.

| | Number (*n*) | Percentage (%) |
|---|---|---|
| **In which intelligence area do you consider yourself more competent?** | | |
| Bodily-kinesthetic intelligence | 28 | 6.5 |
| Interpersonal intelligence | 95 | 22.0 |
| Intrapersonal intelligence | 36 | 8.4 |
| Linguistic-verbal intelligence | 50 | 11.6 |
| Logical-mathematical intelligence | 61 | 14.2 |
| Musical intelligence | 65 | 15.1 |
| Naturalistic intelligence | 89 | 20.6 |
| Visual-spatial intelligence | 7 | 1.6 |

TABLE 3 The scores university students obtained from the twenty-first-century kills scale and the foreign language learning technologies scale.

| Twenty-first century skills scale and its subscales | *n* | | s | Min | Max |
|---|---|---|---|---|---|
| Communication and collaboration subscale | 431 | 4.15 | 0.62 | 2.15 | 5.00 |
| Creativity and innovation subscale | 431 | 4.11 | 0.70 | 2.00 | 5.00 |
| Critical thinking and problem solving subscale | 431 | 4.00 | 0.75 | 1.50 | 5.00 |
| Reflection and awareness subscale | 431 | 4.10 | 0.63 | 2.05 | 5.00 |
| Total twenty-first-century skills scale | 431 | 4.06 | 0.72 | 1.00 | 5.00 |
| Foreign Language Learning Technologies Scale | 431 | 2.72 | 0.86 | 1.00 | 5.00 |

*H*-test, the researchers investigated whether there were significant differences between dominant intelligence types in terms of levels of foreign language-learning technologies. The Kruskal–Wallis *H*-test was run twice. Similarly, the Kruskal–Wallis *H*-Test was used to investigate whether there were significant differences between dominant intelligence types in terms of levels of twenty-first-century skills. In each analysis, a Kruskal–Wallis *H*-test was run with one independent and one dependent variable. In cases where there was a significant difference, pairwise comparisons were performed with the Mann–Whitney *U*-test.

## 5. Results

In this section, first, descriptive statistics of the variables searched in the study are presented. Then, the results of the statistical analyses conducted are given in accordance with the order of the research questions presented in the Introduction. As seen in Table 2, the number of university students who stated that they are more competent in interpersonal and naturalistic intelligence areas was the highest, whereas the number of university students who stated that they are more competent in intrapersonal and visual-spatial intelligence areas was the lowest. As seen in Table 3, university students obtained

TABLE 4 Kruskal Wallis *H*-test analysis of the scores university students obtained from the foreign language learning technology scale scores by the self perceived intelligence types.

| | Perceived intelligent types | *N* | Mean rank | | *df* | *P* |
|---|---|---|---|---|---|---|
| Foreign language learning technologies scale | Bodily-kinesthetic intelligence | 28 | 216.57 | 12.534 | 7 | 0.084 |
| | Interpersonal intelligence | 95 | 247.28 | | | |
| | Intrapersonal intelligence | 36 | 229.93 | | | |
| | Linguistic intelligence | 50 | 228.84 | | | |
| | Logical-mathematical intelligence | 61 | 192.57 | | | |
| | Musical intelligence | 65 | 192.07 | | | |
| | Naturalistic intelligence | 89 | 203.48 | | | |
| | Spatial intelligence | 7 | 211.29 | | | |

higher scores on the twenty-first-century skills scale. As for the scores obtained from the subscales of the twenty-first-century skills scale, it was observed that the scores obtained from the critical thinking and problem-solving subscale were the lowest. On the other hand, the analysis of the scores obtained from the foreign language-learning technologies scale indicated that the students used technology at a moderate level (x = 2.72/5). The first research question aimed to reveal whether the foreign language-learning technologies scale scores differed by self-perceived intelligence type. As can be seen in Table 4, university students' foreign language-learning technology scale scores did not differ significantly by perceived intelligence type.

The second research question examined whether scores on twenty-first-century skills varied according to the type of self-perceived intelligence. As can be seen in Table 5, university students' total twenty-first-century skills scale scores as well as the scores they obtained from the critical thinking and problem solving, and reflection and awareness subscales of the twenty-first-century skills scale differed significantly by the perceived intelligence types. Pairwise comparisons were made with Mann–Whitney *U*-test. The results of these comparisons are given in Tables 6–8. In order for the Mann–Whitney *U* tables not to be too long, only the results with significant differences are included. When Table 6 is examined, according to Mann–Whitney *U* analysis, critical thinking and problem-solving subscales scores of students with self-perceived musical intelligence type were significantly lower than students with all other intelligence types. Similarly, in the critical thinking and problem-solving subscale, the scores of students with self-perceived spatial intelligence type were significantly lower than those of students with many intelligence types (except naturalistic and musical). In Table 7, according to Mann–Whitney *U* analysis, reflection and awareness subscale scores of students with self-perceived musical intelligence type were significantly lower than students with most of the self-perceived intelligence types (except spatial, bodily and kinesthetic). Similarly, in the reflection and awareness subscale, the scores

TABLE 5 Kruskal Wallis *H*-test analysis of the scores university students obtained from the twenty-first-century skills scale by the self perceived intelligence types.

| | Perceived intelligence type | *N* | Mean rank | *df* | $X^2$ | *p* |
|---|---|---|---|---|---|---|
| Communication and collaboration subscale | Bodily-kinesthetic intelligence | 28 | 202.63 | | | |
| | Interpersonal intelligence | 95 | 233.04 | | | |
| | Intrapersonal intelligence | 36 | 220.78 | | | |
| | Linguistic intelligence | 50 | 223.74 | | | |
| | Logical-mathematical intelligence | 61 | 233.87 | 7 | 11,366 | 0.123 |
| | Musical intelligence | 65 | 173.12 | | | |
| | Naturalistic intelligence | 89 | 216.41 | | | |
| | Spatial intelligence | 7 | 195.71 | | | |
| Creativity and innovation subscale | Bodily-kinesthetic intelligence | 28 | 202.07 | | | |
| | Interpersonal intelligence | 95 | 222.09 | | | |
| | Intrapersonal intelligence | 36 | 231.36 | | | |
| | Linguistic Intelligence | 50 | 225.04 | | | |
| | Logical-mathematical intelligence | 61 | 239.02 | 7 | 9,790 | 0.201 |
| | Musical intelligence | 65 | 181.02 | | | |
| | Naturalistic intelligence | 89 | 215.64 | | | |
| | Spatial intelligence | 7 | 174.21 | | | |
| Critical thinking and problem solving subscale | Bodily-kinesthetic intelligence | 28 | 218.38 | | | |
| | Interpersonal intelligence | 95 | 235.26 | | | |
| | Intrapersonal intelligence | 36 | 241.36 | | | |
| | Linguistic intelligence | 50 | 235.35 | | | |
| | Logical-mathematical intelligence | 61 | 227.19 | 7 | 18,318 | 0.011* |
| | Musical intelligence | 65 | 166.07 | | | |
| | Naturalistic intelligence | 89 | 210.31 | | | |
| | Spatial intelligence | 7 | 115.00 | | | |
| Reflection and awareness subscale | Bodily-kinesthetic intelligence | 28 | 196.48 | | | |
| | Interpersonal intelligence | 95 | 231.36 | | | |
| | Intrapersonal intelligence | 36 | 231.19 | | | |
| | Linguistic intelligence | 50 | 222.29 | | | |
| | Logical-mathematical intelligence | 61 | 241.84 | 7 | 15,601 | 0.029* |
| | Musical intelligence | 65 | 170.03 | | | |
| | Naturalistic intelligence | 89 | 219.33 | | | |
| | Spatial intelligence | 7 | 121.93 | | | |
| Twenty-first century skills scale | Bodily-kinesthetic intelligence | 28 | 203.52 | | | |
| | Interpersonal intelligence | 95 | 231.48 | | | |
| | Intrapersonal intelligence | 36 | 233.07 | | | |
| | Linguistic intelligence | 50 | 227.26 | | | |
| | Logical-mathematical intelligence | 61 | 238.57 | 7 | 18,124 | 0.011* |
| | Musical intelligence | 65 | 169.08 | | | |
| | Naturalistic intelligence | 89 | 214.88 | | | |
| | Spatial intelligence | 7 | 141.00 | | | |

*$p < 0.05$.

TABLE 6 Pairwise comparisons of the critical thinking and problem solving subscale scores of university students obtained from the twenty-first-century skills scale by the self-perceived intelligence types.

| Domain binary | *N* | Mean rank | Sum of ranks | *U* | *p* |
|---|---|---|---|---|---|
| Bodily kinesthetic | 28 | 55.55 | 1,555.50 | 670.5 | 0.044* |
| Musical | 65 | 43.32 | 2,815.50 | | |
| Bodily kinesthetic | 28 | 19.82 | 555.0 | 47.000 | 0.034* |
| Spatial | 7 | 10.71 | 75.0 | | |
| Interpersonal | 95 | 90.56 | 8,603.5 | 2,131.5 | 0.001* |
| Musical | 65 | 65.79 | 4,276.5 | | |
| Interpersonal | 95 | 53.37 | 5,070.5 | 154.5 | 0.018* |
| Spatial | 7 | 26.07 | 182.5 | | |
| Intrapersonal | 36 | 62.25 | 2,241.0 | 765.0 | 0.004* |
| Musical | 65 | 44.77 | 2,910.0 | | |
| Intrapersonal | 36 | 24.04 | 865.50 | 52.500 | 0.015* |
| Spatial | 7 | 11.50 | 80.50 | | |
| Linguistic | 50 | 68.31 | 3,415.5 | 1,109.5 | 0.004* |
| Musical | 65 | 50.07 | 3,254.5 | | |
| Linguistic | 50 | 30.96 | 1,548.0 | 77.000 | 0.017* |
| Spatial | 7 | 15.0 | 105.0 | | |
| Logical-mathematical | 61 | 72.81 | 4,441.5 | 1,414.5 | 0.005* |
| Musical | 65 | 54.76 | 3,559.5 | | |
| Logical-mathematical | 61 | 36.45 | 2,223.5 | 94.500 | 0.016 |
| Spatial | 7 | 17.50 | 122.5 | | |
| Musical | 65 | 68.05 | 4,423.0 | 2,278.0 | 0.024* |
| Naturalistic | 89 | 84.40 | 7,512.0 | | |

*$p < 0.05$.

of students with self-perceived spatial intelligence type were significantly lower than students with many intelligence types (except naturalistic, musical, and bodily-kinesthetic). In Table 8, the Mann–Whitney *U*-test was performed on the basis of the twenty-first-century scale total scores and significant differences were observed in self-perceived musical and spatial intelligence scores. The twenty-first-century total scores of the students with the self-perceived musical intelligence type were significantly lower than the scores of the students with the other five intelligence types (interpersonal, intrapersonal, linguistic, logical mathematical, naturalistic). In the self-perceived spatial intelligence type, there was a significant difference in a single intelligence type. The total scores of twenty-first-century skills of the students with the self-perceived logical mathematical intelligence type were significantly higher than the students with the self-perceived visual intelligence type.

The third research question aimed to reveal the correlations between the scores obtained from the total twenty-first-century skills scale, from the subscales of the twenty-first-century skills scale and from the foreign language-learning technologies scale. As seen in Table 9, there was a weak yet statistically significant correlation between the total twenty-first-century skills scale scores and the foreign language-learning technologies scale scores in the positive direction. For a correlation coefficient to be interpreted, the $p < 0.05$. In this study, $r < 0.2$ was found and there was a very weak correlation (Akoglu, 2018). Accordingly, as students' twenty-first-century skills scale scores increased, their foreign language-learning technologies scale scores also increased. No statistically significant correlation was found between the scores obtained from the subscales of the twenty-first-century skills scale and the foreign language-learning technologies scale scores.

## 6. Discussion

This section is structured in two parts. The first section evaluates and discusses the issue of self-perceived intelligence kinds in teaching Turkish as a foreign language in light of the study's findings. In this section, both the subjects of the self-perceived intelligence type and twenty-first-century skills, and the self-perceived intelligence type and foreign language-learning technologies results are discussed in the context of teaching Turkish as a foreign language. The second part is introduced below.

TABLE 7 Pairwise comparisons of the reflection and awareness subscale scores of university students obtained from the twenty-first-century skills scale by the self perceived intelligence types.

| Domain binary | *N* | Mean rank | Sum of ranks | *U* | *p* |
|---|---|---|---|---|---|
| Interpersonal | 95 | 89.44 | 8,496.5 | 2,238.5 | 0.003* |
| Musical | 65 | 67.44 | 4,383.5 | | |
| Interpersonal | 95 | 53.18 | 5,052.5 | 172.5 | 0.033* |
| Spatial | 7 | 28.64 | 200.5 | | |
| Intrapersonal | 36 | 60.64 | 2,183.0 | 823.0 | 0.013* |
| Musical | 65 | 45.66 | 2,968.0 | | |
| Intrapersonal | 36 | 24.07 | 866.5 | 51.5 | 0.014* |
| Spatial | 7 | 11.36 | 79.5 | | |
| Linguistic | 50 | 66.24 | 3,312.0 | 1,213.0 | 0.019* |
| Musical | 65 | 51.66 | 3,358.0 | | |
| Linguistic | 50 | 30.75 | 1,537.5 | 87.500 | 0.032* |
| Spatial | 7 | 16.5 | 115.5 | | |
| Logical-mathematical | 61 | 74.15 | 4,523.0 | 1,333.0 | 0.001* |
| Musical | 65 | 53.51 | 3,478.0 | | |
| Logical-mathematical | 61 | 36.48 | 2,225.5 | 92.500 | 0.014* |
| Spatial | 7 | 17.21 | 120.5 | | |
| Musical | 65 | 67.62 | 4,395.0 | 2,250.0 | 0.018* |
| Naturalistic | 89 | 84.72 | 7,540.0 | | |

*$p < 0.05$.

TABLE 8 Pairwise comparisons of the twenty-first-century skills scale scores of university students obtained from the twenty-first-century skills scale by the self perceived intelligence types.

| Domain binary | *N* | Mean rank | Sum of ranks | *U* | *p* |
|---|---|---|---|---|---|
| Interpersonal | 95 | 89.66 | 8,518.0 | 2,217.0 | 0.002* |
| Musical | 65 | 67.11 | 4,362.0 | | |
| Intrapersonal | 36 | 60.72 | 2,186.0 | 820.0 | 0.013* |
| Musical | 65 | 45.62 | 2,965.0 | | |
| Linguistic | 50 | 66.76 | 3,338.0 | 1,187.0 | 0.013* |
| Musical | 65 | 51.26 | 3,332.0 | | |
| Logical-mathematical | 61 | 73.84 | 4,504.0 | 1,352.0 | 0.002* |
| Musical | 65 | 53.80 | 3,497.0 | | |
| Logical-mathematical | 61 | 36.36 | 2,218.0 | 100.0 | 0.022* |
| Spatial | 7 | 18.29 | 128.0 | | |
| Musical | 65 | 67.96 | 4,417.5 | 2,272.5 | 0.023* |
| Naturalistic | 89 | 84.47 | 7,517.5 | | |

*$p < 0.05$.

## 6.1. Discussion related with self-perceived intelligence types

As can be seen in Table 2, the number of university students who stated that they are more competent in self-perceived interpersonal intelligence and naturalistic intelligence was the highest. This result is compatible with the finding that the highest mean score was obtained from the communication and collaboration subscale of the twenty-first-century skills scale ($x = 4.15$) (see Table 3). The finding that university students thought they were more competent in self-perceived interpersonal intelligence, which implied that they are open to social learning, should be taken into account in the foreign language-learning process. Hence, activities involving group work should be incorporated into the foreign language education curriculum. Along these lines, Tekiner (2005) found that

interpersonal intelligence was the most dominant intelligence type in university students learning a foreign language and concluded that interpersonal intelligence is directly related to group learning activities. Similarly, the university students who participated in this study stated that they are more competent in interpersonal intelligence followed by naturalistic intelligence (see Table 2). This is a remarkable finding since it demonstrates the importance of learning experiences outside the classroom in foreign language teaching. Recently, Mousa (2022) revealed the positive effects of out-of-class teaching activities on foreign language learning. Several studies reported the positive effects of extracurricular and social activities on student motivation and language learning in the context of teaching Turkish language to foreigners (Kinay, 2017; Saydam and Çangal, 2018). These findings indicate that learning activities that activate the interpersonal intelligence type positively affect learning Turkish as a foreign language.

In this study, the number of students whose verbal-linguistic intelligence is dominant is less than the students whose interpersonal intelligence, nature intelligence, musical intelligence and mathematical intelligence are dominant (See Table 2). According to the results of the research, the rate of students whose verbal-linguistic intelligence is dominant is ∼12%. However, in the literature, it is emphasized that students with verbal-linguistic intelligence will be more successful in the studies related to the foreign language learning process and the multiple intelligence areas of the students. Özkan (2008) and Trilling and Fadel (2012) determined that language abilities and potential were best expressed by verbal intelligence. Moreover, verbal intelligence predicts the flexibility, communication and cooperation skills included in the "Life and Career Skills" category of twenty-first-century skills, which prompt individuals to come together and share ideas. In parallel, in a study where collaborative tasks that can be applied in the online environment in teaching Turkish as a foreign language were emphasized, Inan (2021) found that collaborative dialogues performed in the target language in order to prompt learners talk to each other, listen to each other and write together helped learners control and support each other's language learning. In this way, the targeted acquisitions in learning the Turkish language were achieved with activities that enable both verbal intelligence and collaborative skills.

The results of this study indicated that interpersonal intelligence significantly affected the problem-solving dimension of twenty-first-century skills (See Tables 5, 6). Similarly, Kiremitçi et al. (2014) found a statistically significant positive relationship between university students' multiple intelligence areas and problem-solving skills. They determined that students with interpersonal intelligence in addition to logical and mathematical intelligence had better problem-solving skills. On the other hand, Kiremitçi et al. (2014) found that people with logical–mathematical intelligence in addition to verbal–linguistic, bodily–kinesthetic and naturalistic intelligence were more successful in solving problems. Students with high perceived levels of interpersonal intelligence were found to also have high critical thinking skills in this study (See Tables 5, 6). In parallel, Sardogan et al. (2006) found that students with high problem-solving skills also have high personal and social adaptation skills. Similarly, Dündar (2009) found a positive relationship between personal adjustment and problem-solving skills. In addition, in a study that investigated the relationship between teachers' multiple intelligence domains and their problem-solving skills, Genç (2012) found a positive correlation between teachers' intrapersonal intelligence and their problem-solving skills.

In this study, the twenty-first-century skill scores of the students in the two intelligence types related to art (musical and spatial) were lower than the students in the other intelligence areas (see Tables 6–8). Based on this result, it is necessary to examine the extent to which the definitions of twenty-first-century skills overlap with artistic development or artistic competencies. Although Erdoğan (2020) states that creativity, which is an important dimension in twenty-first-century skills, is also related to art, she emphasizes that the relationship between twenty-first-century skills and artistic skills needs to be examined in detail.

According to the findings in Table 4, it was revealed that having different intelligence areas did not differentiate the use of technology in learning Turkish as a foreign language. A similar result in this study, that there is no significant difference in the use of learning technologies according to multiple intelligence types, is in line with the findings of Balakrishnan and Gan's (2016) study. Balakrishnan and Gan (2016) investigated the effectiveness of students' learning styles, i.e., intelligence types, on technology use, and found that there are many different factors affecting it. In line with the results of Balakrishan and Lay's study, it was found in this study that the foreign language-learning technologies scale scores of the students did not differ significantly by perceived intelligence type. Studies addressing the theory of multiple intelligences in combination with learning technologies have generally focused on how the theory of multiple intelligences can be integrated into technology-oriented teaching. In these studies, it is emphasized that each student can be more successful in the learning process if enriched and various learning technologies that address all intelligence types are used (Gardner and Veenema, 1996). In parallel, Sahin Timar (2010) determined that the materials and web-based environments prepared in accordance with the theory of multiple intelligences thus addressed the dominant intelligence types of the students and increased students' success, by assisting in students' understanding of the subject, increasing students' interest in lessons, prompting students' active participation in classes and facilitating learning.

## 6.2. Discussion related with twenty-first-century skills and language learning technologies

In the second part of this discussion, the relationship between two dependent variables, except for the self-perceived intelligence type variable, is discussed. The relationship between twenty-first-century skills and language learning technologies in the context of teaching Turkish as a foreign language is examined on the basis of research findings. As can be seen in Table 2 in the Results section, university students scored an average of 4 points in the items of all the subscales of the five point Likert type twenty-first-century skills scale, indicating that they possess the necessary twenty-first-century skills. In parallel, Engin and Korucuk (2021) determined that the twenty-first-century skills of university students were

TABLE 9 Correlations between the total twenty-first-century skills scale scores, scores obtained from the subscales of the twenty-first-century skills scale, and the foreign language learning technologies scale scores.

| Foreign Language Learning Technologies Scale scores | | | |
|---|---|---|---|
| Twenty-first century skills | *r* | *p* | *N* |
| Communication and collaboration subscale | 0.072 | 0.137 | 431 |
| Creativity and innovation subscale | 0.088 | 0.069 | 431 |
| Critical thinking and problem solving subscale | 0.069 | 0.154 | 431 |
| Reflection and awareness subscale | 0.091 | 0.058 | 431 |
| Total twenty-first-century skills scale | 0.114* | 0.018 | 431 |

*$p < 0.05$.

high. Similarly, in several other studies university students were found to possess high twenty-first-century skills (Erdogan, 2018; Kozikoglu and Altunova, 2018). All kinds of learning activities that university students participate in throughout their lives in order to develop their knowledge, skills, interests and competencies, i.e., lifelong learning skills, can form the basis of twenty-first-century skills. The fact that university students were found to possess high twenty-first-century skills was attributed to lifelong learning (Erdogan, 2018; Kozikoglu and Altunova, 2018). In another study, it was emphasized that lifelong learning skills not only positively affect twenty-first-century skills but also increase academic success (Demirel, 2009). In contrast, some studies suggested that academic skills are slightly related or not related at all to twenty-first-century skills. In one of these studies, Göktepe Yildiz (2020) found that students' academic achievement levels were weakly correlated with some twenty-first-century skills including entrepreneurship-innovation, information technology literacy and career awareness, but not with other twenty-first-century skills including critical thinking, problem-solving, social responsibility and leadership. Similarly, as shown in Table 1, almost half of the university students who were learning Turkish as a foreign language stated that the final grade they received from the Turkish course was less than the passing grade (60 out of 100), which indicated that their academic success was low even though they were found to possess high twenty-first-century skills. The discrepancies between the findings of these studies can be attributed to the fact that foreign language learning requires lifelong learning skills, establishing connections between the content taught and daily life, and having learning experiences outside the classroom.

The results of this study revealed that university students have been using technology in foreign language learning at a moderate level (see Table 3, x = 2.72). The fact that the mean foreign language-learning technology scale score was found to be at a medium level despite the mean twenty-first-century skills scale score being high was attributed to students' use of technology for entertainment and killing time and not using it for learning. In parallel, Coşkun et al. (2007) found that even university students studying in academic programmes that require higher academic skills, such as medical education, use technology primarily for entertainment (42%), secondarily to communicate with each other (38%), and only tertiarily for learning, working on projects and homework (30%). The reasons underlying students' lower use of technology for learning are worth investigating, since understanding these reasons may guide educators and policymakers. Özdal et al. (2022) found that students who set learning goals, have the motivation and make the effort to develop learning strategies, and seek help to eliminate all kinds of problems they face during the process were more successful than other students in the online learning process. Thus, they concluded that the development of online self-regulation skills for students is as important as teachers' guidance in the use of technology for learning purposes.

The results of the correlation analyses given in Table 3 revealed that although the learning and innovation skills scores of the students were high, these scores did not relate to technology use for foreign language learning. More specifically, the 4Cs, namely, critical thinking, collaboration, communication and creativity skills, do not significantly affect the use of technology in the foreign language-learning process. This finding demonstrated the necessity of involving social skills in the development of technological skills. The effectiveness of technology-supported learning environments that involve social skills has been brought to the forefront in some studies (Nevgi et al., 2006; Günindi, 2014).

## 6.3. Limitations of the study

One of the study's limitations was that foreign students' Turkish language proficiency was assessed as a whole. Hence, further studies that address the effects of twenty-first-century skills, perceived intelligence types and learning technologies separately for each language skill, i.e., listening, speaking, reading and writing, in the context of teaching and learning Turkish as a foreign language would be useful. Secondly, the fact that only one dimension (learning and innovation skills-4C) of twenty-first-century skills was addressed in this study may be seen as another limitation. Therefore, further studies may address the other two dimensions (life and career skills and information, media and technology skills). Thirdly, the fact that the dominant intelligence types of the university students who participated in this study were determined based on students' own experiences and perceptions may be considered another limitation. The intelligence types of the university students who participated in this study were identified based on their perceptions (or self-assessment) did not allow detailed analysis of the results. The use of technology in language learning can be addressed in detail with a valid and reliable scale that assesses different intelligence types. In this way, it may be possible to further evaluate the relationship between intelligence types and the use of technology in foreign language learning.

# 7. Conclusions and recommendations

In conclusion, the results of the study revealed a weak yet statistically significant correlation between twenty-first-century skills and foreign language-learning technologies usage. Future studies may focus on the relationships between the other categories

of twenty-first-century skills, namely, life and career skills and information, media and technology skills, and usage of foreign language-learning technologies. Additionally, students' scores in twenty-first-century skills differed significantly, whereas their scores for foreign language-learning technology did not, according to their perceived intelligence types. Based on the finding that the type of perceived intelligence makes a difference in twenty-first-century skills but not in language-learning technologies, it is important to in-crease the number of studies on the effectiveness of intelligence in learning a foreign language.

The other twenty-first-century skills that are thought to be related to intelligence in the literature should be researched in light of the fact that twenty-first-century skills and language-learning technologies have a poor relationship based on these research findings. It is important to develop twenty-first-century skills by associating them with all learning processes both in the school environment and outside the school environment, rather than thinking of them as skills to be taught. It can be inferred that the theory of many intelligences is related to talents that are valued today based on the conclusion that different types of intelligence make a difference in twenty-first-century skills. Thus, the theory of multiple intelligences is still up to date. In this respect, it can be said that teaching methods in multiple intelligence theory, textbooks and measurement–evaluation approaches can be developed and used in teaching Turkish as a foreign language. In this study, it was found that perceived intelligence type does not affect the use of learning technologies in learning Turkish as a foreign language. Thus, it can be suggested that research should be conducted based on other variables. For example, whether the environment in which the person lives or their willingness to learn affects the use of foreign language-learning technology can be investigated.

The pedagogical implications of this study can be summarized as follows. The research's findings indicate that students in higher education possess twenty-first-century skills. Based on this finding, it is possible to engage students in the courses and accomplish effective foreign language acquisition if foreign language education is carried out in accordance with modern methodologies and based on twenty-first-century abilities. According to the research findings, the students' use of technology in foreign language education is at a moderate level. Foreign language education courses could be planned for how students can use technology more effectively and more frequently while learning a language. Considering that there is a difference between using technology in daily life and using it for educational purposes, both scientific and applied studies should be carried out especially on technology-supported language education. In the study, it was observed that the type of interpersonal intelligence was high. It has been revealed in this study that it is important to include social learning rather than individual and competitive learning in foreign language education classes. Based on the high level of natural intelligence of the students, the necessity of conducting foreign language education lessons outside the classroom has emerged.

## 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 study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Cyprus International University (EKK21-22/011/0010 and 18.03.2022) for studies involving humans. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

EKK: conceptualization, resources, visualization, methodology, formal analysis, and data curation. EKK and AG: investigation, writing—original draft preparation, and writing—review and editing. All authors contributed to the article and approved the submitted version.

## Conflict of interest

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

## Publisher's note



## References

Adedoyin, O. B., and Soykan, E. (2020). Covid-19 pandemic and online learning: the challenges and opportunities. *Interact. Learn. Environ.* 1–13. doi: 10.1080/10494820.2020.1813180. [Epub ahead of print].

Ahmadi, M. R. (2018). The use of technology in english language learning: a literature review. *Int. J. Res. Eng. Educ.* 3, 115–125. doi: 10.29252/ijree.3.2.115

Akdag, S., and Altinay, Z. (2021). Learning through digital stories for safe school environment. *Front. Psychol.* 12, 738954. doi: 10.3389/fpsyg.2021.738954

Akoglu, H. (2018). User's guide to correlation coefficients. *Turk. J. Emerg. Med.* 18, 91–93. doi: 10.1016/j.tjem.2018.08.001

Amini, D., and Amini, M. (2017). Integrating emotional, motivational, and cognitive engagement to create state of flow in language learning tasks. *J. Lang. Commun.* 4, 297–309.

Anagün, S. S., Atalay, N., Kiliç, Z., and Yaşar, S. (2016). The development of a 21st century skills and competences scale directed at teaching candidates: validity and reliability study. *Pamukkale Univ. J. Educ.* 40, 160–175. doi: 10.9779/puje768

Ananiadou, K., and Claro, M. (2009). 21st century skills and competences for new millennium learners in OECD countries. *OECD Educ. Work. Pap.* 41, 1–34. doi: 10.1787/19939019

Aslan, M. (2019). The relationship between prospective teachers' emotional intelligence and problem-solving approaches. *J. Educ. Kastamonu* 27, 2481–2492. doi: 10.24106/kefdergi.3357

Atasoy, G. (2021). *Material Design Based on 21st Century Skills and Renewed Bloom Taxonomy in Teaching Turkish to Bilinguals*. (PhD Thesis), Istanbul: Yildiz Technical University (Turkey).

Bacon-Shone, J. (2013). *Introduction to Quantitative Research Methods*. Hong Kong: Graduate School, The University of Hong Kong.

Balakrishnan, V., and Gan, C. L. (2016). Students' learning styles and their effects on the use of social media technology for learning. *Telemat. Inform.* 33, 808–821. doi: 10.1016/j.tele.2015.12.004

Berk, R. R., and Açik, F. (2021). The effects of podcasts on listening comprehension and listening anxiety in teaching, Turkish as a foreign language. *Bayburt J. Educ.* 16, 329–348. doi: 10.35675/befdergi.747319

Bican, G. (2021). "Expression skills in the mother tongue in the context of 21st century skills," in *21st Century Skills and Turkish Education*, ed H. E. Çocuk, and T. Sosyal (Ankara: Nobel).

Birinci, F. G. (2020). A review on information technologies used in teaching Turkish as a foreign language. *Sakarya Univ. J. Educ.* 10, 350–371. doi: 10.19126/suje.685534

Bozavli, E. (2017). Learners' perceptions for the development of verbal language skills in foreign language through social media. *J. Hum. Soc. Sci. Res.* 6, 627–640. Available online at: http://www.itobiad.com/tr/pub/issue/27443/292223

Breakspear, A. (2013). A new definition of intelligence. *Intell. Natl. Sec.* 28, 678–693. doi: 10.1080/02684527.2012.699285

Chang, C. P., and Chen, J. (2015). Correlation between pre-service teachers' information technology integration attitude and creative teaching behavior. *Creat. Educ.* 6, 1802–1814. doi: 10.4236/ce.2015.616184

Çinkiliç, I., and Soyer, F. (2013). An investigation the relation between multiple intelligence areas of pre-service physical education teacher and their problem solving skills. *Sports Manage. Inform. Technol.* 8, 1306–4371.

Çökmez, N. (2017). *Analyzing and evaluating the multiple intelligences theory in Turkish as a foreign language coursebooks*. (Master Thesis), Istanbul: Istanbul Universities (Turkey).

Çokyaman, M., and Çelebi, M. (2021). The effect of digital storytelling (dst) on academic achievement in foreign language teaching. *Kirsehir Ahi Evran Univer. J. Educ.* 22, 994–1035. doi: 10.29299/kefad.993465

Coşkun, G., Kaymakoglu, B., and Gök, E. (2007). "The internet usage of medical school students and their attitudes towards e-learning: An application for Baskent University," in *4th National Medical Informatics Congress* (Antalya), 72–77. Available online at: https://turkmia.net/kongre2007/cd/pdf/72-77.pdf (accessed November 26, 2022).

Cox, C. B., and Montgomery, C. (2019). A study of 21st century skills and engagement in a university Spanish foreign language classroom. *For. Lang. Ann.* 52, 822–849. doi: 10.1111/flan.12426

Creswell, J. W. (2002). *Educational Research: Planning, Conducting, and Evaluating Quantitative, Vol. 7*. Upper Saddle River, NJ: Prentice Hall.

Damari, R. R., Rivers,W. P., Brecht, D. R., Gardner, P., Pulupa, C., and Robinson, J. (2018). The demand for multilingual human capital in the U.S. labor market. *For. Lang. Ann.* 50, 13–37. doi: 10.1111/flan.12241

Daniel, J. (2020). Education and the COVID-19 pandemic. *Prospects* 49, 91–96. doi: 10.1007/s11125-020-09464-3

Demirel, M. (2009). Lifelong learning and schools in the twenty-first century. *Proc. Soc. Behav. Sci.* 1, 1709–1716. doi: 10.1016/j.sbspro.2009.01.303

Dündar, S. (2009). Analyzing the relationship between problem solving skills and personality characteristics of university students. *Dokuz Eylul Univ. Fac. Econ. Admin. Sci. J.* 24, 139–150.

Dündar, S. A., and Polat, A. (2021). Examining the curriculum of teaching Turkish as a foreign language in the scope of 21st century skills. *Int. J. Karamanoglu Mehmetbey Educ. Res.* 3, 66–77. doi: 10.47770/ukmead.995111

Edwards, B. I., and Cheok, A. D. (2018). Why not robot teachers: artificial intelligence for addressing teacher shortage. *Appl. Art. Intell.* 32, 345–360. doi: 10.1080/08839514.2018.1464286

Eker, E. (2020). *Exploring EFL Teachers' Perceptions on 21st Century Skills: A Case Study*. Mersin: Institute of Social Sciences, Çag University.

Engin, A. O., and Korucuk, M. (2021). Investigation of students' 21st century skills in terms of various variables. *Gazi Univer. J. Gazi Educ.* 41, 1081–1119. doi: 10.17152/gefad.875581

Epçaçan, C. (2013). The effect of teaching applications based on multiple intelligence theory on students' reading comprehension skills. *Turk. Stud.* 8, 1335–1353. doi: 10.7827/TurkishStudies.4336

Erdogan, D. (2018). *Investigation of the Relationship Between Turkish Pre-Service Teachers 21st Century Skills and Lifelong Learning Tendencies*. Zonguldak: Zonguldak Bülent Ecevit University.

Erdoğan, F. (2020). The relationship between prospective middle school mathematics teachers' critical thinking skills and reflective thinking skills. *Participat. Educ. Res.* 7, 220–241. doi: 10.17275/per.20.13.7.1

Erice, D., and Ertaş, A. (2011). The impact of e-portfolio on foreign language writing skills. *Ankara Univer. J. Fac. Educ. Sci.* 44, 73–94. doi: 10.1501/Egifak_0000001225

Ertmer, P. A., and Ottenbreit-Leftwich, A. T. (2010). Teacher technology change: how knowledge, confidence, beliefs, and culture intersect. *J. Res. Technol. Educ.* 42, 255–284. doi: 10.1080/15391523.2010.10782551

Eubanks, J., Yeh, H., and Tseng, H. (2018). Learning Chinese through a twenty-first century writing workshop with the integration of mobile technology in a language immersion elementary school. *Comp. Assist. Lang. Learn.* 31, 346–336. doi: 10.1080/09588221.2017.1399911

Frith, E., Elbich, D. B., Christensen, A. P., Rosenberg, M. N., Chen, Q., Kane, M. J., et al. (2021). Intelligence and creativity share a common cognitive and neural basis. *J. Exp. Psychol. Gen.* 150, 609–632. doi: 10.1037/xge0000958

Fullan, M. (2013). *Great to Excellent: Launching the Next Stage of Ontario's Education Agenda*. Available online at: http://www.michaelfullan.ca (accessed November 26, 2022).

Gandasari, A., Wahyudin, D., Abdulhak, I., and Djohar, A. (2022). The national curriculum implementation into the operational curriculum based on multiple intelligences theory (research dissemination before pandemic Covid-19). *Indones. J. Soc. Sci.* 14, 14–25. doi: 10.20473/ijss.v14i1.33003

Garba, S. A., Byabazaire, Y., and Busthami, A. H. (2015). Toward the use of 21st century teaching-learning approaches: the trend of development in Malaysian schools within the context of Asia Pacific. *Int. J. Emerg. Technol. Learn.* 10, 72–29. doi: 10.3991/ijet.v10i4.4717

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligence*. New York, NY: Basic Books.

Gardner, H. (1999). *The Theory of Multiple Intelligences*. New York, NY: PSIA—Rocky Mountain Division—AASI.

Gardner, H., and Veenema, S. (1996). Multimedia and multiple intelligences. *Am. Prosp.* 7, 1554.

Geisinger, K. F. (2016). 21st century skills: what are they and how do we assess them? *Appl. Measure. Educ.* 29, 245–249. doi: 10.1080/08957347.2016.1209207

Geisler, M., Kramsch, C., McGinnis, S., Patrikis, P. P., Louise, M., Karin, R., et al. (2007). Foreign languages and higher education: new structures for a changed world. *Profession* 234–245. doi: 10.1632/prof.2007.2007.1.234

Genç, M. (2012). Investigating of the relation between the teachers' multiple intelligences and problem solving skills. *Bartin Univ. J. Fac. Educ.* 1, 77–88. doi: 10.14686/201212001

Göktepe Yildiz, S. (2020). An investigation of 21st century skills of high school students in terms of some demographic variables. *Ulakbilge J. Soc. Sci.* 51, 884–897. doi: 10.7816/ulakbilge-08-51-02

Güler, T., and Kalin Sali, M. (2021). The use of social networks in teaching Turkish as a foreign language: a case study. *Int. J. Innov. Approach. Educ.* 5, 236–254. doi: 10.29329/ijiape.2021.374.1

Gülmez, A. (2018). *Views of academicians about multicultural education: North Cyprus case*. (Master Thesis), Near East University, Nicosia. Retrieved from http://docs.neu.edu.tr/library/6738831624.pdf in (accessed May 14, 2022).

Güneş, F. (2011). Language teaching approaches and their applications in teaching Turkish. *Mustafa Kemal Univ. J. Inst. Soc. Sci.* 8, 123–148.

Güngör, H. (2021). "Teaching Turkish as a foreign language in the context of 21st century skills," in *21st Century Skills and Turkish Education*, ed H. E. Çocuk, and T. Sosyal (Ankara: Nobel).

Günindi, Y. (2014). Usage of mobile technologies in social skills development. *Creat. Educ.* 5, 1576. doi: 10.4236/ce.2014.516174

Güntaş, S., Güneyli, A., and Gökbulut, B. (2021). Assessment of the effectiveness of blended learning in foreign language teaching: Turkish language case. *Laplage Rev.* 7, 468–484. doi: 10.24115/S2446-622020217Extra-B926p.468-484

Hamarat, E. (2019). 21. yüzyil becerileri odaginda Türkiye'nin egitim politikalari. *Seta Analiz* 272, 1–24.

Hayta, F. (2014). *An examination of language learning strategies with reference to computer and mobile phone technology*. (Master's Thesis), Diyarbakır: Graduate

School of Educational Sciences English Language Teaching Department; Dicle University (Turkey).

Hong, S. B., Park, J., Moon, Y., Grandmason, M. L., Nowrowski, D. P., and Moon, M. (2017). Learning a foreign language in adulthood using principles of neuroscience. *ARC J. Neurosci.* 2, 10–13. doi: 10.20431/2456-057X.0201003

Inal, E., and Arslanbaş, F. (2021). In teaching Turkish as a foreign language communication focused web 2.0 tools and application examples. *Bayburt J. Educ.* 16, 228–249. doi: 10.35675/befdergi.850781

Inan, K. (2021). Online collaborative task design in teaching Turkish as a foreign language. *Bayburt J. Educ.* 16, 74–90. doi: 10.35675/befdergi.846707

International Society for Technology in Education (2017). *ISTE Standards for Educators*. Available online at: https://www.iste.org/standards/for-educators (accessed November 26, 2022).

Ipekşen, S. (2019). *Investigation of the relation between preservice science teachers' 21st century learner skills and learning formats, learning styles and multiple intelligence fields*. (Masters Thesis), Kutahya Dumlupinar University, Kutahya (Turkey).

Kalemkuş, F., and Özek, M. B. (2021). Research trends of the 21th century skills: 2000-2020 (january). *Manas J. Soc. Stud.* 10, 878–900. doi: 10.33206/mjss.774848

Karadag, Y., and Baştug, G. (2018). Recommendations and ethical issues experienced in the process of intelligence assessment in Turkey. *Ankara Univ. J. Health Serv.* 17, 46–57.

Kazazoglu, S., and Bilir, S. (2021). Digital storytelling in l2 writing: the effectiveness of storybird Web 2.0 tool. *Turk. Online J. Educ. Technol.* 20, 44–55.

Keskin, H. (2019). *Examination of "Yedi Iklim Turkish Teaching Set" Which Is Used in Teaching Turkish as a Foreign Language in Terms of Multiple Intelligence Theory*. Bitlis: BEÜ Institute of Social Social; Department of Turkish.

Kim, M., and Han, S.-J. (2015). Nursing students' emotional intelligences and coping strategies. Advanced science and technology letters. *Healthc. Nurs.* 8, 53–56. doi: 10.14257/astl.2015.88.11

Kim, Y. (2019). Global convergence or national variation? Examining national patterns of classroominstructional practices. *Global. Soc. Educ.* 17, 1–25. doi: 10.1080/14767724.2019.1584033

Kinay, D. E. (2017). The usage of social activities in the teaching of Turkish as a foreign language. *Hacettepe Univ. J. Turk. Stud. For. Lang.* 3, 53–75. Retrieved from https://dergipark.org.tr/tr/pub/huydotad/issue/37786/436243

Kiremitçi, O., Meliha, C., and Yildiz, Y. (2014). Evaluation of problem solving skills and the multiple intelligence domains: a study on physical education teacher candidates. *J. Res. Educ. Teach.* 3, 17–24.

Koh, J. L., Sing, C. C., Wong, B., and Huang-Yao, H. (2015). Technological pedagogical content knowledge (TPACK) and design thinking: a framework to support ICT lesson design for 21st century learning. *Asia Pac. Educ. Res.* 24, 535–543. doi: 10.1007/s40299-015-0237-2

Kozikoglu, I., and Altunova, N. (2018). The predictive power of prospective teachers' self-efficacy perceptions of 21st century skills for their lifelong learning tendencies. *J. High. Educ. Sci.* 8, 522–531. doi: 10.5961/jhes.2018.293

Kurudayioglu, M., and Taşkin, S. (2019). An Evaluation of 2018 Turkish course curriculum in terms of 21st century skills. *Ahi Evran Univ. Inst. Soc. Sci. J.* 5, 483–496. doi: 10.31592/aeusbed.621132

LaForce, M. (2018). *How Is Technology Used in Inclusive STEM Schools? STEM School Study*. Chicago, IL: Outlier.

Lei, D. Y., Cheng, J. H., Chen, C. M., Huang, K. P., and James Chou, C. (2021). Discussion of teaching with multiple intelligences to corporate employees' learning achievement and learning motivation. *Front. Psychol.* 12, 770473. doi: 10.3389/fpsyg.2021.770473

Liu, B., and Xin, L. (2018). Research on analysis and application of educational big data based on smart class. *J. Dist. Educ.* 3, 84–93. doi: 10.15881/j.cnki.cn33-1304/g4.2018.03.009

Lomicka, L., and Lord, G. (2018). Ten years after the MLA report: what has changed in foreign language departments? *ADFL Bull.* 44, 116–120. doi: 10.1632/adfl.44.2.116

Looney, D., and Lusin, N. (2018). *Enrollments in Languages Other Than English in United States Institutions of Higher Education, Summer 2016 and Fall 2016: Preliminary Report*. Modern Language Association. Available online at: https://eric.ed.gov/?id=ED590075 (accessed November 26, 2022).

Louis, R. C. (2012). *A Case Study Exploring Technology Integration and Incorporation of 21st Century Skills in Elementary Classrooms*. Boston, MA: College of Professional Studies Northeastern University.

Mettursun, T. (2018). "Multiple intelligence theory in teaching Turkish as a foreign language," in *4th International Student Symposium* (Erzurum: Proceedings Book), 201–220.

Ministry of National Education and Culture of North Cyprus (2021). *2020–2021 Statistical Yearbook*. Department of common services for education, Nicosia. Retrieved from http://eohd.mebnet.net/sites/default/files/2020-2021_Yilligi. pdf (accessed May 14, 2022).

Moeller, A. J., and Abbott, M. G. (2018). Creating a new normal: language education for all. *For. Lang. Ann.* 51, 12–23. doi: 10.1111/flan.12310

Mohr, K., and Welker, R. W. (2017). *The role of integrated curriculum in the 21st century school.* (Dissertation), University of Missouri, St. Louis, MO (USA).

Morska, L., Skibska, J., Sulym, V. T., and Masztalir, V. V. (2018). Didactic potential of the integrated approach to teaching future programmer's professional communicative competence in a foreign language. *Inform. Technol. Learn. Tools* 64, 1–12. doi: 10.33407/itlt.v64i2.2174

Mousa, M. O. (2022). The role of outdoor language arts activities in language learning. *J. Lang. Linguist. Stud.* 18, 461–471.

Ndawo, G. (2021). Facilitation of emotional intelligence for the purpose of decision-making and problem-solving among nursing students in an authentic learning environment: a qualitative study. *Int. J. Africa Nurs. Sci.* 15, 100375. doi: 10.1016/j.ijans.2021.100375

Nevgi, A., Paivi, V., and Hannele, N. (2006). Supporting students to develop collaborative learning skills in technology-based en-vironments. *Br. J. Educ. Technol.* 37, 937–947. doi: 10.1111/j.1467-8535.2006.00671.x,

OECD and PISA (2018). *Database, Table B.1.* Available online at: https://webfs.oecd.org/pisa2018/foreignlanguages-and-careers.xlsx (accessed November 26, 2022).

Osmanli, Ü. (2018). A research on social adaptation of the international students: case of European University of Lefke. *LAÜ Sosyal Bilimler Dergisi* 9, 49–74. Retrieved from https://dergipark.org.tr/tr/pub/euljss/issue/38233/442123 (accessed April 22, 2021).

Özdal, H., Özden, C., Güneyli, A., and Atasoy, R. (2022). Effectiveness of self-regulated Learning skills on web-based instruction at-titudes in online environments. *Pegem J. Educ. Instruct.* 12, 182–193. doi: 10.47750/pegegog.12.01.18

Özkan, H. H. (2008). The relation of multiple intelligence theory and curriculum components. *E J. New World Sci. Acad.* 3, 333–344.

Özkan, T., Demir, K., and Özdemir, A. (2017). Innovative and applicable methodology for teachers teaching Turkish as a foreign language. *J. Int. Ling. Soc. Educ. Sci.* 3, 143–150.

Park, H., and Hiver, P. (2017). Profiling and tracing motivational change in project-based L2 learning. *System* 67, 50–54. doi: 10.1016/j.system.2017.04.013

Parmaxi, A. (2020). Virtual reality in language learning: a systematic review and implications for research and practice. *Interact. Learn. Environ.* 1–13. doi: 10.1080/10494820.2020.1765392. [Epub ahead of print].

Partnership for 21st Century Learning (2021). *P21 Framework Definitions.* Available online at: https://static.battelleforkids.org/documents/p21/P21_Framework_Definitions_New_Logo_2015_9pgs.pdf (accessed November 26, 2022).

Philp, J., and Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Ann. Rev. Appl. Linguist.* 36, 50–72. doi: 10.1017/S0267190515000094

Plucker, J. A., Karwowski, M., and Kaufman, J. C. (2020). "Intelligence and creativity," in *The Cambridge Handbook of Intelligence*, ed R. J. Sternberg (Cambridge: Cambridge University Press), 1087–1105. doi: 10.1017/9781108770422.046

Quicios, J. D. (2018). The North American intelligence community: language management as a vital tool in generating safe and effective future intelligence. *Int. J. Intell. Sec. Public Aff.* 20, 132–154. doi: 10.1080/23800992.2018.1484236

Ravitz, J., Hixson, N., English, M., and Mergendoller, J. (2012). "Using project based learning to teach 21st century skills: findings from a statewide initiative," in *Conference of American Educational Research Association* (Washington, DC).

Repetto, C., Natale, A. F., Villani, D., Triberti, S., Germagnoli, S., and Riva, G. (2021). The use of immersive 360° videos for foreign language learning: a study on usage and efficacy among high-school students. *Interact. Learn. Environ.* 1–16. doi: 10.1080/10494820.2020.1863234. [Epub ahead of print].

Saban, A. (2010). *Multiple Intelligence Theory and Its Presentation in Turkish Education System, 6th Edn.* Ankara: Nobel Publications.

Sahin Timar, Z. (2010). *The evaluation of web based material that is prepared in accordance with the multiple intelligence theory*. (Masters Thesis), Trabzon: Karadeniz Technic Universities (Turkey).

Salman, U., Simşek, A., Turfanda, M., and Salman, A.B. (2017). Intelligence scales used in Turkey. *Istanbul Bilim Univ. Floren. Nightingale Med. J.* 3, 87–89. doi: 10.5606/fng.btd.2017.018

Sardogan, M., Karahan, F., and Kaygusuz, C. (2006). An investigation of university students' indecisiveness strategies in level of the problem solving skill, gender, level of class, and kind of the faculty. *Mersin Univ. J. Educ.* 2, 78–97.

Sarigül, K. (2021). Online formative assessment tools in teaching Turkish as a foreign language. *Int. J. Karamanoglu Mehmetbey Educ. Res.* 4, 56–80. doi: 10.47948/efad.922748

Saydam, M., and Çangal, Ö. (2018). The effect of extra-curricular activities with active participation on the student motivation in teaching Turkish to the foreigners. *J. Theor. Educ. Sci.* 11, 342–356. doi: 10.30831/akukeg.384971

Selçuk, Z., Kayili, H., and Okut, L. (2004). *Multiple Intelligence Activities*. Ankara: Nobel.

Stein-Smith, K. (2016). "The extent and impact of the U.S. foreign language deficit," in *The U.S. Foreign Language Deficit: Strategies for Maintaining a Competitive Edge in a Globalized World,* ed K. Stein-Smith (Cham: Palgrave Macmillan), 13–22. doi: 10.1007/978-3-319-34159-0_2

Sternberg, R. J. (2015). "Multiple intelligences in the new age of thinking," in *Handbook of Intelligence: Evolutionary Theory, Historical Perspective, and Current Concepts*, eds S. Goldstein, D. Princiotta, and J. A. Naglieri. p. 229–241. doi: 10.1007/978-1-4939-1562-0_16

Taylan, U. (2018). VoScreen online foreign language learning environment. *J. Educ. Technol. Online Learn.* 1, 60–69. doi: 10.31681/jetol.376754

Tekiner, A. (2005). *The relationship between perceptual and social learning styles and multiple intelligence and their effects on English Proficiency of Turkish young adults learning English as a foreign Language*. (Unpublished Master Thesis), Middle East Technical University, Ankara (Turkey).

Tilbe, S. (2006). *The effect of Turkish teaching prepared according to multiple intelligence theory on student access and permanence*. (Masters Thesis). Gazi University, Ankara (Turkey).

Toharudin, U. (2017). Critical thinking and problem solving skills: how these skills are needed in educational psychology? *Int. J. Sci. Res.* 6, 2004–2007. doi: 10.21275/ART20171836

Trilling, B., and Fadel, C. (2012). *21st Century Skills: Learning for Life in Our Times*. San Francisco, CA: Jossey-Bass.

Tuzlukova, V., and Prabhukanth, K. (2018). Critical thinking and problem solving skills: english for science foundation program students' perspectives. *Collect. Pap. Fac. Philo.* 48, 37–60. doi: 10.5937/ZRFFP48-18664

Uçar, Y. B. (2021). *The role of emotional intelligence and creativity on entrepreneurial orientation*. (Master Thesis), Ankara: Hacettepe University (Turkey).

Ural, K. (2016). *The use of instruction technologies and language teaching materials in Turkish as a foreign language*. (Masters Thesis), Istanbul Universities (Turkey).

Ustabulut, M. Y., and Keskin, S. (2020). Social media uses-gratification motivations of international students who learning Turkish as foreign language, for the purpose of improving learning effectiveness. *On Dokuz Mayis Univer. J. Educ.* 39, 155–173. Available online at: https://dergipark.org.tr/tr/pub/omuefd/issue/53755/657701

Van Lieshout, C., and Cardoso, W. (2022). Google Translate as a tool for self-directed language learning. *Lang. Learn. Technol.* 26, 1–19. Available online at: http://hdl.handle.net/10125/73460

Waterhouse, L. (2010). Inadequate evidence for multiple intelligences, Mozart effect, and emotional intelligence theories. *Educ. Psychol.* 41, 247–255. doi: 10.1207/s15326985ep4104_5

Wreede, P. (2022). *Utilizing the multiple intelligences theory in classrooms*. (Thesis), Concordia University, St. Paul, CA (USA). Retrieved from: https://digitalcommons.csp.edu/teacher-education_masters/64 (accessed January 7, 2023).

Xu, X., Liu, W., and Pang, W. (2019). Are emotionally intelligent people more creative? A meta-analysis of the emotional intelligence–creativity link. *Sustainability* 11, 6123. doi: 10.3390/su11216123

Yeni, G. (2018). *The effect of 21st century skills training on foreign language teachers' perceptions regarding their educational technology and materials development competencies*. (Masters Thesis), Istanbul Universities (Turkey).

Yilmaz, F., and Babacan, G. (2015). The use of podcasts in teaching Turkish as a foreign language. *Int. Period. Lang. Lit. Hist. Turk. Turkic* 10, 1153–1170. doi: 10.7827/TurkishStudies.7837

Yilmaz, Ö., Gokkaya, G., and Yilmaz, R. (2022). Investigation of factors affecting problem solving skills in technologically rich environments according to PIAAC 2016 data. *J. Inform. Commun. Technol.* 4, 87-109. doi: 10.53694/bited.774159

Yilmaz, R., Karaoglan Yilmaz, F. G., and Keser, H. (2020). Vertical versus shared e-leadership approach in online project-based learning: a comparison of self-regulated learning skills, motivation and group collaboration processes. *J. Comp. High. Educ.* 32, 628–654. doi: 10.1007/s12528-020-09250-2

Young, J. S. (2012). Linking learning: connecting traditional and media literacies in 21st century learning. *J. Media Lit. Educ.* 4, 70–81. doi: 10.23860/jmle-4-1-7

Yücel, H. (2018). *analyzing self-esteem and multicultural perspectives of University students*. (Master Thesis), Near East University, Nicosia. Retrieved from http://docs.neu.edu.tr/library/6684861323.pdf in (accessed May 14, 2022).

Zebari, S., Allo, H., and Mohammedzadeh, B. (2018). Multiple intelligences-based planning of EFL classes. *Adv. Lang. Lit. Stud.* 9, 98–103. doi: 10.7575/aiac.alls.v.9n.2p.98

Zhang, L., Cao, T., and Wang, Y. (2018). The mediation role of leadership styles in integrated project collaboration: an emotional intelligence perspective. *Int. J. Proj. Manage.* 36, 317–330. doi: 10.1016/j.ijproman.2017.08.014

Zhao, B., and Tianyuan, H. (2019). Higher education and its change in the era of artificial intelligence. *Fudan Educ. Forum* 3, 18–25. doi: 10.13397/j.cnki.fef.2019.04.004

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 15 March 2023
DOI 10.3389/fpsyg.2023.1113973



OPEN ACCESS

EDITED BY
Jinfen Xu,
Huazhong University of Science and Technology, China

REVIEWED BY
Rendong Cai,
Guangdong University of Foreign Studies, China
Yuhua Liu,
South China Agricultural University, China

*CORRESPONDENCE
Guiying Jiang
✉ gyjiang118@126.com

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

RECEIVED 01 December 2022
ACCEPTED 10 February 2023
PUBLISHED 15 March 2023

CITATION
Wang D, Jiang G and Zheng Y (2023) Walking out of the light verb jungle: Exploring the translation strategies of light verb constructions in Chinese–English consecutive interpreting. *Front. Psychol.* 14:1113973. doi: 10.3389/fpsyg.2023.1113973



# Walking out of the light verb jungle: Exploring the translation strategies of light verb constructions in Chinese–English consecutive interpreting

Di Wang[1], Guiying Jiang[1]*and Yufan Zheng[2]

[1]College of Foreign Languages and Cultures, Xiamen University, Xiamen, China, [2]School of the English Language and Culture, Xiamen University Tan Kah Kee College, Xiamen, China

Cross-linguistic features of light verb constructions (LVCs) profile a major facet of the typological difference between Chinese and English. By adopting a theory-driven, context-based interpreting task, this study explores the effectiveness and variability of translation strategies in dealing with 12 target LVCs extracted from a Chinese–English Consecutive Interpreting test to capture effective translation strategies fit for Chinese English-as-foreign-language (EFL) learners ($N$=66). Appropriate rates and entropy values denoting variability of strategy selection are calculated by using 12 LVC segments and nine strategies, respectively. A correlation test is also carried out for vocabulary knowledge and the appropriate rates of LVCs to assess the efficacy of learners' vocabulary knowledge in interpreting performance. Results show the general preferences for strategy selection among Chinese EFL learners as well as typical structural patterns in LVC translation. The degree of lightness of the light verbs exerts a reverse effect on the appropriate rates and consistency of strategy selection, and the positive correlation between vocabulary knowledge and LVCs' appropriate rates suggests the need to incorporate the constructional teaching into the EFL learning curriculum. Thus felicitous conditions of applying the strategies have been proposed.

KEYWORDS

**light verbs, light verb constructions, translation strategies, Chinese-English consecutive interpreting, variability of strategy selection**

## 1. Introduction

Light Verb Constructions (LVCs; Grimshaw and Mester, 1988; Rosen, 1989; Butt, 1995; etc.) are basically constituted by a semantically bleached verb (e.g., make, have, get, give) and the action nominal complement (e.g., *answer, advice, help*), most typicssally denoting a motion event or state as in (1a). Its meaning is equivalent to the counterpart synthetic verb derived from the nominal complement as in (1b). In addition, LVCs are analogous to the ditransitive clause in syntactic form as in (1c).

(1) a. Mike gave a kiss to his mother.
b. Mike kissed his mother.
c. Mike gave a book to his brother.

This construction exhibits a cross-linguistic feature, such as Mandarin Chinese (Huang et al., 2014), Japanese (Grimshaw and Mester, 1988), Indo-European languages (e.g., Butt, 1995; Golshaie, 2016; Sundquist, 2018), etc. The common features mainly include: (1) verbo-nominal

combination; (2) semantically semi-compositional; (3) 'light' verbs; and (4) 'heavy' nominal complements.

Yet cross-linguistic studies suggest that the licensing conditions of LVCs vary from language to language. Nagy et al. (2019) focus on the automatic detection of LVCs in four languages, namely, English, German, Spanish and Hungarian; they generalize both common and specific linguistic features from a typological perspective distinguished by five categories, namely, statistical, lexical, morphological, syntactic, and orthographic. Furthermore, the difference can be detected across regional varieties of English. Ronan and Schneider (2015) carried out an automated parser-based study to detect LVCs by inter-variety comparison in the Great Britain component and Ireland component in International Corpora of English. In their study, light verbs with higher frequency are more often used in British English, whereas light verbs with lower frequency tend to be more actively used in Irish English (Ronan and Schneider, 2015). Similar cross-regional investigations of Asia English versions have been carried out by Hoffmann et al. (2011) and Mehl (2017), but show different results in onomasiological preferences. These previous studies illustrate that exploring the common and specific features of LVCs across languages or varieties is weighted toward LVC detection in machine translation and L2 acquisition.

Similar to English, Chinese possesses a robust distribution of LVCs (Huang, 2009), especially in registers, such as official public speeches, legislation and science and technology texts (Wang and Zhang, 2014). As two typologically different languages, English and Chinese have fundamental differences that are embodied in cross-linguistic features of LVC usage, such as the presence/absence of inflectional markers, flexibility of modifications, position of PPs, fixedness of the combinations, etc. These discrepancies increase the difficulties in acquisition for Chinese EFL learners, especially evident in training their interpreting skills, which require the intellectual capacity to instantly transform idioms, colloquialisms and collocations into the equivalent information in the target language. Available Chinese–English comparative studies in LVCs are mainly carried out by Chinese researchers in the field of comparative linguistics, such as semantic properties (Chou, 2019), formal features applied in NLP (Wang and Zhang, 2014; Bai and Xue, 2015); syntactic formation in relation to argument structures (Zhu, 2019), etc. These studies compare five properties of LVCs from different aspects in the two linguistic systems. However, despite their prevalence in both languages, LVCs have not received much attention as formulaic sequences in EFL learners' translations. The fixity of LVCs is presented as a gradient ranging from rigid to free, depending on the syntactic variability and the degree of lexical opacity. Most of the LVCs are not fixed enough to qualify as idioms, but the combinations of the components and modification used are constrained. The semantic complexities of light verbs show "greater cross-linguistic variability than nominal one" (Foley, 2010, p. 84). Such variability across the two languages might be the main cause of difficulties faced by Chinese EFL learners, which attracts interest in the inappropriate usage of LVCs and motivates explorations of corresponding strategies. Light verbs and other high frequency verbs are thought to be a significant barrier for EFL learners because of their limited knowledge (Altenberg and Granger, 2001). However, such knowledge of LVCs is limited in terms of Chinese–English translation strategies in the context of consecutive interpreting performance. Opting for appropriate strategies of the interpretation during learning can save processing and production efforts when retrieving the translation equivalents (McDonald and Carpenter, 1981; Gile, 1995), thus effortlessly bringing about better performance. In addition, locating LVC segments in the production of interpreting text-based sources may truly reflect LVCs usage in a linguistic context. Considering the gaps in the literature and necessity of the research in L2 language acquisition, this study aims to investigate effective translation strategies fit for Chinese EFL learners based on a Chinese–English consecutive interpreting (CECI) test. The study addresses the following three questions:

a. What are the common translation strategies adopted by the professional interpreters when treating LVC segments in consecutive interpreting?
b. How are translation strategies distributed when Chinese EFL learners deal with LVCs in the consecutive interpreting test?
c. How does L2 learners' vocabulary knowledge of English LVCs affect the translation of the target LVC segments in the consecutive interpreting test?

The answers to these questions can help Chinese EFL learners better understand LVCs in both languages and improve their performances in interpreting.

## 2. Literature review: Comparative studies on LVCs between English and Chinese

In this study, the literature review primarily concentrates on comparative studies on LVCs between Chinese and English because translation strategies of LVCs between these two languages have not been addressed in previous studies. This review can serve as the foundation for the discussion of translation strategies by identifying the typical syntacto-semantic features of LVCs in the two target languages.

Studies on English light verbs can be traced back to nearly a century ago. Poutsma (1929) described complex predicate constructions as 'group verbs', including but not limited to light verb constructions, which since then has begun to be noticed for its syntacto-semantic idiosyncrasy. The term 'light verb' was first coined by Jespersen (1954, p. 117) to denote semantically low-content verbs, and 'light verb construction' has since become a commonly accepted term for a bipartite complex predicate (e.g., Grimshaw and Mester, 1988; Rosen, 1989; Butt, 1995, 2003, 2010). Major disputes rest on whether the verb must be pertinent to an isomorphic (zero-derived) form (e.g., *drink* in *to have a drink*), a derivative (e.g., *decide* (v.)—*decision* (n.) in *to make a decision*), or a verbal noun (e.g., *effort* in *to make an effort*; e.g., Wierzbicka, 1982; Quirk et al., 1985; Algeo, 1995; Allerton, 2002; Dixon, 2005), and whether the direct object in

Abbreviations: ASP, aspectual particle; C-E, Chinese-English; CECI, Chinese-English consecutive interpreting; CL, classifier; COMP, complement; L, LVC segment; LV, Light Verb; LV-ed, the past participle form of the light verb; LVC, Light Verb Construction; MOD, modifier; MC, modern Chinese; N, noun; NA, not available; NP, noun phrase; OC, old Chinese; PAR, particle; Prep., preposition; PP, prepositional phrase; SD, standard deviation; TEM4, Test for English Major Grade 4; V, Verb; VK, Vocabulary knowledge.

the construction must be analysed as verbs or as nouns (e.g., Hoffmann et al., 2011).

Studies on Chinese LVCs occurred much later and has reached no consensus on terminology and classification of the light verb. Some use 'formal verb' to highlight its purely formal functions, as the meaning of the light verb is impoverished with no semantic contribution to the clause (e.g., Lv, 1980; Fan, 1981). Others (e.g., Yuan and Xia, 1984; Zhu, 1985) proposed 'delexical verb' or 'quasi-predicate verb', holding that the nominal complement collocating with a light verb must be a two-syllable verb-derived noun or modifier-head combination. Recent Chinese LVC studies mainly center on its syntactic representation and semantic features within Chomskyian generative linguistic sphere (Wen and Cheng, 2007; Zhang, 2013).

A limited number of Chinese–English comparative studies by Chinese researchers concentrate on delineating formal features of LVCs between the two languages (Wang and Zhang, 2014; Bai and Xue, 2015; Zhu, 2019). Under the guidance of Hierarchical Network of Concepts Theory (Huang, 1998), Wang and Zhang (2014) targeted Chinese–English machine translation of LVCs, taking the typical Chinese light verb *jìnxíng* 'make' type as an example. Three linguistic factors have been found to modulate the choice of syntactical structure in Chinese–English translation as follows: the semantic category of the light verb's nominal complement, the syntactical form of LVCs in the target language, and the function of LVCs in the clause where they are located. Three translating rules have been put forward, but their effectiveness and applicability need to further testing. Bai and Xue (2015) attempted to distinguish predicative verbs with vague meaning from the true light verbs by analyzing syntactic and semantic features and argument assignments of the latter. This approach is analogous to Kearns (2002), in which the true light verb (e.g., *give a groan*) and the vague action verb (e.g., *give a demonstration*) are differentiated by testing their passivisation, WH-movement, and pronominalisation, etc. These two studies imply that both languages are involved in the issue of delimitation of the light verb, and syntactic and semantic functions of the nominal complement are the key determining factors to distinguish these two kinds of verbs.

The issue of delimitation of the light verb is aligned with grouping their different types by means of shared semantic attributes. Bai and Xue (2015, p. 11) rank the degree of 'lightness' of five major types of Chinese predicative verbs as shown in (2) (from light to heavy), and thus must be treated differently when annotated.

(2) GIVE > CAUSE = DO > BE > BECOME.

In the same vein, Feng (2016, p. 141) attempted to sequence five types of Chinese light verbs in terms of the degree of grammaticalization as in (3).

(3) CAUSE > TAKE/GET > DO > BE > BE (become/be-with)

Though the two orderings partly overlap, they differ in the research materials, naming of the light verb groups, testing methods, and semantic classification of major types; consequently, the results are different. Similarly, Zhu (2019, p. 155–156) classifies light verbs into four groups, namely, DO, CAUSE, CONSIDER, PREP (a provisional term, indicating a type of covert light verbs that can be used with prepositional phrases), but without sequencing likewise. Furthermore, the naming method that mixes syntactic and semantic attributes obscures classification.

Notably, Feng's (2016) finding mainly comes from his substantial diachronic analysis of old Chinese (OC) compared with modern Chinese (MC). Interestingly, compared with LVCs in MC, those in OC are more similar to those in English (Feng, 2016, p. 114). For example, the light verbs in both OC and English are silent (without phonetic realization), which trigger syntactic shifts and give rise to semantic changes *via* denominalisation or causativisation, as shown in (4–5) (Huang, 2009, p. 2–3):

(4) English and OC denominals: *yú* 'fish or to fish', *shí* 'food or eat', *fàn* 'rice or have rice', *yī* 'clothes or to clothe', *yǐn* 'drink or to drink', etc.
$[_{VP}[_{V}\text{DO}][_{NP}[_{N'}[_{N}$ *yú* 'fish']]]].

(5) English and OC causatives: *bài* 'lose or defeat', *pò* 'break', *hǎo* 'good or to like', *wáng* 'king, to regard as a king', etc.
$[_{VP}[_{V}\text{CAUSE}][_{VP}[_{V}$ *pò* 'break']]].

However, in MC, the silent light verb becomes overt and fills the hypothesized position originally occupied with a silent category (Huang, 2009, p. 2). Thus, the examples in (4–5) can be instantiated by a light verb *dǎ* 'hit' in MC as in (6).

(6) *dǎ yú dǎ fàn dǎ bài dǎ pò*
hunt for fish buy meal hit defeat hit break
'to do fishing to buy meals to cause to defeat to cause to break'
(Huang, 2009, p. 2)

As the two languages evolve, their typological differences emerge. For example, the analytic/synthetic account reveals that English has a large number of bound morphemes to denote the word property. Several action nominal complements in English LVCs are derived from verbs with bound morphemes, such as *contribute* (v.) converted into (*make a*) *contribution* (n.), or *investigate* (v.) into (*give an*) *investigation* (n.), while Chinese uses more free morphemes and has no inflectional markers to distinguish the parts of speech. As such, the nominal complements in Chinese LVCs are often in isomorphic (zero-derived) form. For example, *diàochá* (n.) 'investigation' in (2a) and *diàochá* (v.) 'investigate' in (7b) are identical in form.

(7) a. *zuòle yíge diàochá*
make-ASP one-CL investigation
'made an investigation'
b. *diàochále yíge ànzi*
investigate-ASP one-CL case
'investigated a case'

In addition, paratactic/hypotactic difference indicates that the relationship between the components or clauses is loose and flexible in the Chinese language but conforms to strict order with connectives in English. In Example (8), the theme argument introduced by the Chinese prepositional case marker *duì* 'to' can be placed either before (8a) or after (8b) the subject. However, the syntactic structure in (8b) is not acceptable in English. Such difference can be explained by topic−/subject-dominant account, that topic plays a fundamental role in Chinese clausal constructions. In addition to the canonical syntactic structure SVO like that in English, SOV and OSV are also prevalent in Chinese.

(8) a. *CSI duì zhège ànzi jìnxíngle diàochá.*
CSI to this-CL case proceed-ASP investigation.
'CSI made an investigation of this case.'

b. *duì zhège ànzi, CSI jìnxíngle diàochá.*
to this-CL case CSI proceed-ASP investigation.
'*Of this case, CSI made an investigation.'
(Kuo and Ting, 2007, p. 351)

Such flexibility is similarly exemplified by 'separable LVCs', i.e., a verbo-nominal combination is at times embedded with other attributive components, showing greater syntactical flexibility.

(9) *zuò-le yí-gè quánmiàn-de zǒngjié*
make-ASP one-CL comprehensive-MOD summary
'give a comprehensive summary'

In (9), *quánmiàn* 'comprehensive' separates the light verb *zuò* 'make' and its nominal complement *zǒngjié* 'summary'. This term can also be substituted by more modifiers, either simple adjectives or complex attributive clauses. However, such property varies in degree of flexibility and accessibility among different types of light verbs. Kuo and Ting (2007) divided the light verbs into two types, namely, MAKE group (e.g., *jìnxíng* 'proceed', *zuò* 'do') and GIVE group (e.g., *jiāyǐ* 'give', *yùyǐ* 'give', *gěiyǔ* 'give'). It is claimed that only the insertion of modification by the MAKE group is allowed, and not with the GIVE group. In general, the common type of modification– such as articles, quantifiers, possessors, or adjectives–is placed before the nominal complement in Chinese LVCs. By contrast, Claridge (2000) dichotomizes English LVCs into LV + NP pairings and LV + NP + PP, such as *to run the risk of*. Unlike English, Chinese has no post modifier PP like that in English. The similar component in Chinese LVC sentences often functions as a theme argument projected by the deverbal noun, introduced with or without a preposition. For example,

(10) a. *gōng'ānjú duì zhège ànjiàn jìnxíngle diàochá*
The police office to$_{prep.}$ this-CL case proceed-ASP investigation
'The police office proceeded an investigation to this case.'
b. *gōng'ānjú jìnxíngle zhège ànjiàn-de diàochá*
The police office proceed-ASP this-CL case-MOD investigation
'The police office proceeded an investigation to this case.'

In (10a), the PP *duì zhège ànjiàn* 'to this case' comes before the complex predicate. By comparison, in (10b), the theme argument *zhège ànjiàn* 'this case', as a pre-nominal modifier, is placed directly before the nominal complement *diàochá* 'investigation'.

While Chinese LVCs allow more diversified modifiers, the article usage, modification, and pluralization in English LVCs tend to be more fossilized (Brinton, 2008). Such phenomena can be traced back to the Middle English period, during which adjectival modifications were confined to a small range of adjectives (Matsumoto, 1999, p. 83). For several combinations, no modification has been identified, i.e., LVCs in this situation are lexicalized into fixed expressions, such as *take effect*, *lose sight of*, *give rise to* (Claridge, 2000, p. 157–158). However, Claridge (2000, p. 158) also indicated that, though rather rare, modification is found in well-established units such as *take* full *place*, or *find* so much *fault*, because the noun is salient and independent enough to be modified.

In addition, syntactic operation such as passivisation can render the light verb and its nominal complement separable and inverted, and is sometimes treated as a testing method to distinguish true light verb from vague action verb as mentioned earlier (Saito and Hoshi, 2000; Kearns, 2002; Kuo and Ting, 2007). Notably, the light verb and its nominal complement in a true light verb construction cannot be passivized. Such property is observed in both Chinese and English.

Another issue relates to the semantic difference between an LVC and its simplex predicate verb. It is generally agreed that LVCs are 'semantically more lightweight than the same word would have been in a normal context'. (Allerton, 2002, p. 172) It is the same case in Chinese as is shown in example (11).

(11) a. *zuò-le yí-gè quánmiàn-de zǒngjié*
make-ASP one-CL comprehensive-MOD summary
'give a comprehensive summary'
b. *zǒngjié-de hěn quánmiàn*
summarize-PAR very comprehensively-MOD
'to summarize fully'

In (11a), *Quánmiàn-de* 'comprehensive' means that an overall summing-up has been completed while *quánmiàn* 'comprehensively' in (11b) denotes one typical property of the action 'summarize'. For the record, not all nominal complements can be converted into the counterpart synthetic verb. For example, the change of *effort* into *to make an effort* cannot be used in the form **to effort*. In this case, *make* is a light verb and *effort* denotes an abstract event. Chinese also has a group of specialized event nouns that cannot be converted into the counterpart synthetic verbs, such as *zhànzhēng* 'warfare', *yíshì* 'ceremony', *shǒushù* 'surgery' (Lu, 2012).

In summary, from a broader perspective, the canonical order of LVCs basically shared by Chinese and English is 'the light verb + nominal complement', with or without modifiers such as articles, quantifiers, possessors, adjectives inserted in between. Besides, both languages are characterized by the categorization of true light verbs and vague action verbs, classification of event nouns, semantic differentiation between an LVC and its simplex predicate verb. The differences are reflected mainly in the presence or absence of inflectional markers, flexibility of modifications, position of PPs, fixedness of the combination, etc. These differences pose difficulties in the comprehension of LVCs and lead to various problems in conversion across languages. The present investigation attempts to be carried out with this line of research to delve into translation strategies of LVCs in the CECI test.

# 3. Research methodology

## 3.1. Research design

To capture the effective translation strategies fit for Chinese EFL learners, we follow a theory-driven, top-down procedure in this study. Two criteria are set, which are the guiding principles of extracting LVCs and the common translation strategies extracted from the professional interpreting work as the baseline for comparison. By comparing Chinese EFL learners' performance with professional interpreting work, the predilection of translation strategy selection can be exhibited.

## 3.2. Participants

Sixty-six juniors (all are Chinese natives) majoring in English from a comprehensive university in China are selected for the study.

The score of the National Test for English Major Grade 4 (henceforth TEM4) taken by the end of the second academic year is adopted to measure the participants' general English proficiency. The passing rate of the participants is 57% ($N = 66$, Mean score = 61.92, SD = 7.90), slightly higher than the national average level (=52.69%). To guarantee normal distribution of the target data, we eliminate scores that are three standard deviations above or below the mean as outfielders. At the time of taking the interpreting test designed for the study, the participants had taken 2 years' interpreting training, and are thus presumed to have generally acquired basic interpreting skills, including taking notes, memorizing strategies, analyzing, reconstructing languages. The study protocol is approved by the ethics review board of the university where the tests are carried out. Written informed consent is obtained from all participants. All of the procedures are performed in accordance with the Declaration of Helsinki and relevant policies in China.

## 3.3. Materials

The testing material is a public speech at a press conference and derived from available resource in the public domain[1]. The length of the speech is approximately 3 min, 20 s with 550 Chinese characters. The speed of speech is moderate and the articulation of the speaker is clear.

To provide a full picture of the common strategies adopted in the authentic context of interpreting, we manually collect the translating strategies adopted by the professional interpreters from a self-built small-size Chinese–English parallel corpus of speeches at press conferences held by China's Ministry of Foreign Affairs from 1 July to 31 July, 2021 (85,495 Chinese characters, 64,456 English words, 2,658 sentence pairs). The materials are openly accessible in[2].

## 3.4. Procedure

The interpreting test was arranged in an audio classroom. The test was arranged as one part of the final exam of the interpreting course at the time. The participants were first fully informed with instructions by the course teacher. Participants were allowed to take notes while listening to the soundtrack. The play was paused when the speech reached the natural end, and the participants then began to translate in the target language. The translation works were automatically recorded.

Given the importance of determining whether the related LVC expressions in target language are ready for use in interpreting, an after-test questionnaire about vocabulary knowledge is arranged right after the interpreting test to ensure consistency of the experiment. The items in the questionnaire are all related to the possible English versions of the 12 target LVC segments. The participants are expected to respond with their knowledge about those items. Based on the "vocabulary knowledge scale" designed by Paribakht and Wesche (1997), each item is rated in five scales: a. I have never met the expression before; b. I have seen it before, but I do not know its meaning; c. I have seen it before, and I think I may know its meaning; d. I know it. Its meaning is___ (paraphrase or translation); e. I know how to use it to make up a sentence, for example (if you choose this one, please fill in the blank in d, too.) The full design of the after-test questionnaire is presented in Appendix 1.

## 3.5. Extraction of target LVC segments and annotation principles

Given the limited consensus on the defining features—such as lightness of the light verb, properties of the nominal complement, and degree of modification—the current study follows three guiding principles in selecting and comparing relevant LVCs for the empirical analysis. The principles are generally acknowledged by previous LVC studies in both Chinese and English (Nagy et al., 2019; Zhu, 2019):

a. The choice of target LVCs is limited to the overt light verb and nominal complement combination, considering the types of modification;
b. The light verb is grammaticalized to the extent that attributes its semantically lighter meaning and major syntactical function to the construction as a whole;
c. The nominal complement typically denotes an action or an event, assigning theta-roles of the arguments in the clause.

We firstly generate a frequency wordlist of verbs ($n = 1{,}424$) by using the online word parsing and processing tool *Weiciyun*[3]. Two coders use the three guiding principles to select the light verbs from the wordlist exhaustively and generated 928 concordance lines that were manually scrutinized as target LVC segments in Mandarin in the self-built Chinese–English parallel corpus translated by professional interpreters. Their English counterparts are marked for retrieving translation strategies. Finally, nine strategies are set as the benchmark for comparison with learners' versions, as listed below:

Type A: Literal translation (or transliteration).
LV + (MOD) + N --> LV + (MOD) + N.

(12) *dáchéngle zhòngyào gòngshí*
reach-ASP important common understandings
'reached important common understandings'

Literal translation is preferred when an English LVC equivalent to the Chinese counterpart is available. However, most cases are far more complex due to cross-linguistic differences. Therefore, further translation strategies are necessary to meet the needs, which are illustrated as follows:

Type B: VP conversion.
LV + (MOD) + N --> N-derived V (+MOD).

(13) *zuòle yígè héxīn guīnà*
do-ASP one-CL core summary
'summarized them up into one thing'

1 http://www.gov.cn/gongbao/content/2019/content_5377102.htm
2 https://www.fmprc.gov.cn/
3 www.weiciyun.com

Type C: Verb missing.
LV + (MOD+) N --> (MOD+) N.

(14) *bǎochí wěndìngde shuāngbiān(zhōngměi) guānxī*
maintain steady-MOD bilateral (China-US) relationship
'steady growth of China–US relationship'

Type D: Passive voice.
LV + (MOD+) N --> (MOD+) N + be + LV-ed.

(15) *qǔdéle fēngshùo chéngguǒ*
get-ASP a great deal accomplishments
'a great deal has been accomplished'

Type E: Inverted LVC.
N + (MOD+) LV --> (MOD+) N + LV (+MOD).

(16) *máodùn jiūfēn yě jīngcháng tūchūde biǎoxiànchūlái*
Conflict disputes too often distinctly-MOD present out-COMP
'from time to time problems and difficulties may have occurred'

Note that Type E is derived from a canonical LVC with the light verb and its nominal complement inverted to form an unaccusative clause, and the verb is often modified by an adverb of degree indicating the gradient property. As both Chinese and English have such syntactic structure, literal translation can be used.

Type F: Copula construction.
LV + N1(MOD) + N2 --> N1(MOD) + be + N2.

(17) *bǎochí zhōngměi guānxi zǒngtǐ wěndìng*
keep China–US relationship general stable
'China–US relationship is stable'

Type G: There be construction.
LV + N1(MOD) + N2 --> There be + N2 + Prep. + N1(MOD).

(18) *jìnxíng yìmiáo hùrènde tǎolùn*
proceed vaccine mutual recognition-MOD discussion
'there are discussions on mutual recognition of vaccines'

Type H: PP conversion.
LV + N --> Prep. + N.

(19) *zuòle yígè héxīn guīnà*
do-ASP one-CL core summary
'in conclusion'

Type I: DUI argument shifting.
DUI-NP1 + LV + NP2 --> NP2-derived V + NP1.

(20) *duì xiāngguān jīzhuāngxiāng jìnxíngle xiūfù he qiánghuà*
$to_{prep.}$ Relevant containers proceed-ASP repair and strengthen
'(it) has repaired and strengthened the relevant containers'

The nine strategies summarized above are used as the baseline to identify learners' deviation from the standard or appropriate translation strategies and to observe the different preferences in dealing with LVCs.

In line with the three guiding principles for selection, we also extracted 12 Chinese LVC segments categorized into three major types by their semantic attributes from the testing material, namely, DO, BE, BECOME (Feng, 2005, 2016) from the testing material. Table 1 show the segment classifications.

The English version provided here is translated on site by a Chinese professional interpreter.

The testees' interpreting works are recorded and transcribed after the test (about 22,000 words in total). The basic information listed in the transcribed texts includes the student number, name, class, and test score. In addition, 12 target LVC segments of each testee's version are manually tagged by the types of strategy as stated earlier in this section, and separately marked with tick or cross to indicate translating appropriateness by the two raters. The general plan is to track the proportion of frequency and the acceptability rate of each strategy type adopted by the 12 target LVC segments in response to the first two research questions.

## 3.6. Reliability

The evaluation criterion of this test follows the assessment for Chinese undergraduate students based on interpreting process (Chen, 2017). To guarantee the reliability of the result, two professional interpreting teachers were invited as raters to evaluate the appropriateness of target LVCs. A reliability test is carried out and a high degree of consensus ($r = 0.967$) shows to ensure the consistency of the rating system. In cases of disagreement, a third independent rater can be invited to determine discrepancies to settle the disagreements.

## 3.7. Analyzing methods

Entropy computation is used to measure variability by the 12 LVC segments and the nine strategies, in addition to the regular descriptive statistics to summarize the general features of the current data set, The entropy H of a variable quantifies the degree of randomness or variability (Cover and Thomas, 2005). The formula is presented as below:

$$H(x) = \sum_{x=1}^{n} \mathrm{p}(\mathrm{x}) \log \mathrm{p}(\mathrm{x})$$

For the 12 LVC segments, $x$ denotes each of the nine strategies, and $p(x)$ is estimated with the proportion of strategies that participants have adopted for translating a given LVC segment. The entropy score per LVC segment is a measure of how diverse strategies are used to translate a given LVC segment. The entropy value close to zero indicates either the translating strategies used for a given LVC segment is relatively consistent. In contrast, high entropy means more options have been taken in treating a given LVC segment.

For the nine strategies, $x$ denotes each of the 12 target LVC segments, and $p(x)$ is estimated with the proportion of target LVC segments that a given strategy is applied to. The entropy score per strategy is a measure of consistency of strategy applications. The entropy value close to zero indicates the strategy is consistently

TABLE 1 Target Chinese LVC segments (L1–L12).

| NO | CH | EN | TYPE |
|---|---|---|---|
| L1 | zuòle yígè héxīn guīnà | Sum them up | DO |
| | Make-ASP one-CL core summary | | |
| L2 | qǔdéle fēngshuò chéngguǒ | A great deal has been accomplished | BECOME |
| | Get-ASP rich achievement | | |
| L3 | yǒuzhe guǎngfànde gòngtóng lìyì | There is a broad common interest | BE |
| | Have-ASP broad-MOD common interest | | |
| L4 | bǎochí wěndìngde shuāngbiān guānxi | Steady growth of China-US relationship | BE |
| | Keep steady-MOD bilateral relationship | | |
| L5 | bǎochí zhōngměi guānxi zǒngtǐ wěndìng | Maintaining the overall stability of China-US relationship | BE |
| | Keep China-US relationship overall stability | | |
| L6 | máodùn jiūfēn yě jīngcháng tūchūde biǎoxiàn chūlái | Problems and difficulties may have appeared | BECOME |
| | Problem dispute too often prominently-MOD present out-COMP | | |
| L7 | dáchéngle zhòngyào gòngshí | Reached important common understandings | BECOME |
| | Reach-ASP important common sense | | |
| L8 | jìnxíng chuōshāng | Consultations between the two sides on economic and trade issues are still under way | BE |
| | Make consultation | | |
| L9 | shíxiàn hùlì gòngyíng | Deliver win-win and mutual benefits to the two countries | BECOME |
| | Achieve mutual benefit win-win | | |
| L10 | tuīdòng zhōngměi guānxide fāzhǎn | Continue to grow China-US relationship, including their economic and trade ties | BECOME |
| | Promote China-US relationship-MOD development | | |
| L11 | jìnxíng huàjiě guǎnkòng | Defuse their differences and manage them properly | BE |
| | Proceed solution control | | |
| L12 | tuīdòng fúhé shìjièchǎoliúde zhōngměiguānxi wěndìng fāzhǎn | Pursue steady and sound growth of China-US relationship | BECOME |
| | Push fit world trend-MOD China-US relationship steady growth | | |

applied to relatively fewer limited LVC segments. In contrast, high entropy implies that the strategy is employed by a wider range of LVC segments.

Entropy enjoys increasing use in the language sciences (e.g., Montemurro and Zanette, 2011; Gries, 2012) for its preponderance in variability computation. It can be used for computing both categorical and continuous variables. In addition, it is comparable across individuals and categories due to a specific quantification of the variability value.

In answer to the third research question, a correlation test is carried out between English vocabulary knowledge and the appropriate rates of LVCs in the interpreting test.

## 4. Results

In response to the three research questions, three sets of data are reported: (1) distribution of translation strategies (including the proportion of frequency and entropy values) used by the participants for the 12 LVC segments; (2) the appropriate rates of the 12 LVC segments and the nine translation strategies, as well as their interrelation with the semantic attributes of light verbs; and (3) the correlation between vocabulary knowledge of LVCs and the interpreting score of the 12 target LVC segments.

In general, literal translation type (A) takes up the highest proportion in selection (54.92%), followed by Types C (10.61%), B (6.44%) and G (6.31%; see Table 2). In terms of entropy value by LVC segments (Figure 1), the scores of L8, 9, 10, 12 are relatively low, which implies that when they are translated into English, the translation strategies used by the participants are relatively consistent and mostly centralized in Type A. These LVC segments mainly pertain to BE and BECOME grouped by semantic attributes of the light verb. By contrast, L1 (DO group) shows the highest entropy ($H = 0.643$) indicating high variability in strategy selection distributed in almost all strategies except Types E and F. For entropy value by strategies (Figure 2), the scores of Types H, I and E are relatively low, implying that those strategies are applied to a limited number of LVC segments. By contrast, Type A shows the highest entropy ($H = 1.02$) indicating its wide application in various LVC segments.

The appropriate rates of the 12 LVC segments (Figure 3) and of the nine translation strategies (Figure 4) are unevenly distributed. A relatively higher rate is achieved in L3, 4, 5, 7, 10 and in Types A, B, C, respectively, (given that only one case is translated as Type H, the result can be ignored).

To obtain a closer look at the appropriate rate by semantic attribute groups, we compute the grand means of appropriate rate using the three semantic attribute groups, as shown in Figure 5. The BECOME group has the highest rate, and the DO group has the lowest.

TABLE 2 Raw data of translation strategies for the 12 LVC segments.

| LVC strategy | A | B | C | D | E | F | G | H | I | NA[a] | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| L1 | 21 (13) | 14 (12) | 6 (4) | 1 (1) | 0 (0) | 0 (0) | 3 (3) | 0 (0) | 0 (0) | 21 (0) | 66 (33) |
| L2 | 43 (29) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 5 (1) | 3 (2) | 0 (0) | 0 (0) | 15 (0) | 66 (32) |
| L3 | 38 (35) | 0 (0) | 10 (4) | 0 (0) | 1 (1) | 2 (0) | 9 (8) | 0 (0) | 0 (0) | 6 (0) | 66 (48) |
| L4 | 35 (30) | 1 (0) | 16 (14) | 1 (1) | 0 (0) | 4 (2) | 0 (0) | 0 (0) | 0 (0) | 9 (0) | 66 (47) |
| L5 | 34 (28) | 2 (2) | 8 (6) | 0 (0) | 0 (0) | 11 (11) | 3 (1) | 1 (1) | 0 (0) | 7 (2) | 66 (51) |
| L6 | 11 (9) | 0 (0) | 1 (1) | 0 (0) | 14 (11) | 8 (4) | 30 (20) | 0 (0) | 0 (0) | 2 (0) | 66 (45) |
| L7 | 25 (19) | 1 (1) | 37 (34) | 0 (0) | 0 (0) | 1 (0) | 0 (0) | 0 (0) | 0 (0) | 2 (0) | 66 (54) |
| L8 | 56 (41) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 1 (0) | 0 (0) | 0 (0) | 9 (0) | 66 (41) |
| L9 | 57 (45) | 0 (0) | 5 (1) | 0 (0) | 1 (0) | 0 (0) | 1 (0) | 0 (0) | 0 (0) | 2 (0) | 66 (46) |
| L10 | 55 (47) | 1 (1) | 0 (0) | 0 (0) | 1 (0) | 5 (2) | 0 (0) | 0 (0) | 0 (0) | 4 (0) | 66 (50) |
| L11 | 4 (0) | 31 (24) | 1 (1) | 2 (0) | 0 (0) | 0 (0) | 0 (0) | 0 (0) | 24 (9) | 4 (0) | 66 (34) |
| L12 | 56 (40) | 1 (1) | 0 (0) | 0 (0) | 0 (0) | 3 (2) | 0 (0) | 0 (0) | 0 (0) | 6 (0) | 66 (43) |
| Total | 435 (336) | 51 (41) | 84 (65) | 4 (2) | 17 (12) | 39 (22) | 50 (34) | 1 (1) | 24 (9) | 87 (2) | 792 (524) |
| Percent (%) | 54.92 | 6.44 | 10.61 | 0.51 | 2.15 | 4.92 | 6.31 | 0.13 | 3.03 | 10.98 | 100 |

The correct number is shown in the bracket.[a]NA refers to omissions in interpreting.

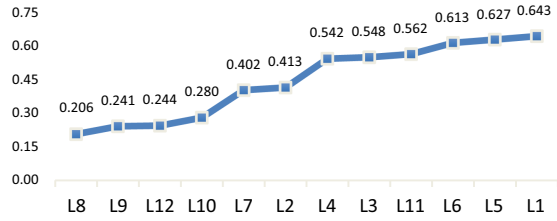



FIGURE 1
Entropy value by LVC segments.

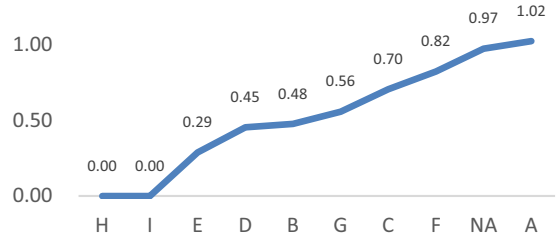



FIGURE 2
Entropy value by strategies.

We also compute the grand means of entropy value by semantic attribute groups in a similar manner, and the results are given in the reverse order than those of the appropriate rates ($Mean_{DO} = 0.643$; $Mean_{BE} = 0.497$; $Mean_{BECOME} = 0.366$). These results indicate that as the light verb becomes more abstract, the translation appropriate rate decreases and the variability increases (or consistency decreases) in strategy selection. In summary, the degree of lightness of the light verbs exerts a reverse effect on the appropriate rate and consistency of strategy selection.

Furthermore, a correlation test is carried out between the English vocabulary knowledge related to the 12 target LVCs and the appropriate rates of the 12 target LVC segments in the interpreting test. Given that both of the testing variables are normally distributed, and the assumption of linearity is not markedly violated, Pearson correlations are computed to examine the two variables. The result shows a significant correlation ($r = 0.272$, $p < 0.05$). According to Cohen (1988), the vocabulary knowledge and the appropriate rates of the 12 target LVC segments have a positive correlation, which is considered as a close medium effect size ($r = 0.3$). This finding means that students with a good command of target topic-related lexical knowledge are likely to have high LVC scores in the interpreting test.

## 5. Discussion

In this section, under the background of typological differences and syntactic and semantic properties of the construction, we mainly discuss the achieved results from three motivating factors proposed by the research questions: a. preferences of strategy selection; b. structural patterns in LVC translation; and c. relations of lexical knowledge and appropriate use of LVCs. Thereupon, the felicitous conditions of the nine translation strategies are summarized along with related pedagogical implications.

### 5.1. Preferences of strategy selection

Different processing inclinations are identified and discerned by observing the participants' preferences for different translation strategies in treating LVCs during the C-E interpreting test.

Based on our observation from the data, the most frequently-used strategy is literal translation. The reason is obvious: literal translation saves the processing effort so as to allocate more attention where necessary. However, the misuse of light verbs frequently occurs, as examples in (21):

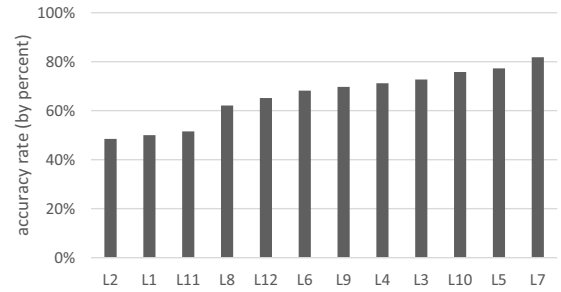



FIGURE 3
The appropriate rates of the 12 LVC segments (by percent).

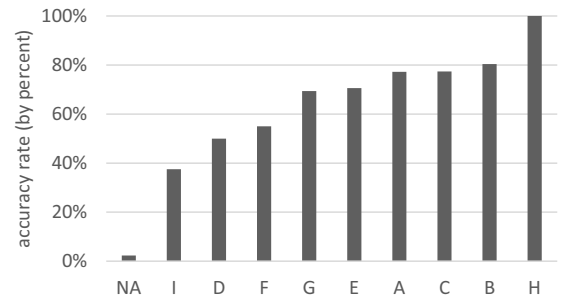



FIGURE 4
The appropriate rates of the nine translation strategies (by percent).

(21) a. * 'to have an conclusion'
'to draw a conclusion'
b. * 'to move stable and healthy relationship'
'to maintain stable and healthy relationship'

Without knowledge of an equivalent LVC, the testee would preferably choose Types B or C, which are retaining the nominal complement only or converting the noun into a verb, respectively. However, the potential problem of choosing Type C is that, when the light verb is removed, no predicate is left in the clause. If the testee cannot find a syntactic predicate for the nominal complement at the time, an unacceptable expression may be produced. Here is an ill version in (22):

(22) *tuīdòng fǔhé shìjiè cháoliúde zhōngměi guānxi wěndìng fāzhǎn*
push fit world trend-MOD China-US relationship steady growth
*'…the world relationship and our development to the health and steady…'

In the translated version of L12, only a nominal complement *fāzhǎn* 'development' is retained. The testee cannot reorganize the sense relations of the source LVC segment with limited processing capacity, and fails to properly translate the light verb *tuīdòng* 'push' and the complex attribute clause that modifies the noun phrase.

The choice of Type B indicates that the participants assume that LVCs and their counterpart synthetic verbs are interchangeable. If no matchable LVC is available for conversion, a nominal-derived verb might be used. However, different from the simplex verbal predicate,

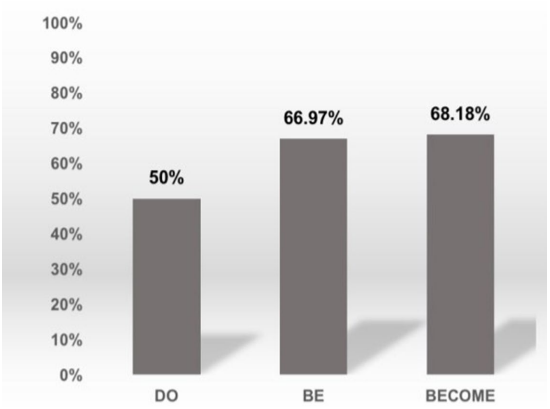



FIGURE 5
The appropriate rates grouped by semantic attributes.

the light verb in LVCs is proposed to serve an aspectual function (Wierzbicka, 1982), and thus the two forms are actually not identical. However, a corpus study indicates that it is the ease and variety with the usage rather than the semantic minuscule difference that motivates the use or disuse (Bonial and Pollard, 2020). The data of the current study validate the statement, given that the choice of Types B or C shows no regular tendency.

Besides Types B and C, Type G (*There be construction*) is another frequently-used strategy in translation of the target LVC segments, especially in L6, but its appropriate rate seems low (=63.18%), as in (23):

(23) *máodùn jiūfēn yě jīngcháng tūchūde biǎoxiànchūlái*
problem dispute too often prominently-MOD present out-COMP
'There are conflicts and disputes that arise frequently'

This unaccusative clause is inverted from a canonical verbo-nominal word order. The statistics shows that a preferable option to translate such derived LVC is *There-be construction* for lack of a volitional subject. However, the misuse of the construction mainly lies in confounding the existential predicate *There be* with the light verb. If a testee fails to arrange the two into an appropriate hierarchical structure, then errors will follow, such as:

(24) *'there have a lot of conflicts'
*'there are appear some disputes'
*'there are still some frictions emerge'

## 5.2. Structural patterns in LVC translation

The statistics indicate that the appropriate rates of the 12 LVC segments (Figure 3) are relatively high in L3, 4, 5, 7,10, and low in L1, 2, 6, 8, 9, 11, 12. Further observation of participants' performances reveals that the main reason for mistranslation is the differences of modification patterns in the two languages. Though internal grammatical modification is a common feature in both Chinese and English LVCs, the intricate difference may cause inappropriate translation.

One major difference is that Chinese LVCs allow more diversified modifiers, while English LVCs tend to be more fossilized. The tighter connection of such LVCs becomes more unified such as a simplex verb, increasing the probability of taking a post adverbial component. In this case, strategy B is often adopted in C–E conversion. That is to say, the Chinese separable LVCs with adjectival modification in between may be converted into a synthetic verb counterpart with a post adverbial modification in English as in (11), if the target language has no equivalent LVC available in the target language. The LVC segment L1 selected from the testing material is a similar case, as illustrated in (24).

(25)a. *zuò-le yí-gè héxīn guīnà*
make-ASP one-CL core-MOD summary.
'sum them up'.
b. *do a core summary.

In 25(a), the verb phrase 'sum up' is used to translate the LVC expression in the professional version. However, a participant who fails to find the appropriate equivalence adopts a literal translation and overused the delexical verb 'do' to form an unacceptable LVC expression. Moreover, cautious consideration is needed in that prenominal modifications in Chinese LVCs are not always adjectival but rather may be nominal or an expression with dual semantic properties, such as *héxīn* 'core' in Example (25). The nominal modification functions as a genitive case and its thematic role is assigned by the verbal noun rather than the light verb. For example:

(26) *bǎochí zhōngměi guānxi zǒngtǐ wěndìng*
keep China-US relationship overall stability
'maintaining the overall stability of China–US relationship'

In (26), the verbal noun *wěndìng* 'stability' (NP2) retains its semantic relationship with the nominal modifier *zhōngměiguānxi* 'China–US relationship' (NP1), forming an internal semantic relation--assigner (NP2) - assignee/patient (NP1). When converting into English, the common treatment is to leave NP1 behind NP2 led by a preposition, i.e., LV + NP2 + Prep. + NP1. Given that post modification is rare in Chinese NP, if an L2 learner is not aware of this cross-linguistic distinction, then an erroneous translation is likely to occur in limited working time.

Another major difference in this regard is that Chinese LVCs are at sometimes separated by prenominal adjectival components. However, in English, the presence/absence of the prenominal adjectival modification may be affected by frequency and syntactic fixity of the collocation. The nominal modifier in a Chinese separable LVC is frequently represented by 'DUI-insertion' (Zhu, 2019). 'DUI' represents all the prepositional case markers, such as *duì* 'to', *duìyú* 'as for', *bǎ* 'about', or *gēn* 'along with'. This approach is equivalent to the function of a preposition inserted in an English LVC, introducing the patient of the verbal noun. 'DUI' is also often inserted between the subject and the LVC, but sometimes it may move forward to the head as the topic of the sentence as in (27).

(27) *duìyú máodùn hé fēnqí jìnxíng huàjiě guǎnkòng*
as for conflict and disagreements proceed solution control
'As for the differences and disagreements, we have confidence to defuse their differences and manage them properly.'

In this sentence, the theme argument *máodùn hé fēnqí* 'differences and disagreements' of LVC introduced by a preposition *duìyú* 'as for' is topicalized and projected by the LVC segment *jìnxíng huàjiě guǎnkòng* 'proceed solution and control'. In fact, the topicalized theme argument semantically functions as the direct object of verbal noun phrase *huàjiě guǎnkòng* 'solution and control'. However, different from Chinese which is a topic-prominent language, English is subject-prominent. Hence, the common practice in dealing with the above Chinese LVC is to translate into a canonical SVO English structure with the verb derived from the nominal complement and the object attained from a prepositional argument. Example (20) of Type I illustrates this approach. Our data show that the appropriate rate of this strategy is only 37.50%. In addition, the participants do not favour Type I ($n = 24$; percentage≈3.03%), which implies that most testees are not used to applying the strategy into C–E interpreting.

Another observation is that most testees prefer attributive to adverbial modifications in their English versions. The main reason is that literal translation is most frequently used than other strategies. In Chinese, an attributive modification is normally placed before its head noun, and the English version is likely to follow such word order. Similar preference is also discussed by Fleischhauer and Neisani (2020) in Persian separable LVCs: although their study focuses on a different language, their findings present significance in understanding LVCs in general. Like Persian, many Chinese adverbials are either overtly or non-overtly derived from adjectives, as given in Example (11). No inflectional marker is available to use to distinguish adjectives and adverbials, which share identical lexical forms that are not clear-cut in most cases. One typical test for distinction is to use the post-modification particle *de* (的) for adjectives and *de* (地) for adverbials. The internal attributive modifications in Chinese LVCs are mainly regarded as adjectives. However, internal modifiers in LVCs do not always share similar semantic functions with those in counterpart synthetic verbs, as L7 shows in (28).

(28) *dáchéngle zhòngyào gòngshí*
reach-ASP important common understanding
'reached important common understandings'

Some argue that the internal modifiers in Chinese LVCs basically modify the whole construction rather than the nominal components. However, in (27), 'important' modifies the nominal component 'common understandings', for the adjective specifies the importance of understanding, i.e., the mutual goal shared by the political leaders is crucial for the future friendly negotiation. This part cannot be paraphrased as 'understand importantly'. Therefore, learners need to know the felicitous conditions of the modification in the target language to achieve proper translation.

## 5.3. Relations of lexical knowledge and appropriate use of LVCs

The correlation test performed between the target LVCs knowledge and the appropriate rates of the 12 target LVC segments shows a significant positive result. Thus EFL learners with a good command of target topic-related lexical knowledge are very likely to

have high LVC score in the interpreting test. Furthermore, the grand means of entropy value by semantic attribute groups shows that the degree of lightness of the light verbs exerts a reverse effect on the appropriate rate and consistency of strategy selection. Specifically, as the light verb becomes more delexicalized, the appropriate rate in translation decreases, and the variability increases (or consistency decreases) in strategy selection.

This study groups 12 Chinese LVC segments into three types based on their semantic attributes, i.e., DO, BE, BECOME, illustrated in Table 1. According to Feng (2016), the degree of grammaticalization can be ordered from high to low as in (28) below.

(29) DO>BE>BECOME

The appropriate rates grouped by semantic attributes (Figure 5) are consistent with the order of the degree of grammaticalization: the former decreases as the latter increases, and vice versa. This result implies that the lightness of the light verb affects the equivalence of interpreting. In addition, the entropy value computed in the same fashion further indicates that the lightness of the light verbs might affect strategy selection. Being semantically light yet functionally complex, such light verb can have multiple treatments in translation. Given that most light verbs are polysemous, and LVCs are typical for their complex predicates with complex meanings, unskillful learners may be easily confused in discriminating 'light' usage from the 'heavy' sense. Naturally, intermediate EFL learners may have difficulties in such indirect and obscure matching process. As argued by Butt (2010), the light verb in LVCs contributes a generic meaning rather than an actual motion concept, namely, a full verb. Therefore, the degree of lightness of the light verbs exerts a reverse effect on the appropriate rates and consistency of strategy selection.

## 5.4. Felicitous conditions of the strategies and implications on formulaic language learning

Considering the three aspects discussed above, applicable conditions of the nine LVC strategies are briefly shown in Figure 6.

If an LVC expression is available in both Chinese and English, then Type A is the best option. If no direct equivalence is available in the target language, then the major semantic bearer—the verbal noun—may play a key role in the conversion (Types B, C, H), or syntactic transformation may be considered (Types D, F, G). Both Chinese and English have inverted LVC, and thus literal translation (Type E) can be used. As for the special case of Chinese non-canonical LVC with DUI-insertion, a canonical SVO English structure (Type I) can be used with the verb derived from the nominal complement and the object attained from the prepositional argument.

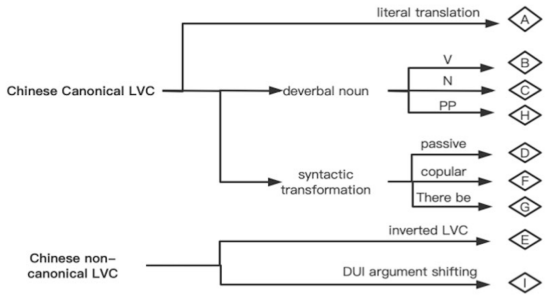



FIGURE 6
Felicitous conditions of LVC translation strategies.

In addition to the common translating issues illustrated above, other minor problems occur in this test, such as morphological misuse (especially between the nominal verbs and deverbal nouns), syntactic variation of LVCs (such as passivity in L6), and PP shift (as illustrated in Types H and I). Moreover, the strategy selections illustrated above are not mutually exclusive, but work together in translation.

Apart from major linguistic factors, one's formulaic knowledge may also affect the selection of translation strategies. The proper use of formulaic sequences in the target language may bridge the gap between native speakers and L2 learners (e.g., Wray, 2002; Wood, 2010), and a few illuminating attempts to quantify pattern variability of fomulaicity in text registers (e.g., Roemer, 2010; Forsyth and Grabowski, 2015). Langacker (2008) considered 'formulaic representation' as the basic tool to coginitively understand language. The concept of 'slate' in language is represented in various linguistic levels such as pronunciation, lexical, syntax. These modular slate structures are believed to be the smallest unit of English language communication. Formulaicity forms a gestalt as a holistic representation. As mentioned in Section 5.2, English LVCs tend to be more fossilized in contrast to those of Chinese. The fixed combination calls for L2 learners' awareness of formulaicity in the target language. Formulaic sequence is neither a word nor a syntactical structure but a lexical–syntactical continuum to construct a text. It has low decomposability and high cohesion, which renders it considerable advantages in bilingual transformation. However, most frequency-driven approaches have not yet been widely applied in translation and interpreting studies. Possibly, different formulaic constructions are treated with no difference. The correlation test between the after-test questionnaire on vocabulary knowledge and the LVCs appropriate rates in the current study supports the claim that formulaic use improves fluency and accuracy of interpreting by helping to alleviate the limited cognitive load and promote his processing capacity to save time (Tang and Li, 2013, 2016).

# 6. Conclusion

This study explores the appropriateness and variability of translation strategies in dealing with the 12 target LVC segments by L2 learners to capture effective translation strategies fit for Chinese learners of English in this regard. The results show that the frequency of use and the appropriate rates of nine types of translation strategies are distributed unevenly; as such, preference of strategy selection conforms to the economy principle, i.e., saving the processing effort to allocate more attention to wherever necessary. Moreover, the consistency between the appropriate rates and of strategy selection grouped by semantic attributes and the order of the degree of grammaticalization implies that the lightness of the light verb affects the appropriateness and strategy selection of translation, i.e., the degree of lightness of the light verbs exerts a reverse effect on the appropriate rates and consistency of strategy selection. Meanwhile, the positive correlation between the after-test questionnaire and interpreting score shows that a good command of target topic-related lexical knowledge helps to improve the interpreting performance. The

teacher needs to determine whether the L2 learners have or have not appropriately acquired the target expressions to determine whether it is the strategy itself or unfamiliarity with the target expression that leads to failure in bilingual translation. The felicitous conditions of the nine translation strategies are thus delineated. The findings support that cultivating the awareness of formulaicity and acquiring translation strategies help L2 learners develop a set of matching system across languages, and improve fluency and accuracy of C-E interpreting.

As an attempt to explore the interpreting strategies of LVC, the current study only offers data observed from Chinese to English translation, and not the other way around. A full picture can be obtained if bidirectional interpreting tasks can be designed. We hope that future studies will further probe into the issue by providing more comprehensive experimental support.

## 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 Xiamen University Tan Kah Kee College. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

DW: the conception and design of the work, the acquisition, analysis, and interpretation of data for the work, drafting the original version. GJ: the conception and design of the work, revising the draft critically for important intellectual content. YZ: data collection, revising the draft critically. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the Humanities and Social Science General Program sponsored by the Ministry of Education of the People's Republic of China (grant number 18YJA740047).

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1113973/full#supplementary-material

## References

Algeo, J. (1995). "Having a look at the expanded predicate," in *The verb in contemporary English: theory and description*. eds. B. Aarts and C. F. Meyer (Cambridge: Cambridge University), 203–217.

Allerton, D. J. (2002). *Stretched verb constructions in English*. London: Routledge.

Altenberg, B., and Granger, S. (2001). The grammartical and lexical patterning of MAKE in native and non-native student writing. *Appl. Linguist.* 22, 173–195. doi: 10.1093/applin/22.2.173

Bai, X., and Xue, N. (2015). A study on argument structure of 'V + event NP-Obj' construction in Chinese from a multi-lingual perspective. *J. Foreign Lang.* 38, 2–13.

Bonial, C., and Pollard, K. A. (2020). Choosing an event description: what a PropBank study reveals about the contrast between light verb constructions and counterpart synthetic verbs. *J. Linguist.* 56, 577–600. doi: 10.1017/S0022226720000109

Brinton, L. J. (2008). "Where grammar and lexis meet," in *Theoretical and empirical issues in grammaticalization*. eds. E. Seoane and M. J. López-Couso (Amsterdam: John Benjamins), 33–53. doi: 10.1075/tsl.77.04bri

Butt, M. (1995). *The structure of complex predicates in Urdu*. San Antonio, TX: CLSI Stanford.

Butt, M. (2003). "The light verb jungle," in *Papers from the GSAS/Dundley house workshop on light verbs (Harvard working papers in linguistics 9)*. eds. G. Aygen, C. Bowern and C. Quinn (Cambridge, MA: Harvard University Press), 1–50.

Butt, M. (2010). "The light verb jungle: still hacking away" in *Complex predicates: cross-linguistic perspectives on event structure*. eds. M. Amberber, B. Baker and M. Harvey (Cambridge, UK: Cambridge University Press), 48–78.

Chen, Y. (2017). Assessment standard construction in the course of consecutive interpreting for undergraduates oriented from interpreting process. *TEFLE* 144, 74–81.

Chou, W. (2019). *A comparative study of the English and Chinese Delexical construction from the perspective of cognitive linguistics*. Shanghai, China: Shanghai Foreign Language Education Press.

Claridge, C. (2000). *Multi-word verbs in early modern English: a corpus-based study*. Amsterdam: Rodopi.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. (*2nd ed.*). Hillsdale, NJ: Erlbaum.

Cover, T. M., and Thomas, J. A. (2005). *Elements of information theory*. Hoboken, NJ: Wiley.

Dixon, R. M. W. (2005). "She gave him a look, they both had a laugh and then took a stroll: give a verb, have a verb and take a verb constructions," in *A semantic approach to English grammar*. ed. R. M. W. Dixon (Oxford: Oxford University Press), 459–483.

Fan, X. (1981). *A general account of verb in Chinese*. Shanghai: Shanghai Education Press.

Feng, S. (2005). Light verb movement in modern and classical Chinese. *Linguist. Sci.* 14, 3–16.

Feng, S. (2016). *A preliminary theory of diachronic syntax in Chinese*. Shanghai: Shanghai Educational Publishing House.

Fleischhauer, J., and Neisani, M. (2020). Adverbial and attributive modification of Persian separable light verb constructions. *J. Linguistics.* 56, 45–85. doi: 10.1017/S0022226718000646

Foley, W. A. (2010). "Events and serial verb constructions" in *Complex Predicates: Cross-linguistic Perspectives on Event Structure*. eds. M. Amberber, B. J. Baker, and M. Harvey (Cambridge University Press), 79–109.

Forsyth, R. S., and Grabowski, L. (2015). Is there a formula for formulaic language? *Poznań Stud. Contemp. Linguist.* 51, 511–549. doi: 10.1515/psicl-2015-0019

Gile, Daniel (1995). *Basic concepts and models for interpreter and translator training*. Philadelphia, PA: John Benjamins.

Golshaie, R. (2016). A corpus study on identification and semantic classification of light verb constructions in Persian: the case of light verb xordan 'to eat/collide'. *Lang. Sci.* 57, 21–33. doi: 10.1016/j.langsci.2016.05.002

Gries, S. (2012). Frequencies, probabilities, and association measures in usage−/exemplar-based linguistics: some necessary clarifications. *Stud. Lang.* 36, 477–510. doi: 10.1075/sl.36.3.02gri

Grimshaw, J., and Mester, A. (1988). Light verbs and θ-marking. *Linguist. Inquiry* 19, 205–232.

Hoffmann, S., Hundt, M., and Mukherjee, J. (2011). Indian English an emerging epicentre? A pilot study on light verbs in web-derived corpora of south Asian Englishes. *Anglia J. English Philol.* 129, 258–280. doi: 10.1515/angl.2011.083

Huang, Z. (1998). *HNC (hierarchical network of concepts) theory*. Beijing: Tsinghua University Press.

Huang, C.-T. J. (2009). Lexical decomposition, silent categories, and the localizer phrase. *Essays Linguist.* 39, 86–122.

Huang, C., Lin, J., Jiang, M., and Xu, H. (2014). "Corpus-based study and identification of mandarin Chinese light verb variations," in *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*. eds. M. Zampieri, L. Tan, N. Ljubešić and J. Tiedemann (Dublin, Ireland: VarDial Workshop), 1–10.

Jespersen, O. (1954). *A modern English grammar on historical principles, part VI: Morphology*. London: Bradford and Dickens.

Kearns, K. (2002). Light verbs in English. Master's Thesis. MIT.

Kuo, P., and Ting, J. (2007). "Light verb, heavy verb and verbal noun in mandarin Chinese," in: *Proceedings of the 9th Seoul international conference on generative grammar (SICOGG 9), Seoul*, 349–357.

Langacker, R. W. (2008). *Cognitive grammar: a basic introduction*. New York: Oxford University Press.

Lu, B. (2012). The semantic characteristics of 'event nouns' in Chinese and English. *Contemp. Linguis.* 14, 1–11.

Lv, S. (1980). *Eight hundred words in modern Chinese*. Beijing: Commercial Press.

McDonald, J. L., and Carpenter, P. A. (1981). Simultaneous translation: idiom Interpreation and parsing heuristics. *J. Verbal Learn. Verbal Behav.* 20, 231–247. doi: 10.1016/S0022-5371(81)90397-2

Matsumoto, M. (1999). "Composite predicates in Middle English" in *Collocational and idiomatic aspects of composite predicates in the history of English*. eds. L. J. Brinton, and M. Akimoto (Amsterdam: John Benjamins), 59–95.

Mehl, S. (2017). Light verb semantics in the international corpus of English: onomasiological variation, identity evidence and degrees of lightness. *English Lang. Linguist.* 23, 55–80. doi: 10.1017/S1360674317000302

Montemurro, M. A., and Zanette, D. H. (2011). Universal entropy of word ordering across linguistic families. *PLoS One* 6:e19875. doi: 10.1371/journal.pone.0019875

Nagy, T. I., Rácz, A., and Vincze, V. (2019). Detecting light verb constructions across languages. *Nat. Lang. Eng.* 26, 1–30. doi: 10.1017/S1351324919000330

Paribakht, T. S., and Wesche, M. B. (1997). Reading comprehension and second language development in a comprehension based ESL programme. *TESL Canada J.* 11, 09–27. doi: 10.18806/tesl.v11i1.623

Poutsma, H. (1929). *A grammar of late modern English (part I: The sentence)*. Groningen: P. Noordhoff.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Roemer, U. (2010). Establishing the phraseological profile of a text type. The construction of meaning in academic book reviews. English text. *Construction* 3, 95–119. doi: 10.1075/etc.3.1.06rom

Ronan, P., and Schneider, G. (2015). Determining light verb constructions in contemporary British and Irish English. *Int. J. Corpus Linguist.* 20, 326–354. doi: 10.1075/ijcl.20.3.03ron

Rosen, S. T. (1989). Argument structure and complex predicates. PhD dissertation. Waltham, MA: Brandeis University.

Saito, M., and Hoshi, H. (2000). "The Japanese light verb construction and the minimalist program," in *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*. eds. R. Martin, D. Michaels and J. Uriagereka (Cambridge, MA: MIT Press), 261–295.

Sundquist, J. D. (2018). A diachronic analysis of light verb constructions in old Swedish. *J. Ger. Linguist.* 30, 260–306. doi: 10.1017/S1470542717000137

Tang, F., and Li, D. (2013). Explicitation in Chinese-English consecutive interpreting: A comparative study of professional and student interpreters. *Foreign Lang. Teach. Res.* 45, 442–452.

Tang, F., and Li, D. (2016). Explicitation patterns in English-Chinese consecutive interpreting: differences between professional and trainee interpreters. *Perspectives* 24, 235–255. doi: 10.1080/0907676X.2015.1040033

Wang, H., and Zhang, K. (2014). Light verb construction in Chinese-English translating study based on corpus--take '*jinxing*' as an example. *J. PLA Univ. Foreign Lang.* 37, 62–68+144.

Wen, B., and Cheng, J. (2007). On syntactic nature of the light verb v. *Modern Foreign Lang.* 30, 17–20.

Wierzbicka, A. (1982). Why can you have a drink but you can't have an eat? *Language* 58, 753–799. doi: 10.2307/413956

Wood, D. (2010). *Formulaic language and second language speech fluency. Background, evidence and classroom applications*. London: Continuum.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yuan, J., and Xia, Y. (1984). A general account of delexical verbs. *Second. Lang. Res.* 7, 31–40.

Zhang, Z. (2013). Syntax and semantic study of light verb. *Foreign Lang. Educ.* 34, 17–20.

Zhu, D. (1985). Delexical verbs and nominal verbs in modern written Chinese. *J. Peking Univ.* 5, 3–8.

Zhu, L. (2019). *A contrastive analysis of light verbs in English and Chinese*. Beijing: Foreign Language Teaching and Research Press.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 30 March 2023
DOI 10.3389/fpsyg.2023.1143031



OPEN ACCESS

EDITED BY
Jinfen Xu,
Huazhong University of Science and Technology, China

REVIEWED BY
Rob Kim Marjerison,
Kean University-Wenzhou, China
Shuo Lillian Yang,
Kean University-Wenzhou, China
Quanguo Liu,
Shaanxi Normal University, China

*CORRESPONDENCE
Jin Chen
✉ chenjinecho@scut.edu.cn

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

RECEIVED 12 January 2023
ACCEPTED 06 March 2023
PUBLISHED 30 March 2023

CITATION
Xu Y, Huang M, Chen J and Zhang Y (2023) Investigating a shared-dialect effect between raters and candidates in English speaking tests.
*Front. Psychol.* 14:1143031.
doi: 10.3389/fpsyg.2023.1143031



# Investigating a shared-dialect effect between raters and candidates in English speaking tests

Ying Xu, Mengjia Huang, Jin Chen* and Yaqing Zhang

School of Foreign Languages, South China University of Technology, Guangzhou, China

This study set out to examine existence of a shared-dialect effect,a phenomenon that when a rater shares the same dialect with a candidate, the rater is more likely to give the candidate a higher score in English speaking tests. Ten Cantonese-speaking raters and ten Mandarin-speaking raters were selected to assess forty Cantonese-accented and forty Mandarin-accented candidates' oral performance in the retelling task of the Computer-based English Listening and Speaking Test (CELST). Besides, seven raters from each group participated in the stimulated recall stage aiming to reveal their thought process. Quantitative results suggested that the two rater groups were comparable in terms of internal consistency. There were no significant differences in the scores of both candidate groups awarded by both rater groups. The effect of interaction between candidates' dialect and raters' dialect was not statistically significant, indicating non-existence of such effect. Qualitative results showed that some raters attended to candidates' accents, and indicated that awareness of accents and their familiarity with the accents affected their comprehension of the speech samples and potentially influenced their scoring process. The findings are discussed with reference to rater training, rating scale, raters' familiarity with candidates' accents, raters' attitudes toward candidates' accents and the task type. The main implication of this study is that recruiting both group raters in domestic English speaking tests is warranted if the shared-dialect effect could be duly managed.

KEYWORDS

**English speaking tests, shared-dialect effect, accent, stimulated recall, retelling**

## Introduction

Raters' judgment plays an indispensable part in oral performance assessments, which may be easily affected by construct-irrelevant factors (e.g., rater bias) and bring a detrimental effect on test fairness. However, there is a possibility that raters are affected by various background factors that are not involved in the rating criteria during the rating process, and those factors could introduce unwanted, construct-irrelevant variance into ratings, thus interfering with the measurement of speakers' actual performance and unrightfully contributing to the score variance (McNamara, 1996; Winke, 2012).

Those construct-irrelevant variations in scores such as bias against non-native accents (Lindemann, 2017), also termed rater effects, must therefore be minimized to avoid unduly influencing raters' scoring decisions, especially in large-scale and high-stakes speaking assessments (Winke et al., 2013; Isaacs, 2016; Kang et al., 2019a). Given the importance of validity and reliability in assessment design and development, raters are supposed to refrain from any performance-irrelevant judgments and research on the impact of rater background characteristics on their ratings have captivated many researchers' attention, as the understanding of raters' rating behaviors is conducive to elucidating "why raters assign ratings the way they do and what attributes or elements they still need to improve their rating performance" (Kim, 2015, p. 241). A large volume of research has been carried out in an attempt to investigate the influence of various rater background characteristics on raters' rating behaviors and cognitive process in speaking assessments, such as raters' linguistic background (Zhang and Elder, 2011; Gui, 2012; Wei and Llosa, 2015), rater experience (Isaacs and Thompson, 2013; Kim, 2015) and rater training (Weigle, 1998; Xi and Mollaun, 2009; Davis, 2015), but no definite conclusion can be reached due to the complex nature of raters' decision-making process. Among the listener background variables, familiarity with the accent has increasingly garnered research interest because listeners are likely to be confronted with a range of native speaker accents and diverse non-native accents in various language use contexts (Canagarajah, 2006). A considerable amount of literature in the field of speech processing and language assessment has investigated the impact of accent familiarity on listeners' perceptions and judgments (Xi and Mollaun, 2011; Winke et al., 2013; Huang et al., 2016; Park, 2020). Nonetheless, there have been few empirical investigations into the effect of accent familiarity with a certain dialect on raters.

China is a multi-ethnical country with numerous dialects. There are at least eight main dialects in China (Li, 1989), among which Cantonese and Mandarin are most widely used (Lee et al., 1996). Mandarin and Cantonese are tone languages belonging to the Sino-Tibetan language family (Yang et al., 2020). Mandarin is the official language of China with more than 1 billion speakers worldwide. Cantonese, as one of the most well-known Chinese dialects, is conservatively estimated to have over 60 million speakers in the world and is mainly spoken in Guangdong province, the Southeast region of Guangxi Province, Hongkong, and Macau (Han et al., 2015). In the field of applied linguistics, Mandarin and Cantonese are generally treated as two distinct dialects for their disparate phonological systems. For example, in Munro et al.'s (2006) study, Cantonese was treated as the phonetic counterpart of Mandarin. A more radical view even deems them as two languages with representative phonetic features which can reach the international standard of bilinguals (Xing et al., 2021). Some apparent phonological discrepancies were found between Mandarin and Cantonese in terms of tones, vowel and consonant sounds, as well as accents, evidenced by the fact that there are five tones in Mandarin whereas Cantonese has practically nine tones. When it comes to the phonetic symbol system, the Mandarin vowel system comprises nine monophthongs, nine diphthongs, and four triphthongs, while Cantonese has 11 monophthongs and 11 diphthongs in its vowel system and 19 initial consonants, and six final consonants in its consonant system (Law and So, 2006). Cantonese-accented English is characterized by some typical pronunciation errors, such as sounds' swallowing (/d/,/t/,/k/,/z/), sounds' addition, phoneme error, word or chunk error and word stress errors (Xu and Zeng, 2015). Nevertheless, the following acoustic properties such as the devoicing of a word-final stop consonant (Hayes-Harb et al., 2008), the mispronunciation of/ae/,/ɛ/,/ʌ/, and the improperly perceptual distance of tense and lax vowels (Barkana and Patel, 2020) are labeled Mandarin-accented English. Winke et al. (2013) claimed that Cantonese and Mandarin as different language varieties might affect raters' judgment toward speakers' oral performance. Therefore, of interest to this study is whether raters assign scores to examinees' performance differentially as a result of different dialects spoken by the rater and the candidate.

In 2011, the National Matriculation English Test (NMET) of Guangdong province incorporated a separate component of oral test, the Computerized English Listening and Speaking Test (CELST), which purports to gauge candidates' oral English ability. The annual number of candidates for the CELST in Guangdong NMET amounts to over 800,000 and they are from all over the province (Zhang, 2014). Speech produced by these candidates would inevitably be stamped with their native dialects, and human raters invited to assess candidates' oral performance may also be affected by the dialects they routinely speak. Thus, there is an urgent need to scrutinize whether there exists a shared-dialect effect, analogous to a shared-L1 effect (Harding, 2012), between raters and examinees in oral English proficiency tests, and whether raters are aware of any influences of such effect on their scoring.

To the best of the authors' knowledge, any research that has specifically explored whether raters' background difference in dialects could cause noticeable differences in their ratings or make a difference to their decision-making process in English speaking tests is yet to be carried out in the Chinese context, which is demonstrably a crucial question requiring an immediate answer. Therefore, it is urgent to inquire existence of a shared-dialect effect and to elucidate whether certain candidates are thus advantaged or disadvantaged. The research results could advance our knowledge of such effect and shed light on the recruitment of raters with regional dialects in English speaking tests.

## Literature review

### Empirical studies on the shared-L1 effect

In the field of language testing, the term shared-L1 effect refers to the phenomenon that a group of candidates who share the same L1 with the speaker of the test recording can find the listening materials more comprehensible and give a better performance on the test (Harding, 2012; Dai and Roever, 2019), which is analogous to the *interlanguage speech intelligibility benefit* (ISIB) in the speech processing literature (Bent and Bradlow, 2003). A probable explanation for this phenomenon might be that being exposed to a speaker's accent repeatedly contributes to familiarity with that accent, which in turn facilitates comprehension of the speaker (Bradlow and Bent, 2008; Stevenage et al., 2012). Although the possibility of a shared-L1 advantage and the potential for a shared-L1 effect have aroused considerable interest in areas

including speech perception, L2 listening comprehension and language testing, current research has yielded mixed results.

The term ISIB refers to the benefit of a shared language background between non-native listeners and speakers. It was proved by Bent and Bradlow (2003) which performed perception tests on native speakers of Chinese, Korean, English, and other language backgrounds, asking them to listen to sentences read in English with Chinese, Korean, and English accents. Results indicated that native English listeners had higher word recognition rates for sentences spoken by native than non-native speakers. However, the non-native listeners found high-proficient non-native speakers of the same L1 equally as intelligible as the native English speakers. Interestingly, there seems to be an assumption that the L2 proficiency of listeners and speakers may play a role in modulating the ISIB (Xie and Fowler, 2013). A few studies suggested that the interlanguage speech intelligibility benefit existed more in low-proficiency learners. For example, in a follow-up study of Bent and Bradlow (2003), Stibbard and Lee (2006) reported that there were no significant differences in intelligibility scores for high proficiency non-native speakers and native speakers within each listener group. Native speakers were not more intelligible than non-native speakers even to their fellow native listeners. The non-native listeners found that the non-native low-proficiency speakers who shared their own first language was not as significantly unintelligible as those who did not share the first language, which suggested that the shared-L1 effect may only be taking hold when listeners heard lower-proficiency speakers.

Additionally, from the perspective of language features, the sheer volume of studies provides partial or little evidence to support a shared-L1 advantage phenomenon. Some studies such as Harding (2012), Dai and Roever (2019) found positive evidence of the shared-L1 effect in Mandarin-L1 candidate groups through the comparison of candidates' performance in English listening tests conducted with various accents. The Mandarin Chinese-L1 listeners were found distinctly advantaged on several test items featuring a speaker with Mandarin Chinese accent. However, studies conducted by Major et al. (2002), Kang et al. (2019b) failed to support the shared-L1 effect argument for the Chinese-L1 listeners scored significantly lower than other listener groups when listening to passages recorded by speakers who shared their native language. Other studies, such as Munro et al. (2006) have shown the facilitative effect of L1 accent on the Japanese listeners group on account of the fact that researchers found speeches produced by speakers of their own language background were easier to understand than speeches by Cantonese, Polish and Spanish speakers. Nonetheless, in Harding's (2012) study, the effect of shared-L1 was not clearly observed when investigating a shared-L1 advantage to the Japanese. The mixed findings grounded in the above two languages could not offer full support for the existence of the shared-L1 effect. Regarding other languages that have been investigated, Abeywickrama (2013) found no evidence of a shared-L1 effect by the measurement of three other language groups' (Korea, Sri Lanka, and Brazil) comprehension of shared-L1 accent speech *via* a multiple-choice (MC) TOEFL listening test whose speech stimulus were recorded by speakers with Chinese, Korean, Sri Lankan, and American accent. Test-takers' comprehension scores on the MC listening assessment were not significantly affected by speakers' accents and they had comparable performance even when the input was delivered by speakers who shared the same native languages, suggesting that there is no shared-L1 effect. Besides, the shared-L1 advantage has not yet been found in the French-language background. For example, Crowther et al. (2016) examined how listeners' status (native, non-native) and language background (French) influenced the raters' (French, Mandarin) L2 comprehensibility and accentedness. Analyses of the global ratings demonstrated that when rating the L2 speakers from the French-language background, the French listener group did not benefit from the shared language background compared to the Mandarin listener group, contradicting the shared-L1 advantage.

These mixed findings reported on various languages have shown the indeterminacy of the existence of the shared-L1 effect, which suggests that the effect is not consistent across language variables. The question of whether shared-L1 could impact candidates' performance in listening tests still remains unknown. More importantly, prior studies predominantly focused on the language of a certain country without the consideration of its regional varieties' effect on research findings (Winke et al., 2013). In China, Mandarin Chinese speakers might put on diverse local accents across the country. For fairness reasons and positive washback of language tests, there is a necessity to examine the effect of Chinese dialects on the interactions between listeners and speakers under Chinese dialect cultural contexts.

## The accent familiarity's effect on raters

Familiarity with a particular accent is conducive to understanding that type of accented speech (Gass and Varonis, 1984; Tauroza and Luk, 1997; Major et al., 2002; Dai and Roever, 2019). To date, several studies have examined the influence of accent familiarity of certain languages on raters' rating process and behaviors (Carey et al., 2011; Huang, 2013; Winke et al., 2013; Park, 2020). Results of following studies suggested that raters' familiarity with examinees' accents affects the rating of pronunciation and general speaking ability. For example, Carey et al. (2011) demonstrated that raters who were familiar with the candidates' accent were more likely to assign favorable higher pronunciation scores than raters who had little or no familiarity with that accent, and they also tended to score candidates from their own home country higher than candidates from a different country. Their findings were similar to Winke and Gass (2013) which delved into raters' cognitive process through collecting raters' (Spanish, Chinese, and Korean L2 learners) comments while rating three groups of examinees from Spanish, Chinese, or Korean L1 backgrounds in a qualitative study. Analyses of raters' comments revealed that heritage language speakers had unconscious biases in rating familiar accented speech samples. This result supported the notion that raters' language backgrounds, in particular heritage language backgrounds, could influence their rating decisions. It demonstrated the potential bias of accent familiarity on raters' scoring and also provided evidence supporting Winke et al.'s (2013) hypothesis that accent familiarity could potentially lead to bias, including rating reliability, though the effect may be limited and inconsistent. However, such a clear pattern was not observed by Park (2020) which found that ratings across three teacher groups with different degrees of familiarity with Korean accent (heritage, familiar, and unfamiliar) on the assessment of Korean-accented

English exhibited high interrater reliability, and prior exposure to foreign-accented speech affected their consistency in ratings. In the meanwhile, by comparing the severity of the three groups, the researcher found that non-heritage teachers were less tolerant than heritage teachers in assessing global proficiency and accentedness, even though there was no significant difference in the level of severity between the familiar and unfamiliar teacher groups.

The causes of inconsistent results are likely in part due to different methodological perspectives, the tools used to measure familiarity, and raters' varying perceptions of interlingual and intralingual accents. It should be noted that the first three studies in this section examined raters' familiarity effect on the assessment of L2 pronunciation by comparing rater performance while assessing speech samples with different accents, but these studies did not strictly control raters' familiarity with every accent. Although Park (2020) investigated the familiarity's effect with a simple accent from different levels, the current literature is still limited and further research should give clear evidence to illustrate the familiarity effect on rater bias.

## Raters' perception of candidates' accent

Contrary to the significant effect of accent familiarity manifested in the reviewed studies, other studies failed to detect that effect (Xi and Mollaun, 2009, 2011; Huang, 2013; Wei and Llosa, 2015; Huang et al., 2016) under various conditions. However, it does not mean that raters' decision-making would not be affected by other mediating variables. It's still necessary to further explore the potential effect of raters' complicated psychological course ensconced in digital signals transmitted by scores. The deep-going comportment can reveal the possible factors that would lead to raters' differential assessment decisions with different accents and provide insights into raters' views of the practicality of including non-native accents in English speaking tests.

Xi and Mollaun (2009, 2011) compared the ratings of the TOEFL iBT Speaking test assigned by trained (including how to score English speech samples from native-Indian speakers) and untrained bilingual/multilingual Indian raters. Even though they did not find a significant difference between the numerical ratings, they discovered positive effects from undergoing the training, which helped trained raters guard against what they claimed to be an internal dilemma when rating speakers of familiar accents. Huang's (2013) findings were consistent with those of Xi and Mollaun (2009, 2011), showing no significant differences between the ratings of the three rater groups. With a focus on raters sharing the same L1, Huang (2013) from the angle of teaching experience as well as accent familiarity, investigated the two-fold effect on raters' self-perception. Three groups of raters who varied on familiarity with non-native accents and language teaching experience were recruited to evaluate speech samples spoken by native Chinese speakers on both holistic and analytical dimensions. Results revealed that the speakers' accent together with teaching experience might lead to the potential leniency effect. Given that raters' bias related to the two factors' combining effect, it was unclear whether accent familiarity alone could give rise to rating bias. In the subsequent study of Huang et al. (2016), they only investigated the influence of raters' familiarity with accents on their rating decisions. Three groups of raters with different backgrounds (Spanish Heritage, Spanish Non-Heritage, and Chinese Heritage) rated 28 speech samples on the overall English proficiency and foreign accents. Raters self-reported that their accent familiarity affected their evaluations of accentedness, and might have made them more lenient toward speakers with familiar accents. Besides, they expressed a strong preference for Spanish accents. Results clearly demonstrated that being familiar with a certain type of foreign accent facilitated the identification of that accent and also revealed that more favorable accents in their study were those prevalent in the language speaking country, suggesting that positive contexts of familiarity would lead to positive bias and vice versa (Cargile, 1997; Lindemann, 2005). Similar findings from the quantitative view were also obtained in Wei and Llosa (2015), which examined whether American and Indian raters differed in their scores and scoring process with Indian test-takers' speech samples from the TOEFL iBT speaking tasks. No statistically significant differences were found between Indian and American raters in their use of the scoring criteria, their attitudes toward Indian English, and the internal consistency and severity of the scores. However, in-depth qualitative analysis revealed that some Indian raters even held negative attitudes toward Indian English. The findings of this study manifested that sharing a common language background does not guarantee a positive evaluation of candidates' L2 speaking performance after all.

The inconsistent findings of quantitative and qualitative methods are unsurprising because of the complex development trajectories of cognitive processing in raters from various backgrounds. This mismatch also indicated that rater bias was not fully uncovered in reviewed studies or, alternatively, raters' mental process was not precisely captured for the methodological gap.

In summary, the aforementioned studies have produced somewhat inconclusive results regarding the shared-L1 effect based on different language backgrounds of listeners and speakers, and existence or strength of such effect has not been fully investigated. Scant literature has been found focusing on the impact of sharing the same dialect between raters and candidates in English speaking tests, let alone the potentiality for a shared-dialect effect in the Chinese context. Although the role of raters' accent familiarity of certain languages in speaking assessments has received increasing attention, more empirical research is needed to further probe the effect of accent familiarity of dialects subsumed under one certain language on rating performance and cognition.

## Research questions

The overarching goal of this study was to explore the potential for a shared-dialect effect in English speaking tests in the Chinese context and to investigate whether raters were aware that the shared dialect between raters and candidates may have an influence on their judgment of oral performance. The present study was guided by the following two research questions:

(1) Are there any significant differences in the scores given by Cantonese-speaking and Mandarin-speaking rater groups to the Mandarin candidate group and the Cantonese candidate group on the Retelling task in the CELST?

(2) Are trained raters aware that the shared dialect with the candidates might impact their ratings on the Retelling task in the CELST?

# Methods

## Participants

Thirty-eight postgraduates from different universities in China were recruited as raters. Graduate students were selected because participants from a more diverse population would introduce far more variables (Winke et al., 2013). One half of them were heritage speakers of Mandarin and the other half Cantonese heritage speakers, meaning that they were immersed in the language environment where their family members spoke that language natively and were identified with one particular ethnic group by it. They were all female aging from 22 to 26, none of whom had hearing or speech disorders. All of them had been learning English as a foreign language in China for at least 12 years. Following the six steps training approach proposed by Bachman and Palmer (1996, p. 222) and the calibration standard suggested by Hoskens and Wilson (2001), only 20 raters were accredited and they were classified into two background groups according to their dialects: Group A (including ten Mandarin-speaking raters) and Group B (including ten Cantonese-speaking raters). No participant reported speaking any languages or dialects other than Cantonese, Mandarin and English.

To avoid any influence caused by background variables, the participants were selected on the basis of homogeneity of their educational background, language proficiency and rating experience. First, all raters were postgraduate students studying in the field of applied linguistics. Second, they had all passed Test for English Majors for Grade 8 (TEM8) (Jin and Fan, 2011). TEM8 is a large scale and high-stakes criterion-referenced English test, designed to assess undergraduate English majors' language proficiency at the end of their four years professional learning program (Zou and Xu, 2016), to check whether test-takers' language knowledge and capacities could meet the learning requirements documented in the *Syllabus for Test for English Majors (Grade 8)* (National Advisory Committee for Foreign Language Teaching, 2004). Last, they all had no prior rating experience of any oral assessments. They were informed that they should attend both the training stage and the rating stage. Besides, 14 raters (seven raters from each group) were invited to take part in the stimulated recall stage based on their availability. All participants received certain monetary rewards for their participation. Raters' general background information was collected with an online background questionnaire before training, which would be introduced in the forthcoming section. Some detailed background information of raters is shown in **Table 1**.

An independent *t*-test performed on the familiarity with Cantonese showed that there was a significant difference between the two groups ($t = -13.887$, df = 18, $p = 0.00$). No significant statistical difference was found in the means of age ($t = -0.590$, df = 18, $p = 0.56$) and years of learning English ($t = -0.557$, df = 18, $p = 0.58$). Only two Mandarin-speaking raters reported that they were neutral about Cantonese, but the rest of the raters held a positive attitude toward Cantonese.

## Instruments

### Speech samples

The speech samples for the present study were candidates' performances on the Retelling task in the CELST in Guangdong NMET in 2013 (**Appendix A**). Four subsets of samples in different numbers were purposefully chosen from a pool of sound files by 32 listener judges who were enrolled in a MA programs at a University in Guangdong, with either Cantonese or Mandarin background. In order to strictly control all the speech samples to have a similar degree of accent strength and identifiability, 32 recruited listener judges were required to evaluate the above two mentioned indexes of the provided speech samples with a Strength and Identifiability of Accent Scale (**Appendix B**), which was designed drawing on Ockey and French's (2016) accent scale and the accent strength and identification task used in Dai and Roever (2019). To guarantee reliable accent strength and typicality, only judges who claimed high familiarity with the two dialects in the evaluation process and reported to be apt at dialect judgment and identification were selected. At last, 96 valid speech samples were included and classified into four subsets in the formal experiment.

Subset 1 included four benchmark samples, used as exemplars of each score band of the rating scale, representing a range of proficiency levels and performance types. Subset 2 contained 12 practice samples rated by two expert raters (who were professors of applied linguistics and had more than eight years rating experience of CELST) and used in training. Subset 3 consisted of 80 formal rating samples utilized in the formal rating. Subset 4 comprised four Cantonese-accented speech samples and four Mandarin-accented speech samples, which were purposefully picked out from Subset 3 and used as the prompts in the stimulated recall stage. All raters rated and commented on the same set of speech samples in the experiment. The formal rating samples were counterbalanced in terms of candidates' dialect (Mandarin, Cantonese), gender, and official NMET scores of the Retelling task. Candidates were evenly divided into two groups based on their dialects. Each dialect group had 20 male and 20 female candidates. Ten candidates (five males and five females) were at each of two levels of proficiency (high and low) within each dialect group. The two levels of proficiency were assigned according to candidates' NMET scores of the retelling task. As the maximum score of the task is 24, candidates who received a score higher than 18 were labeled high-level, and those scored lower than 12 low-level. Samples with the same dialect, of the same gender and of the same proficiency did not occur adjacently.

### Background questionnaire

At the beginning of the study, all participants completed a background questionnaire online (**Appendix C**) to obtain participants' demographic information and to explore their language background. By adapting the questionnaire from Wei and Llosa (2015), questions in this instrument aim to solicit information concerning raters' age, gender, dialect, English learning experience and proficiency levels, exposure to Cantonese, rating experience and academic background. In addition, participants' familiarity with Cantonese was also gauged on a Likert-scale ranging from 1 (strongly unfamiliar) to 5 (strongly familiar) after listening to two pieces of speech materials with typical pronunciation characteristics of the Cantonese accent. Similarly, a 1–5 Likert scale (1 = strongly dislike; 5 = strongly like) was employed to investigate

participants' attitudes toward Cantonese. The reason why not tap into raters' familiarity with and attitudes toward Mandarin is that it is the official national language and has been popularized in China for several decades, hence generally Chinese people are much familiar with it and hold a positive view on it.

### The retelling task

The retelling task in the CELST is designed to measure candidates' integrated listening and speaking ability, especially the ability to obtain information from listening materials and to process and reconstruct information. Candidates are required to listen to a 2 min story. While listening, they are presented with a one-sentence hint of the story. The story will be played twice. Candidates are allowed to take notes while listening. After listening and 1 min preparation, they should retell the story by using proper words and sentences within 1 min. The retelling content should cover as much information of the story as possible. The entire process of completing a retelling task in the CELST lasts approximately 6 min.

The present study only concentrated on the retelling task because as a typical integrated task, retelling could reflect candidates' use of second language in the real-life situations and measure candidates' speaking ability validly, thus has been widely used in L2 oral performance assessments (Frost et al., 2012).

### Rating criteria

Rather than using the official rating scale of retelling in the Guangdong NMET (**Appendix D**), the modified version of the rating scale of TEM4 (Test for English Majors for Grade 4) story retelling task developed by Liu (2013) (**Appendix E**) was employed, because the official rating scale (including two dimensions: Content and Holistic) does not require raters provide any score on candidates' pronunciation, which is a major concern of the present study. Instead, Liu's (2013) version is an analytic rating scale, containing four conceptual dimensions: Grammar, vocabulary and expression; Retelling content; Pronunciation and intonation; and Fluency. There are detailed descriptions of four different levels in each dimension. The full mark is 16 points because each dimension spans score bands of 1 (lowest) to 4 (highest) corresponding to the different levels.

### Stimulated recall

Stimulated recall was conducted to trace raters' individual thinking process in assigning scores. As an introspective method, this type of verbal reporting is conducive to probing into the complex nature of the scoring process by providing raters with recently recorded stimulus or cues (Gass and Mackey, 2000). It is generally applied in studies of rater performance in speaking tests (Winke et al., 2013). By replaying the tape-recording or fragments of the recording, stimulated recall could prompt raters to recall and verbalize their concurrent cognitive activity when performing the scoring task.

## Procedures

This study included three stages: training, rating, and stimulated recall. Due to the outbreak of *COVID-19* pandemic, training, and stimulated recall were carried out on *Tencent Meeting*, an application allowing users to attend a real-time interactive online meeting. **Figure 1** illustrates the details.

Before starting the actual rating, each participant should undertake training. At this stage, 38 raters were provided with the original script and the recording file of the story-retelling, the rating criteria, benchmark samples, and practice samples. Firstly, all raters were allocated enough time to familiarize themselves with the task and the rating criteria. Secondly, one researcher introduced the story-retelling task, and then illuminated the rating criteria in detail and explained the rationale for assigning a specific score for speech samples at each score level with the benchmark samples. Raters could ask any questions concerning the rating scale in order to internalize it with the help of benchmark samples. Next, each rater was asked to rate the set of 12 practice samples and provide legitimate reasons for their ratings individually. Later, they compared their rating results and reasons with the scores and rationales provided by the two expert raters. In an attempt to simultaneously guarantee the reliability and validity of the formal rating, only 20 raters who not only reached 80% consensus with the agreed-upon scoring outcomes but also correctly interpreted the rating scale reflected by their reasons for ratings were accredited (Hoskens and Wilson, 2001; Elder et al., 2007; Xi and Mollaun, 2009). Last, seven raters in each group were informed of the procedure for stimulated recall and trained to verbalize their thoughts. After practicing with a sample recording, no participant reported difficulty in verbal reporting.

At the formal rating stage, raters were allowed to complete rating independently with the rating scales at their convenience within a certain time on the computer. They were allowed to listen to the speech samples for more than one time if necessary.

When each of 14 selected raters completed the rating tasks, she was arranged to converse immediately with one researcher to undergo the stimulated recall stage individually. At this stage, the recordings of four Cantonese-accented and four Mandarin-accented speech samples were replayed twice to raters by the researcher. They should award a score to the current candidate's performance in the same way of the rating stage and state the reasons for that score after the first replay. Then after listening for a second time, they were encouraged to recall what they had been thinking about at the time of rating and speak out what came to their minds immediately as much as possible. Leading questions were shown below: (1) What were you thinking about when scoring? (2) What were you thinking when listening to the speech sample? (3) Did you find this sample difficult to understand? Why? What factors affected your understanding? Any further ideas or comments if raters wished to elaborate were welcomed. Raters were free to choose any language to verbalize their thoughts, so that they could express their ideas fluently and clearly. The whole procedure was audio recorded, lasting for approximately 60 min.

## Data analysis

Quantitative and qualitative data have been gathered for the present study. Quantitative data consisted of 1,600 valid ratings that were assigned by 20 raters to 80 speech samples, and qualitative data included the voice recordings of the stimulated recall. A combination of quantitative and qualitative methods was used in the data analysis.

TABLE 1 Raters' background information.

| Group | Dialect | Rater | Age | Years of learning English | Familiarity with Cantonese | Attitude toward Cantonese |
|---|---|---|---|---|---|---|
| A | Mandarin | A1 | 23 | 14 | 2 | 5 |
| | | A2 | 25 | 13 | 2 | 3 |
| | | A3 | 24 | 15 | 1 | 4 |
| | | A4 | 24 | 14 | 2 | 4 |
| | | A5 | 24 | 13 | 2 | 4 |
| | | A6 | 25 | 14 | 1 | 3 |
| | | A7 | 25 | 16 | 2 | 4 |
| | | A8 | 26 | 16 | 2 | 4 |
| | | A9 | 23 | 15 | 1 | 5 |
| | | A10 | 22 | 12 | 2 | 4 |
| B | Cantonese | B1 | 24 | 15 | 5 | 5 |
| | | B2 | 25 | 16 | 5 | 5 |
| | | B3 | 26 | 14 | 5 | 4 |
| | | B4 | 24 | 15 | 4 | 5 |
| | | B5 | 24 | 14 | 5 | 4 |
| | | B6 | 23 | 13 | 5 | 5 |
| | | B7 | 23 | 13 | 4 | 5 |
| | | B8 | 24 | 15 | 5 | 5 |
| | | B9 | 26 | 16 | 5 | 5 |
| | | B10 | 25 | 14 | 4 | 4 |

The referential meanings of the last two columns' numbers are further elucidated in the background questionnaire section.

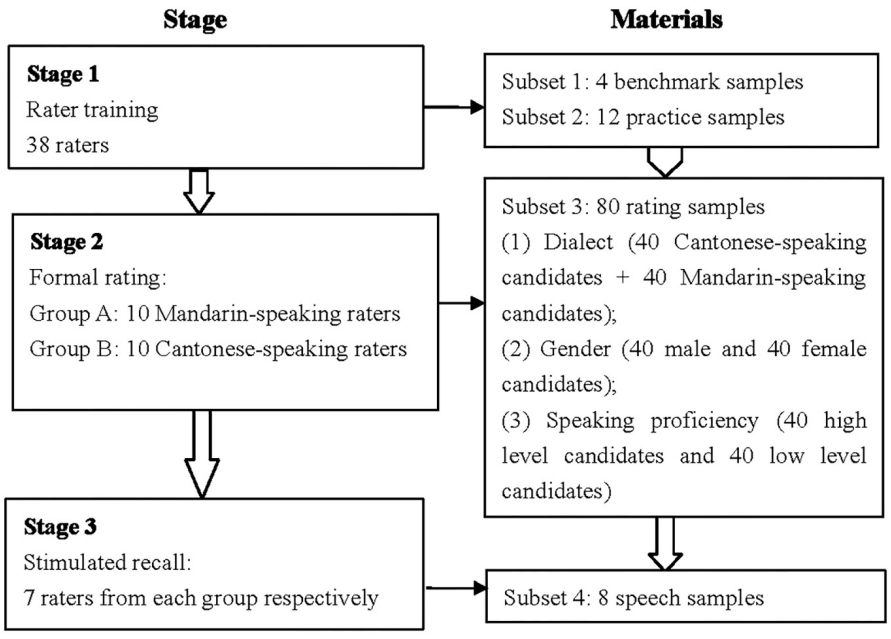



FIGURE 1
The procedure.

To answer RQ 1, raw data was collected by using Microsoft Excel spreadsheets. The interrater reliability statistics were calculated at first and the Two-Way factorial analysis of variance (ANOVA) was performed to examine whether there were any significant differences in the scores of two candidate groups awarded by two rater groups through the software SPSS 19.0.

To address RQ 2, an analytic inductive approach (Thomas, 2006) was adopted to analyze qualitative data collected through stimulated recall. Themes and patterns were expected to emerge from the data. Verbal reports were analyzed in four steps. Firstly, all stimulated recall audio recordings were transcribed verbatim by a research assistant majoring in language testing and double-checked by one researcher. The essential principle of the transcription was faithfulness. In the transcribing process, the playing of recordings of speech samples was omitted and all spoken information on the recordings from the scoring and reporting sessions should be written down as much as possible. Any pause longer than 3 s was marked by a "...". Then the transcribed texts of each rater were entered into the qualitative data analysis software QSR NVivo 11.0. Next, the research assistant segmented those transcripts into idea units (Green, 1998) independently, which were double-checked by the researcher. For the sake of coding reliability, the researcher and the research assistant reread all the idea units and coded them into various themes independently. Finally, they discussed and agreed on specific names for, and operationalization of prominent themes. Through discussion, a consensus was reached on coding.

# Results

## Findings of RQ 1

### Rater reliability and descriptive statistics

An internal consistency was examined by means of reliability analysis. Results showed that Cronbach's coefficient alpha was 0.961 and 0.988 for the scores awarded by Cantonese-speaking and Mandarin-speaking raters, respectively. The reliability statistics indicated that the two rater groups exhibited high internal consistency. Furthermore, descriptive statistics for 80 candidates' scores assigned by the two rater groups were reported in **Table 2**.

It can be seen that for the four dimensions and the total score, the means of scores given by the Cantonese-speaking raters were either slightly higher or lower than those by the Mandarin-speaking raters, yet no statistically significant difference was observed ($p = 0.171 \sim 0.809$).

## ANOVA results

A mixed two-factor ANOVA with raters' dialect (the between-subjects factor) and candidates' dialect (the within-subjects factor) as independent variables and the total score as the dependent variable was initially run to test whether differences in ratings across the two groups were statistically meaningful. It was found that there was non-significant difference for candidates' dialect [$F(1, 1) = 2.078$, $df = 1$, $p = 0.15$] and raters' dialect [$F(1, 1) = 0.059$, $df = 1$, $p = 0.81$]. The interaction effect between raters' dialect and candidates' dialect [$F(1, 1) = 0.000$, $df = 1$, $p = 0.99$] was not statistically significant either.

In order to test whether differences in ratings along four dimensions across the two groups were meaningful, two-factor ANOVA was used four times. Due to multiple comparisons being made, a Bonferroni correction was applied to the $p$-value with a new threshold of 0.0125 set.

First, using scores in *Grammar, vocabulary, and expression* as the dependent variable, it was found that there was no statistically significant main effect of candidates' dialect [$F(1, 1) = 2.397$, $df = 1$, $p = 0.12$], no statistically significant main effect of raters' dialect [$F(1, 1) = 0.099$, $df = 1$, $p = 0.75$], and no statistically significant interaction effect between candidates' dialect and raters' dialect [$F(1, 1) = 0.013$, $df = 1$, $p = 0.91$]. Second, as for the scores in *Retelling content*, the main effect of candidates' dialect [$F(1, 1) = 0.424$, $df = 1$, $p = 0.52$], the main effect of raters' dialect [$F(1, 1) = 1.817$, $df = 1$, $p = 0.17$], and the interaction between candidates' dialect and raters' dialect [$F(1, 1) = 0.004$, $df = 1$, $p = 0.95$] were all not statistically significant. Third, with scores in *Pronunciation and intonation* as the dependent variable, the results of ANOVA showed that the main effect of raters' dialect [$F(1, 1) = 0.523$, $df = 1$, $p = 0.47$], the main effect of candidates' dialect [$F(1, 1) = 5.562$, $df = 1$, $p = 0.02$], and the interaction of the candidates' and raters' dialect [$F(1, 1) = 0.007$, $df = 1$, $p = 0.93$] were all not statistically significant. Last, using scores in *Fluency* as the dependent variable, neither the between-group main effect [$F(1, 1) = 1.484$, $df = 1$, $p = 0.23$], the within-group main effect [$F(1, 1) = 0.089$, $df = 1$, $p = 0.77$], nor the interaction effect of raters' dialect × candidates' dialect [$F(1, 1) = 0.053$, $df = 1$, $p = 0.82$] was found statistically significant.

To summarize, the quantitative analysis indicated that neither the main effect of raters' dialect nor the interaction effect was significant when candidates' total scores and scores on each dimension were employed as the dependent variable.

## Findings of RQ 2

The current study adopted Winke et al.'s (2013) coding scheme and made some necessary adjustments on coding categories for some new features. Comments were drawn from both rater groups for all eight speech samples. One researcher and the assistant read those comments carefully and coded independently. The coding work was done using QSR NVivo 11.0. The initial intercoder agreement reached approximately 87%. For those incongruences between the coding results, the researcher and the assistant discussed thoroughly. Eventually, a perfect agreement between two coders was achieved. Seven broad themes were identified from analysis of raters' comments, including (1) candidates' accent; (2) candidates' heritage status; (3) raters' scoring strategy; (4) comments on pronunciation; (5) affect; (6) candidates' voice; (7) candidates' intonation.

To elaborate, the first theme was raters' comments on candidates' accents, which were further broken down into positive and negative comments. Comments such as "the accent was great" and "it did not impact understanding" were coded as positive. Comments such as "the accent was a bit problematic," "... made it difficult to score," and "... left me an awful impression of the candidate" were categorized as negative. Some other references only mentioned the candidates' accents but without further comment were coded as neutral. Theme two (candidate's heritage status) was relative to comments of guessing where candidates may come from. The third theme showed raters' scoring strategy of paying attention to candidates' pronunciation in the first place. Different from comments on candidates' accents, raters' positive or negative

TABLE 2 Two-factor ANOVA results for within group analysis.

| | Cantonese-speaking rater group | | | | Mandarin-speaking rater group | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dimension | *N* | M | SD | SE | *N* | M | SD | SE | F | Sig. |
| Grammar, vocabulary and expression | 10 | 2.81 | 0.72 | 0.08 | 10 | 2.78 | 0.69 | 0.08 | 0.099 | 0.75 |
| Retelling content | 10 | 2.87 | 0.85 | 0.09 | 10 | 2.69 | 0.82 | 0.09 | 1.314 | 0.17 |
| Pronunciation and intonation | 10 | 2.86 | 0.66 | 0.07 | 10 | 2.93 | 0.69 | 0.08 | 0.523 | 0.47 |
| Fluency | 10 | 2.86 | 0.70 | 0.08 | 10 | 2.90 | 0.68 | 0.07 | 0.089 | 0.77 |
| Total score | 10 | 11.40 | 2.83 | 0.32 | 10 | 11.30 | 2.79 | 0.31 | 0.059 | 0.81 |

*N*, number; M, mean scores; SD, standard deviations.

TABLE 3 Summary of coding themes.

| Coding category | Raters | Cantonese-speaking raters | Mandarin-speaking raters | References | Words |
|---|---|---|---|---|---|
| 1. Candidate's accent | 9 | 4 | 5 | 40 | 2,394 |
| 1.1. Positive | 5 | 2 | 3 | 12 | 571 |
| 1.2. Negative | 5 | 2 | 3 | 17 | 1,174 |
| 1.3. Neutral | 5 | 3 | 2 | 11 | 649 |
| 2. Candidate's heritage status | 7 | 4 | 3 | 14 | 1,209 |
| 3. Rater's scoring strategy | 8 | 3 | 5 | 12 | 622 |
| 4. Comments on pronunciation | 7 | 5 | 2 | 20 | 679 |
| 4.1. Positive | 7 | 4 | 3 | 14 | 408 |
| 4.2. Negative | 4 | 2 | 2 | 6 | 271 |
| 5. Affect | 6 | 3 | 3 | 6 | 223 |
| 6. Candidate's voice | 5 | 1 | 4 | 6 | 256 |
| 7. Candidate's intonation | 4 | 4 | 0 | 7 | 496 |

comments on candidates' pronunciation were coded into the fourth coding category. The coding theme of affect related to how rater felt while listening and rating. The ultimate two coding categories were germane to the candidates' voice and intonation, respectively.

The seven coding themes, the number of raters (including the number of raters from two dialect groups) who made comments associated with the theme, the frequency of references connected to the theme, and the entire numbers of words used in discussing the theme were displayed in Table 3.

The following subsections would probe into three major coding themes relevant to the relationship between raters' and candidates' dialects, including (1) candidates' accent; (2) candidates' heritage status; (3) raters' scoring strategy.

## The candidates' accent

Nine of the fourteen raters reported that they noticed or made further comments on the candidates' accents while listening and rating. Five raters expressed a positive attitude toward accents. Five raters conveyed negative feelings toward the issue of accent. They commented that accents affected their comprehension of the samples and probably influenced their rating decision. Five raters said that they noticed the accents in the candidates' speech, but did not comment on this issue.

Two Cantonese-speaking raters and three Mandarin-speaking raters held a positive view of accent. They indicated that having an accent did not matter a lot as long as it did not interfere with understanding, as demonstrated in (1) and (2).

> [1] *I noticed that he had a strong accent, the Cantonese accent. But I don't think it mattered, as long as it did not challenge my understanding* (B8, Cantonese-speaking).
>
> [2] *When the speaker started to talk, I could easily identify her Mandarin accent. Compared with the last speaker (a Cantonese-accented speaker), I felt more comfortable with her accent* (A9, Mandarin-speaking).

Interestingly, a Cantonese-speaking rater (B6) and her Mandarin-speaking counterpart (A8) seemed to display a feeling of positive bias for familiar accents. They all noted that candidates' speeches as a whole were not extremely difficult to understand because they were familiar with candidates' accents. And due to this familiarity, they became tolerant of various difficulties during the rating process, as shown in (3) and (4).

> [3] *This speech sample as a whole was slightly difficult to understand. Although I have read the script, I don't know why he mentioned the words like "garden" and "milk", so I couldn't*

> *understand what he was saying. But his accent was fine to me, because basically I'm familiar with it* (B6, Cantonese-speaking).
>
> [4] *The candidate spoke slowly and his pronunciation was not as natural as native speakers. When I listened to this sample, the candidate's accent reminded me of my English teacher's speech pattern. Her pronunciation was friendly to me and helped me understand what she was saying* (A8, Mandarin-speaking).

In contrast, two Cantonese-speaking raters and three Mandarin-speaking raters held a negative view on candidates' accents. Seventeen comments were coded into the negative category. Raters expressed a general concern for candidates' pronunciation with accents and conveyed dreadful feelings. As illustrated in the following comments, they reported that the annoying accent made them feel uncomfortable, unpleasant and perplexed, leading to unfavorable impressions. They also noted that accents affected their listening process and impeded comprehension. A representative example can be found in the following comment by A3. She stated that the unfamiliar strong accent made it difficult for her to understand the candidate and thus influenced her rating. Ultimately, she only assigned a passing score because of the strong accent.

> [5] *In terms of pronunciation, his heavy accent and dialect gave rise to all sorts of difficulties. It can significantly affect my understanding, so he only got a passing mark* (A3, Mandarin-speaking).

Five raters commented that candidates' pronunciation was not accurate and excellent. They noticed an issue of accents occurring in candidates' performances. However, the Cantonese-speaking raters were more likely than the Mandarin-speaking raters to recognize the Cantonese accent and Mandarin-speaking raters were better at identifying Mandarin accent. A pair of examples can be found in (6) and (7).

> [6] *What I hear is that he had a Cantonese accent regarding pronunciation. One thing by the way, I think most speech samples have rhymes. I felt that the feature of rhymes at the end of every sentence or word was like a Cantonese accent* (B7, Cantonese-speaking).
>
> [7] *The speaker's Mandarin accent was not native-like, because his pronunciation was not very good, the intonation was basically flat, and his pronunciation seemed to have a strong accent, and some words were not accurately pronounced* (A3, Mandarin-speaking).

## The candidates' heritage status

Seven raters reported wondering about the candidates' language of origin and guessing where they came from. Fourteen comments were coded into this category.

In the following excerpts, one Cantonese-speaking and two Mandarin-speaking raters discussed how they recognized accents. B1 claimed that the Mandarin candidate's accent was identical to one of his friends who did not live in the Cantonese speaking area.

> [8] *I can perceive that her English accent was totally different from the English spoken by the native people of Guangdong province. Her accent was very close to one of my friends, but she did not belong to our ethnic group* (B1, Cantonese-speaking).

A1 mentioned that the typical phonetic error of mixing /n/ with /l/ reminded her of Southern accent, as shown in (9). Her prior experience with individuals who spoke with strong Southern accent also made her identify the candidate's Southern accent.

> [9] *This person had a strong and obvious Southern accent, which had an effect on his pronunciation. For example, it was related to the common pronunciation mistake in South China that mixed /n/ with /l/. A certain phrase did give me a deep impression, I remember it was "there was no answer", in which /n/ is mispronounced as /l/ by him. Anyway, it was possible that the Southern accent had a certain influence on pronunciation* (A1, Mandarin-speaking).

Besides, A6 inferred that a candidate might be from South China from the way how she pronounced.

> [10] *My first thought is that this student's accent suggested that she might be from the South, as her English pronunciation was a little strange, that is, she couldn't pronounce each sound correctly. It seemed that she only used the front part of the tongue, and seldom the back part* (A6, Mandarin-speaking).

What is interesting about several Mandarin-speaking raters who were unfamiliar with Cantonese is that they took some Cantonese candidates for Indian, Thai, black American English speakers, as shown in (11), which was in accordance with Ballard and Winke's (2017) finding that non-native speakers of English always feel difficult to ascertain the origin of an accent.

> [11] *When I listened to it for the first and second time, I thought that the accent of this person was very similar to Thai English. You know, it was really difficult to understand, and it was kind of weird* (A2, Mandarin-speaking).

However, unlike the Mandarin-speaking raters who were unfamiliar with the Cantonese, four Cantonese-speaking raters succeeded in identifying candidates' heritage status, as demonstrated in the following comments. They indicated that notable features in candidates' pronunciation enabled them to determine that the candidates might come from Guangdong province. This can be seen in the example of rater B5. She made a speculation in (12) about where the candidate might come from and confirmed that the candidate was a Cantonese in a short time based on the accented pronunciation of Cantonese.

> [12] *After listening for just 10 s, I could tell that this student must be a Cantonese, because his pronunciation sounded odd, which only exists in Cantonese people. For some words, the /r/ was going to be a little bit skewed toward /l/. For instance, they pronounce "very" as "vely". I think these are typical features of Cantonese*

> *pronunciation, so I probably listened to it for the first 10 s or so and knew he was a Cantonese* (B5, Cantonese-speaking).

## Raters' scoring strategy

Among the raters who participated in the stimulated recall session, half indicated using the strategy of prioritizing pronunciation while scoring. They expressed the belief that if the candidate's pronunciation was accurate and excellent at first, it would leave a pleasant impression on them, hence they would assign a higher score. It appeared that candidates' performance in *pronunciation* had a significant effect on rating, as illustrated in (13) and (14).

> [13] *According to the four dimensions of the rating scale, first of all, I would assess whether his pronunciation is good and accurate as soon as he opens his mouth. I think there was an evaluation standard in my mind* (A7, Mandarin-speaking).
>
> [14] *First of all, if he speaks out, his pronunciation is very good, the first impression will be good, then if his intonation is good and smooth, and I will definitely give him a high score* (B8, Cantonese-speaking).

# Discussion

## Discussion of RQ 1

The quantitative results demonstrated that there were no appreciable differences in the consistency of each rater when judging test performance, which was in agreement with the findings of other studies that acceptable consistency was obtained in the ratings of raters no matter whether they were familiar or unfamiliar with the first language of the speaker being assessed (Xi and Mollaun, 2011; Winke et al., 2013; Park, 2020). From a theoretical perspective, the findings of this study offer evidence against existence of a shared-dialect effect in rating candidates' performance on the retelling task and bridge the gap of empirical study on the shared-L1 effect within assessment context, especially in the Chinese context which has thus far been the focus of little research.

The reasonably high scoring consistency of the two rater groups might be attributed to the training that both rater groups received. All raters were required to participate in the training session, which resulted in their greater understanding of the rationale for each score. Rater training was effective in helping raters to gain consensual understanding of the categories and criteria represented in the rating rubric and to adopt a common frame of reference (Saito, 2008), leading to greater improvement in the level of agreement between raters. Furthermore, rater reliability reflected *via* scores is not necessarily the sole indicator of an accredited rater' assessment literacy. In this research, in order to guarantee the validity of the ratings, the training calibration test standards incorporated expert raters' reasons for ratings into the measurement of their understanding of the rating scale. As a result, the integration of both psychometric approach and hermeneutic approach (Petruzzi, 2008) to rater training substantially improved the consistency and interpretability of ratings. It may not be surprising that raters might be guided by their experience in the rating process in the absence of rater training, and they tended to determine scores differently based on different levels of experience using the language being tested (Winke and Gass, 2013). Overall, rater training seemed to have helped raters score consistently and confidently.

The two-factor ANOVA analysis found no statistically significant interactions between raters' dialect and candidates' dialect in the total score and in each rating category, suggesting that the two rater groups were equivalent in the scores assigned to the two candidate groups. These results rejected the hypothesis that raters who share the same dialect with the candidate would give a higher score to that candidate than those who do not. Neither the Mandarin-speaking raters nor the Cantonese-speaking raters showed a shared-dialect effect. These findings differed from some published studies (Harding, 2012; Dai and Roever, 2019), but they were aligned with previous studies showing inconsistent effects based on a shared language background (Abeywickrama, 2013; Crowther et al., 2016).

Aside from the aforementioned rater training, the null result could be explained by the analytic rating scale employed in the present study. Previous studies have found that both rating criteria and rater training could become a crucial factor in raters' rating outcomes (Xi and Mollaun, 2011). Typically, raters engage in impressionistic judgment when applying a holistic rubric to rate test-takers' overall speech quality (Xi and Mollaun, 2009). Since the present study used an analytic rubric, raters had to adjust their typical rating behavior and resort to more analytic evaluations in judging the speaking proficiency of examinees who shared the same dialect with them, which may have helped them engage in more reliable and valid evaluations. Moreover, the benchmark samples as exemplars of each score level of the rating scale could guide the raters to determine how similar a sample was to the exemplar, which enhanced raters' understanding of descriptions at different score levels. Hence, raters could articulate scores in some way consistent with the rating scale and provide more accurate and consistent assessments.

The findings tend to suggest that both groups of raters were capable of rating reliably and consistently. Evaluations of oral performance by the two rater groups resulted in the same or roughly similar outcome in terms of aggregate scores. While the selection of raters for the current study cannot be deemed to represent the broader population, these findings provide sound grounds for including both Cantonese-speaking raters and Mandarin-speaking raters in assessing speaking ability in the CELST. It seems that the language background of raters may not matter for scoring purposes in a testing context, and raters from different language backgrounds can be employed interchangeably as long as they have been sufficiently trained.

## Discussion of RQ 2

The second research question is whether raters are aware that sharing the same dialect with the candidates might impact

their rating process. Results from the qualitative analysis indicated that most raters could recognize the candidates' native dialect by their accents. On the whole, accent had little effect on their understanding. It appeared that Cantonese-speaking raters were more capable of identifying candidates' Cantonese accent, compared with their Mandarin-speaking counterparts.

Some raters indicated that awareness of accents and their familiarity with that accent played a role in the comprehension, and potentially affected their scoring process. The present findings seem to be consistent with previous research which showed that scores were affected by accent familiarity, resulting in higher scores (Winke and Gass, 2013; Winke et al., 2013). For example, the Cantonese-speaking rater (B6) displayed a feeling of positive bias in rating candidates with familiar accent. She commented that candidates' speech as a whole was not extremely difficult to understand because she was familiar with this accent. There was, therefore, a possibility for her to become tolerant of accented pronunciation.

Although the two rater groups did not significantly differ in numerical ratings of candidates' oral performance, some raters participated in the stimulated recall stage reported that familiarity with candidates' accents potentially affected their rating decisions. This finding was aligned with the results of prior studies showing discrepancies between raters' assigned ratings and self-perceptions (Xi and Mollaun, 2011; Huang, 2013).

These mismatching results might be due to the mediating effect of raters' attitudes as suggested by Huang (2013). In the present study, Mandarin-speaking raters shared similar attitudes toward the Cantonese accent with the Cantonese rater group. The lack of significant difference in numerical rating may therefore be explained by the similar attitude between the two rater groups. Additionally, the present study focused on the retelling task, which is an integrated task rather than a task that lay mere emphasis on the pronunciation. The core of scoring integrated tasks is the overall oral proficiency rather than pronunciation, which might affect raters' scoring decision. It is possible that a shared-dialect effect is more of a concern with tasks that focuses on pronunciation, like reading-aloud, than with tasks that assess comprehensive speaking ability. Previous studies investigating a possible shared-L1 effect in listening tests suggested that a shared-L1 effect seems to exert different impact on various task types (Dai and Roever, 2019). As a result, the role of task type deserves further exploration.

The potential for test bias in English oral assessment featuring raters with regional dialects has been proved from the cognitive perspective. It provides a foundation for further research on the effect of regional dialects in oral tests, and suggests that a shared-dialect effect is more likely to occur. Although such effect may be made 'steerable' *via* rigorous training, the conflicting results still raise a cautionary red flag that raters' bias caused by personal dialects requires careful monitoring.

## Conclusion

In summary, the shared-dialect benefit was neither observed with Mandarin-speaking raters nor with Cantonese-speaking raters, despite that some raters attended to candidates' accent/dialect and indicated that awareness of accents and their familiarity with the accents affected their comprehension of the speech samples and potentially influenced their scoring process. The above findings add to our knowledge of the shared-dialect benefit and support the claim that including both group raters in the CELST is valid on the condition that rigorous rater training has been provided.

The current study is not without limitations. First, the validity of the whole research procedure could be improved if it was done under normal circumstances. The outbreak of the *COVID-19* pandemic definitely reduced the effectiveness and efficiency of the training procedure and the stimulated recall method. Second, the current study only examined differences in rating behavior among raters at the group level through the classical statistical analyses, which may not be powerful enough to detect differences at the individual level. It might be the case that a higher score assigned by one rater to the candidate was offset by a lower score awarded by another rater in the group, and these variations were not captured in the current analysis treating raters as a group (Huang et al., 2016). As such, more sophisticated statistical analyses such as the Multi-Faceted Rasch model should be employed in future to gain more fine-grained insights into the rater variability. Third, there is a small chance that rating decisions were affected by accent familiarity, but the effect did not entirely demonstrate in the present study. In particular, the mismatch between raters' assigned ratings and self-perceptions demands closer examination of raters' decision-making process. More qualitative data should be collected through other methods (like interview) to triangulate the findings. Finally, since the raters in this study were all young inexperienced female postgraduates, future studies could employ male and/or experienced raters to improve generalizability of the present findings.

## 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 Ethics Committee at School of Foreign Languages, South China University of Technology. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

YX: conceptualization, organization, and revision. MH and JC: revision and expansion. YZ: writing—original draft. All authors contributed to the article and approved the submitted version.

## Funding

This study was supported by the National Social Science Foundation of China (grant no. 22BYY089) and the Fundamental Research Funds for the Central Universities (grant no. QNMS202220).

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1143031/full#supplementary-material

## References


Abeywickrama, P. (2013). Why not non-native varieties of English as listening comprehension test input? *RELC. J.* 44, 59–74. doi: 10.1177/0033688212473270

Bachman, L. F., and Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Ballard, L., and Winke, P. (2017). "Students' attitude towards English teachers' accent: The interplay of accent familiarity, comprehensibility, intelligibility, perceived native speaker status, and acceptability as a teacher," in *Second language pronunciation assessment: Interdisciplinary perspectives*, eds T. Isaacs and P. Trofimovich (Bristol: Multilingual Matters), 121–140. doi: 10.21832/ISAACS6848

Barkana, B. D., and Patel, A. (2020). Analysis of vowel production in Mandarin/Hindi/American-accented English for accent recognition systems. *Appl. Acoust.* 162:107203. doi: 10.1016/j.apacoust.2019.107203

Bent, T., and Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *J. Acoust. Soc. Am.* 114, 1600–1610. doi: 10.1121/1.1603234

Bradlow, A., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Canagarajah, S. (2006). Changing communicative needs, revised assessment objectives: Testing English as an international language. *Lang. Assess. Q.* 3, 229–242. doi: 10.1207/s15434311laq0303_1

Carey, M. D., Mannel, R. H., and Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Lang. Test.* 28, 201–219. doi: 10.1177/0265532210393704

Cargile, A. C. (1997). Attitudes toward Chinese-accented speech: An investigation in two contexts. *J. Lang. Soc. Psychol.* 16, 434–443. doi: 10.1177/0261927X970164004

Crowther, D., Trofimovich, P., and Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *J. Second Lang. Pronunciation* 2, 160–182. doi: 10.1075/jslp.2.2.02cro

Dai, D. W., and Roever, C. (2019). Including L2-English varieties in listening tests for adolescent ESL learners: L1 effects and learner perceptions. *Lang. Assess. Q.* 16, 64–86. doi: 10.1080/15434303.2019.1601198

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Lang. Test.* 33, 117–135. doi: 10.1177/0265532215582282

Elder, C., Barkhuizen, G., Knoch, U., and von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Lang. Test.* 24, 37–64. doi: 10.1177/0265532207071511

Frost, K., Elder, C., and Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Lang. Test.* 29, 345–369. doi: 10.1177/0265532211424479

Gass, S. M., and Mackey, A. (2000). *Stimulated recall methodology in second language research*. New Jersey, NJ: Lawrence Erlbaum Associates.

Gass, S. M., and Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Lang. Learn.* 34, 65–87. doi: 10.1111/J.1467-1770.1984.TB00996.X

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.

Gui, M. (2012). Exploring differences between Chinese and American EFL teachers' evaluations of speech performance. *Lang. Assess. Q.* 9, 186–203. doi: 10.1080/15434303.2011.614030

Han, W., Brebner, C. M., and Mcallister, S. (2015). Redefining 'Chinese' L1 in SLP: Considerations for the assessment of Chinese bilingual/bidialectal language skills. *Int. J. Speech Lang. Pathol.* 18, 135–146. doi: 10.3109/17549507.2015.1081285

Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Lang. Test.* 29, 163–180. doi: 10.1177/0265532211421161

Hayes-Harb, R., Smith, B. L., Bent, T., and Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts. *J. Phon.* 36, 664–679. doi: 10.1016/j.wocn.2008.04.002

Hoskens, M., and Wilson, M. (2001). Real-time feedback on rater drift in constructed-response items: An example from the golden state examination. *J. Educ. Meas.* 38, 121–145. doi: 10.1111/j.1745-3984.2001.tb01119.x

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System* 41, 770–785. doi: 10.1016/j.system.2013.07.009

Huang, B., Alegre, A., and Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Lang. Assess. Q.* 13, 25–41. doi: 10.1080/15434303.2015.1134540

Isaacs, T. (2016). "Assessing speaking," in *Handbook of second language assessment*, eds D. Tsagari and J. Banerjee (Berlin: De Gruyter Mouton), 131–146. doi: 10.1515/9781614513827-011

Isaacs, T., and Thompson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Lang. Assess. Q.* 10, 135–159. doi: 10.1080/15434303.2013.769545

Jin, Y., and Fan, J. (2011). Test for English majors (TEM) in China. *Lang. Test.* 28, 589–596. doi: 10.1177/0265532211414852

Kang, O., Rubin, D., and Kermad, A. (2019a). The effects of training and rater differences on oral proficiency assessment. *Lang. Test.* 36, 481–504. doi: 10.1177/0265532219849522

Kang, O., Thomson, R., and Moran, M. (2019b). The effects of international accents and shared first language on listening comprehension tests. *TESOL Q.* 53, 56–81. doi: 10.1002/tesq.463

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Lang. Assess. Q.* 12, 239–261. doi: 10.1080/15434303.2015.1049353

Law, N. C., and So, L. K. (2006). The relationship of phonological development and language dominance in bilingual Cantonese-Putonghua children. *Int. J. Biling.* 10, 405–427. doi: 10.1177/13670069060100040201

Lee, Y. S., Vakoch, D. A., and Wurm, L. H. (1996). Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *J. Psycholinguist. Res.* 25, 527–542. doi: 10.1007/BF01758181

Li, R. (1989). Classification/distribution of Chinese dialects. *Dialect* 4, 241–259.

Lindemann, S. (2005). Who speaks "broken English"? US undergraduates' perceptions of non-native English. *Int. J. Appl. Linguist.* 15, 187–212. doi: 10.1111/j.1473-4192.2005.00087.x

Lindemann, S. (2017). "Variation or 'error'? Perception of pronunciation variation and implications for assessment," in *Second language pronunciation assessment: Interdisciplinary perspectives*, eds T. Isaacs and P. Trofimovich (Bristol: Multilingual Matters), 193–209. doi: 10.21832/9781783096855-013

Liu, M. (2013). Validation of test of English major band 4 (TEM4) story retelling oral test rating scale. *J. Zhejiang Univ.* 43, 187–194.

Major, R. C., Fitzmaurice, S. F., Bunta, F., and Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Q.* 36, 173–190. doi: 10.2307/3588329

McNamara, T. F. (1996). *Measuring second language performance*. Harlow: Longman.

Munro, M. J., Derwing, T. M., and Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Stud. Second Lang. Acquis.* 28, 111–131. doi: 10.1017/S0272263106060049

National Advisory Committee for Foreign Language Teaching (2004). *Syllabus for test for English majors (Grade 8)*. Shanghai: Shanghai Foreign Language Education Press.

Ockey, G. J., and French, R. (2016). From one to multiple accents on a test of L2 listening comprehension. *Appl. Linguist.* 37, 693–715. doi: 10.1093/applin/amu060

Park, M. S. (2020). Rater effects on L2 oral assessment: Focusing on accent familiarity of L2 teachers. *Lang. Assess. Q.* 17, 231–243. doi: 10.1080/15434303.2020.1731752

Petruzzi, A. (2008). Articulating a hermeneutic theory of writing assessment. *Assess Writ.* 13, 219–242. doi: 10.1016/j.asw.2008.10.005

Saito, H. (2008). EFL classroom peer assessment: Training effects on rating and commenting. *Lang. Test.* 25, 553–581. doi: 10.1177/0265532208094276

Stevenage, S. V., Clarke, G., and McNeill, A. (2012). The "other-accent" effect in voice recognition. *J. Cogn. Psychol.* 24, 647–653. doi: 10.1080/20445911.2012.675321

Stibbard, R. M., and Lee, J. (2006). Evidence against the mismatched interlanguage speech intelligibility benefit hypothesis. *J. Acoust. Soc. Am.* 120, 433–442. doi: 10.1121/1.2203595

Tauroza, S., and Luk, J. C. (1997). Accent and second language listening comprehension. *RELC. J.* 28, 54–71. doi: 10.1177/003368829702800104

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *Am. J. Eval.* 27, 237–246. doi: 10.1177/1098214005283748

Wei, J., and Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Lang. Assess. Q.* 12, 283–304. doi: 10.1080/15434303.2015.1037446

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Lang. Test.* 15, 263–287. doi: 10.1177/026553229801500205

Winke, P. (2012). "Rating oral language," in *The encyclopedia of applied linguistics*, ed. C. A. Chapelle (New York, NY: Blackwell Publishing Ltd), 1–8. doi: 10.1002/9781405198431.wbeal0993

Winke, P., and Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Q.* 47, 762–789. doi: 10.1002/tesq.73

Winke, P., Gass, S., and Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Lang. Test.* 30, 231–252. doi: 10.1177/0265532212456968

Xi, X., and Mollaun, P. (2009). *How do raters from India perform in scoring the TOEFL iBT™ Speaking Section and what kind of training helps?* (TOEFL iBT Research Report RR-09-31). Princeton, NJ: Educational Testing Service.

Xi, X., and Mollaun, P. (2011). Using raters from India to score a large-scale speaking test. *Lang. Learn.* 61, 1222–1255. doi: 10.1111/j.1467-9922.2011.00667.x

Xie, X., and Fowler, C. A. (2013). Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English. *J. Phon.* 41, 369–378. doi: 10.1016/j.wocn.2013.06.003

Xing, Q., Wu, X., Wang, J., and Zhang, Z. (2021). The influence of the matching of modality presentation mode and perceptual learning style on the bidialectal switching cost of Cantonese-Mandarin. *Acta Psychol. Sin.* 53, 1059–1070. doi: 10.3724/sp.j.1041.2021.01059

Xu, Y., and Zeng, Y. (2015). How to deal with students' pronunciation errors — A study based on the corpus of test-takers' performance in the Computer-based English listening and speaking test (CELST) of Gaokao (Guangdong Version). *Curric. Teach. Mater. Method* 35, 90–96.

Yang, Y., Chen, S., and Chen, X. (2020). "F0 patterns in Mandarin statements of Mandarin and Cantonese speakers," in *Proceedings of the annual conference of the international speech communication association, INTERSPEECH*, (Shanghai: The Hong Kong Polytechnic University), 4163–4167. doi: 10.21437/interspeech.2020-2549

Zhang, B., and Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Lang. Test.* 28, 31–50. doi: 10.1177/0265532209360671

Zhang, F. (2014). Washback of the listening and speaking test in NMET Guangdong: Teachers' BAK cognition system and test preparation behaviors. *Foreign Lang. Test. Teach.* 3, 44–49.

Zou, S., and Xu, Q. (2016). A washback study of the test for English majors for grade eight (TEM8) in China — From the perspective of university program administrators. *Lang. Assess. Q.* 14, 140–159. doi: 10.1080/15434303.2016.1235170

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 04 April 2023
DOI 10.3389/fpsyg.2023.1142411



OPEN ACCESS

EDITED BY
Hui Zhang,
Nanjing Normal University, China

REVIEWED BY
Kakeru Yazawa,
University of Tsukuba, Japan
Ming-Che Hsieh,
National Applied Research Laboratories,
Taiwan (Province of China)

*CORRESPONDENCE
Xiaodong Fei
✉ feixiaodonghi@bfsu.edu.cn

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

RECEIVED 11 January 2023
ACCEPTED 13 March 2023
PUBLISHED 04 April 2023

CITATION
Song Q, Fei X and Matsumi N (2023) The lexical processing of Japanese collocations by Chinese Japanese-as-a-Foreign-Language learners: An experimental study by manipulating the presentation modality, semantic transparency, and translational congruency.
*Front. Psychol.* 14:1142411.
doi: 10.3389/fpsyg.2023.1142411



# The lexical processing of Japanese collocations by Chinese Japanese-as-a-Foreign-Language learners: An experimental study by manipulating the presentation modality, semantic transparency, and translational congruency

Qichao Song[1], Xiaodong Fei[2]* and Norio Matsumi[1]

[1]Graduate School of Humanities and Social Sciences, Hiroshima University, Hiroshima, Japan, [2]Beijing Center for Japanese Studies, Beijing Foreign Studies University, Beijing, China

**Introduction:** Research on collocations has become an essential issue in L2 acquisition and cognitive psychology. Previous studies have mainly focused on phonographic languages such as English, Swedish, and German, and primarily discussed the effect of semantic transparency and translational congruency. However, these studies have lacked (1) an analysis of the interactions between presentation modalities (visual vs. auditory) and the semantic transparency and translational congruency, and (2) a discussion of an ideographic language, such as Chinese and Japanese.

**Methods:** We conducted an experiment with 36 Chinese Japanese-as-a-Foreign-Language learners to examine the processing of Japanese collocations. In the experiment, we manipulated the presentation modality, semantic transparency, and translational congruency during a lexical judgment task.

**Results:** Data analysis using linear mixed-effects models revealed the following. (1) In both conditions of semantic transparency and translational congruency, the auditory presentation was associated with longer reaction times than the visual presentation. (2) In the visual presentation condition, neither semantic transparency nor translational congruency showed significant effects. (3) In the auditory presentation condition, the reaction time for collocations with high semantic transparency tended to be longer than that for collocations with medium semantic transparency and significantly longer than that for collocations with low semantic transparency. The reaction time for collocations with congruent translation was longer than that for collocations with incongruent translation.

**Discussion:** These results support the dual-route model of Japanese collocational processing by Chinese Japanese-as-a-Foreign-Language learners.

Our findings suggest that whether the analytic or holistic processing dominates is closely related to the learners' knowledge of Chinese and Japanese *Kanji* words and strongly influenced by the presentation modality, semantic transparency, and translational congruency.

KEYWORDS

**Japanese collocational processing, translational congruency, Chinese JFL learners, presentation modality, semantic transparency**

## 1. Introduction

When learners engage in language activities in a second language (L2), grammatical knowledge and vocabulary knowledge play an essential role (Bernhardt, 2005; Jeon and Yamashita, 2014). Collocation is one of the key points in assessing learners' vocabulary proficiency (e.g., Pawley and Syder, 1983; Nation, 2013; Koizumi and In'nami, 2020; Du et al., 2022). However, it has been noted that mastering collocations is challenging, even for advanced foreign language learners (Laufer and Waldman, 2011). Therefore, research on collocations has attracted attention as an essential issue in L2 acquisition and cognitive psychology (Kjellmer, 1991).

Previous studies have shown that factors such as frequency of use, native language (L1), and L2 proficiency affect the processing of L2 collocations (e.g., Sprenger et al., 2006; Yamashita and Jiang, 2010; Wolter and Gyllstad, 2011; Wolter and Yamashita, 2015, 2018; Zhang and Fang, 2020; Fei and Song, 2021; Jiang, 2022; Song and Fei, 2022). In particular, the research focuses on the co-occurrence strength of each constituent word of the collocation and the bilingual translational relationship, i.e., semantic transparency and translational congruency. Semantic transparency is the degree to which the meaning of a collocation can be inferred from its parts, while translational congruency refers to the fact that the collocation meaning can be translated or inferred with the aid of L1 (e.g., Günther et al., 2020; Song and Fei, 2022).

Previous studies have focused on phonographic languages such as English, Swedish and German (e.g., Wolter and Gyllstad, 2011; Garibyan et al., 2022), some involving Japanese learners of English (e.g., Yamashita and Jiang, 2010), and Chinese learners of English (e.g., Zhang and Fang, 2020; Jiang, 2022). However, there is almost no research on Chinese learners of Japanese (Fei and Song, 2021). Since Chinese and Japanese use Chinese characters, an ideographic writing system, it is clear that Chinese learners of Japanese are strongly influenced by their knowledge of Chinese characters in their lexical processing of Japanese (e.g., Matsumi et al., 2012; Hsieh et al., 2017, 2021; Fei et al., 2022). Therefore, the processing of Japanese collocations consisting of several words by Chinese learners is expected to be complex, and it is possible that different mechanisms would be observed during the processing of Japanese collocations in Chinese learners.

Based on the research results of L2 collocational processing, especially the research results of the collocational processing by Chinese English learners and the lexical processing by Chinese Japanese-as-a-Foreign-Language (JFL) learners, this study aims to clarify the effects of semantic transparency, an essential linguistic characteristic, and translational congruency, which is closely related to bilingualism, upon both visual and auditory presentation of test items. In this study, we explore the lexical processing of Japanese collocations, provide new empirical evidence for the study of collocational processing from the perspective of ideographic characters, and offer suggestions for teachers on how to improve Chinese JFL learners' acquisition of Japanese collocations.

## 2. Literature review

### 2.1. Definition and classification of collocations

From a broad perspective, a collocation is defined as the co-occurrence relationship between words (Sinclair, 1991). In other words, collocations include lexical, grammatical, and contextual elements and are characterized by structural stability, formal unity, and usage restrictiveness (Wray, 2002). For example, the Japanese verb "投げる (nageru), throw" is used in collocations such as "石を投げる (ishi-wo-nageru), throw a stone"; "視線を投げる (shisen-wo-nageru), throw a glance"; "筆を投げる (fude-wo-nageru), throw the pen and give up writing"; "匙を投げる (saji-wo-nageru), beyond remedy." All these collocations exist in Japanese but differ in co-occurrence relationship and the relationship between individual constituents and the overall meaning.

According to the definitions in previous studies of phonographic languages (Nattinger and DeCarrico, 1992), collocations in Japanese were classified into three types by Miyoshi (2007). The three types are collocations with high semantic transparency, collocations with medium semantic transparency, and collocations with low semantic transparency. Collocations with high semantic transparency ("石を投げる") are free word combinations characterized by the lowest strength of linkage between the individual constituent words. Collocations with medium semantic transparency ("視線を投げる") are somewhat fixed word combinations, and the strength of linkage between its individual constituents is medium. Collocations with low semantic transparency ("筆を投げる"; "匙を投げる") are wholly fixed word combinations, and the strength of linkage between the individual constituent words is the strongest. Furthermore, collocations with low semantic transparency were subdivided into figurative idioms and genuine idioms. A figurative idiom's meaning can be inferred from the meanings of its individual constituents ("筆を投げる"), while a genuine idiom's meaning cannot be inferred ("匙を投げる").

The above classification is based on the linguistic characteristics of Japanese. How can the above collocations be classified taking into account the relationship between Chinese and Japanese languages? Matsumi et al. (2012) revealed that the morphological, phonological, and semantic information of L1 Chinese characters significantly impacts Chinese learners' processing of Japanese *Kanji* words. Through behavioral experiments and fMRI scanning, Hsieh et al. (2021) also confirmed that Chinese-Japanese cognate word processing showed longer reaction time and greater activation in the supplementary motor area than L2 control word processing. Therefore, it is essential to classify collocations according to whether or not the Japanese meaning can be inferred from the Chinese meanings of the individual constituent words. Fei and Song (2021) categorized collocations in Japanese into those with matching and non-matching bilingual translations between Chinese and Japanese. In the case of the three types of collocations mentioned above, namely, collocations with high semantic transparency, collocations with medium semantic transparency, and figurative idioms (one subtype of the collocation with low semantic transparency), Chinese learners can infer the overall meaning of the collocation with the aid of L1. In contrast, as for genuine idioms (another subtype of the collocation with low semantic transparency), Chinese learners cannot infer the overall meaning of the collocation with the aid of L1.

More specifically, the semantic transparency of the three collocations "石を投げる" "視線を投げる" "筆を投げる" progressively decreases. "石を投げる" can be translated as "扔掉石头 (rengdiao-shitou)," and "視線を投げる" can be translated as "投去视线 (touqu-shixian)," both retaining the complete or partial original meanings of Chinese characters. Therefore, the overall meaning can be inferred based on the knowledge of L1 Chinese. The literal meaning of "筆を投げる" can be interpreted as "弃笔 (qibi), throw away the pen" and it can be further speculated that its overall meaning is "中途弃写 (zhongtu-qixie)." In contrast, the literal meaning of "匙を投げる," which is also of low semantic transparency, can be interpreted as "弃勺 (qishao)," and after further speculation can be interpreted as "不吃饭 (bu-chifan), do not eat." It is impossible to use the knowledge of L1 to guess right about the meaning of this collocation. Thus it belongs to incongruent translational collocation.

According to the above analysis, it is clear that semantic transparency is a linguistic characteristic of the language itself, while translational congruency is a linguistic characteristic involving two languages. On the basis of the existing research results about collocational processing and the specific characteristics of Japanese collocations, this study comprehensively investigates Japanese collocational processing and compares the results with those about the collocational processing of phonographic languages (such as English and German, but not Hindi and Urdu) and the lexical processing of ideographic languages.

## 2.2. Collocational processing

### 2.2.1. Hypotheses about the collocational processing model

Regarding language processing, Sinclair (1991) proposed two principles: the "open choice principle," under which constituent words are processed according to grammatical rules, and the "idiom principle," under which pre-existing linguistic expressions are processed as a single entity. In addition, Wray (2002) identified two patterns of language processing: analytical processing based on syntactic knowledge and holistic processing using formulaic sequences. It is assumed that these two strategies are used properly when processing languages, and the "idiom principle" is said to have the advantage of reducing cognitive burden. Pawley and Syder (1983) argued that since native speakers have more than thousands of formulaic sequences, including the collocations stored in their mental lexicon, they can process language quickly and accurately. In contrast, L2 learners' mental lexicon stores fewer formulaic sequences and is more inclined to follow the "open choice principle" during L2 processing (Jiang, 2022).

Farrokh (2012) pointed out that in L2 acquisition, learners learn collocations analytically or holistically, depending on the level of semantic transparency. This argument was consistent with the two patterns of language processing mentioned above. In previous studies, there has been debate over whether or not collocations are stored as a whole in the learner's mental lexicon. The main arguments are the full-listing model (e.g., Seidenberg and Gonnerman, 2000; Jiang and Nekrasova, 2007; Conklin and Schmitt, 2008), the decompositional model (e.g., Schmitt and Underwood, 2004; Brooks and Cid de Garcia, 2015), and the dual-route model (e.g., Sprenger et al., 2006; MacGregor and Shtyrov, 2013; Chen et al., 2020).

Previous studies have discussed collocational processing, mainly using reaction times and accuracy rates as indices. Suppose that the reaction time for formulaic sequences is shorter than that for atypical expressions. In that case, the formulaic sequences are stored in the learner's mental lexicon as a whole, and then a full-listing model is supported. On the other hand, if there is no significant difference in reaction time for formulaic sequences and atypical expressions, formulaic sequences are also processed based on syntactic knowledge, and the decompositional model is supported. According to Zhang et al. (2021), many studies supported the full-listing model. However, Xu and Wang (2015) argued that the "shorter reaction time" may not be direct evidence of the full-listing model. They pointed out that the highly frequent co-occurrence seen in collocations may result in a stronger linkage between its constituent words, thus improving the efficiency of processing the collocations for L2 learners and reflecting a shorter reaction time. Nevertheless, this does not entirely imply that collocations with short reaction times are stored in the mental lexicon as a lexical entry. Each constituent word may still be stored separately in the mental lexicon, but the processing speed becomes faster because of the strong linkage. This led to the question of the validity of the full-listing model. Therefore, the dual-route model, in which the decompositional model and the full-listing model coexist, was proposed. Many studies tried to find out which processing model is dominant by manipulating various factors (e.g., Sosa and MacFarlane, 2002; Wolter and Gyllstad, 2011; Fang and Zhang, 2021).

Regarding the L2 lexical processing model, Kroll and Stewart (1994) proposed the revised hierarchical model, which has been widely used. This model clearly proposes that the L2 lexicon is independent and shares conceptual representation with L1. This model is suitable for studying languages with relatively independent glyphs and sounds, such as Chinese and Japanese (Fei et al., 2022).

Sharing Chinese character representations may make Japanese collocational processing by Chinese learners more closely related to their L1. Therefore, it is necessary to focus on bilingual representation activation when exploring the Japanese collocational processing model. In addition, previous studies mainly explored the selection of the full-listing model or the decompositional model based on reaction time. The revised hierarchical model is also based on reaction time to show the processing path between two languages. Based on these, the present study aims to systematically explore the model of Chinese JFL's Japanese collocational processing by describing the representational links between Chinese and Japanese bilingual mental lexicon.

### 2.2.2. Factors affecting L2 collocational processing

Factors affecting L2 collocational processing have been divided into two main categories. They are internal factors such as frequency of use, semantic transparency, and translational congruency, and external factors other than the collocation, such as L2 proficiency, context, and presentation modality.

It has already been shown that frequency of use strongly influences the acquisition and processing of L2, and there is a common understanding among studies to date. Research results showed that the frequency of collocation use had an effect, using eye movements as a measure of eye tracking (e.g., Siyanova et al., 2011) and reaction times (e.g., Wolter and Gyllstad, 2013). More precisely, collocations used more frequently showed a processing advantage over collocations used less frequently.

It was found that semantic transparency affects L2 collocational processing. Collocations with low semantic transparency were observed to be associated with shorter reaction times than collocations with high semantic transparency (Libben et al., 2003; Brooks and Cid de Garcia, 2015; Chen et al., 2020). Collocations with low semantic transparency, such as formulaic sequences and idioms, were more likely to be processed by the full-listing model. On the other hand, results were also reported where semantic transparency did not significantly affect L2 collocational processing (Frisson et al., 2008). In this regard, Zhang et al. (2021) conducted a priming task under different SOA (stimulus onset asynchrony) conditions (200 ms, 400 ms, and 600 ms), and their results suggested that the effect of semantic transparency was weakened by the processing time of collocations. Chen et al. (2020) also showed that the effect of semantic transparency is influenced by factors such as the frequency of use. Therefore, the influence of semantic transparency should be examined together with other factors.

Additionally, it has been shown that translational congruency, which reflects the relationship between L1 and L2, affects collocational processing. It was observed that the reaction time was shorter when the bilingual translation of collocation matched (i.e., congruent translational collocation) than when it did not (e.g., Yamashita and Jiang, 2010; Wolter and Gyllstad, 2011; Zhang, 2017). On the other hand, similar to semantic transparency, some studies reported that translational congruency had no effect on reaction time (e.g., Fang and Zhang, 2019; Fei and Song, 2021). It was noted that the effect of translation congruency is affected by the linguistic distance (i.e., degree of the actual difference) between the two languages (Wolter and Yamashita, 2015) and the judgment criteria for translational congruency (Zhang and Fang, 2020). Studies focused on phonographic language learners, and few discussed about ideographic Chinese characters. Considering the sharing of some Chinese character representations and the short linguistic distance (see Chai and Bao, 2023), different results may be found in Chinese-Japanese bilingual research.

In addition to internal factors due to the linguistic characteristics of collocation, the influence of external factors also cannot be ignored. Pawley and Syder (2000) pointed out that the more fluent the speaker is, the more pauses are placed between phrases, and there are almost no breaks in sound due to pauses or hesitation within phrases. It can be inferred that learners' L2 proficiency affects their processing of collocations. In this regard, experimental studies that manipulated learners' L2 proficiency demonstrated that proficiency affects collocational processing and how frequency of use, semantic transparency, and translational congruency function (Wolter and Yamashita, 2018; Zhang and Fang, 2020). Therefore, if language ability is not taken as a factor when exploring collocational processing, it is necessary to control the language ability of participants. Furthermore, it was reported that context affects the influences of semantic transparency and translational congruency on collocational processing (Jiang and Nekrasova, 2007; Cervera and Rosell, 2015). It was shown that context relatively weakly affected semantic transparency's influence, while it significantly affected translational congruency's influence on collocational processing (Song and Fei, 2022).

Whether discussing internal or external factors, most of the studies examined the processing of collocations using the visual presentation condition. A previous study using an auditory presentation condition reported that semantic transparency had a strong effect on the processing of Japanese collocations by Chinese learners of Japanese, while translational congruency had a weak effect (Fei and Song, 2021). Zhang and Fang (2020) pointed out that the influence of each factor depends on the experimental paradigm. Further studies are needed to examine how different presentation modalities change the effects of semantic transparency and translational congruency on collocational processing.

## 2.3. Objectives and issues of this study

According to the literature review of previous studies, the following three points became clear. First, most studies dealt with phonographic languages, and only a few studies dealt with languages that use ideographic Chinese characters. Second, the effect of the frequency of use of L2 was consistently observed in all the studies. Third, semantic transparency and translational congruency affected collocational processing in the visually presented condition, but how they affect collocational processing depended on the context. However, the results of using the auditory presentation modality and the interactions between presentation modalities and the linguistic characteristics of collocations have not been extensively investigated. We believe that collocational processing can be clarified by examining these issues. In particular, exploring these issues can further demonstrate how the decompositional model and the full-listing model can co-exist.

Based on these results, this study examines the following three research questions.

RQ1. How do presentation modality (visual vs. auditory) and semantic transparency interact to affect the semantic processing of Japanese collocations?

RQ2. How do presentation modality and translational congruency interact to affect the semantic processing of Japanese collocations?
RQ3. What kinds of Japanese collocational processing models do Chinese JFL learners follow?

# 3. Materials and methods

## 3.1. Participants

A total of 36 advanced Chinese JFL learners (female, 22; male, 14) participated in the experiment. The average age of the participants was 24.7 years old, and all were enrolled in graduate school in China. All participants began studying Japanese in their first year of college, with an average of 5.9 years of Japanese language study. All participants had normal vision (corrected) and hearing. And all had attained Japanese-Language Proficiency Test (JLPT) N1 certificate (the most difficult level, the ability to understand Japanese used in various circumstances). The participants were randomly divided into an auditory presentation experimental group and a visual presentation experimental group. Ten participants (auditory presentation, 4) had experience attending Japanese universities as exchange students. **Table 1** reported the N1 mean score, the self-report of Chinese and Japanese proficiency, the length of studying abroad, and Japanese usage frequency (referred to The Language History Questionnaire, LHQ-3, see Li et al., 2020) of participants in the case of different presentation modalities. The independent samples *t*-test for every indicator showed no significant difference between the two groups ($ps > 0.05$). Each condition excluded the influence of test differences on the experimental results as much as possible.

After the experimenter's detailed explanation of the study, all participants voluntarily signed the informed consent form, which clearly states that the experimental data will only be used for academic research and the personal information of the participants will never be disclosed to others. At the end of the experiment, all participants received a reward of 30 yuan.

## 3.2. Design

The present study employed a lexical decision task to compare our results with those of previous studies. Using linear mixed-effects models (LMMs), we aimed to examine the effects of semantic transparency and translational congruency on L2 Japanese collocational processing by advanced Chinese JFL learners. In our experimental design the presentation modality and semantic transparency, or the presentation modality and translational congruency, were fixed factors, respectively. To ensure that semantic transparency and translational congruency did not interfere with each other (Fang and Zhang, 2021), the semantic transparency experiment was conducted on collocations with congruent translation, while the translational congruency

TABLE 1 Participants' language proficiency, Japanese usage frequency, and comparisons of different presentation modalities.

| | | | Auditory presentation | Visual presentation | *t*-test results | | |
|---|---|---|---|---|---|---|---|
| | | | | | *t* | *p* | *cohen'd* |
| JLPT N1 score | | | 152.83 (16.53) | 148.72 (19.86) | 0.68 | 0.504 | 0.23 |
| Self-report proficiency scores | L | C | 6.44 (0.51) | 6.39 (0.50) | 0.33 | 0.744 | 0.11 |
| | | J | 5.28 (0.67) | 5.11 (0.83) | 0.66 | 0.512 | 0.22 |
| | S | C | 6.06 (0.64) | 5.83 (0.62) | 1.06 | 0.297 | 0.35 |
| | | J | 4.61 (1.15) | 4.50 (0.71) | 0.35 | 0.728 | 0.12 |
| | R | C | 6.50 (0.51) | 6.33 (0.49) | 1.00 | 0.324 | 0.33 |
| | | J | 5.67 (0.84) | 5.89 (0.76) | 0.83 | 0.411 | 0.28 |
| | W | C | 5.83 (0.79) | 5.78 (0.65) | 0.23 | 0.818 | 0.08 |
| | | J | 4.78 (0.94) | 5.06 (0.87) | 0.92 | 0.365 | 0.31 |
| Length of studying abroad (months) | | | 2.06 (4.22) | 2.83 (4.63) | 0.53 | 0.602 | 0.18 |
| Japanese usage frequency (hours/day) | | | 4.11 (1.08) | 4.39 (1.41) | 0.66 | 0.511 | 0.22 |

The total score of JLPT N1 is 180. L, listening; S, speaking; R, reading; W, writing; C, Chinese; J, Japanese. 1, none ~ 7, near native-like. Results are expressed as mean (*SD*).

experiment was conducted on collocations with low semantic transparency.

## 3.3. Materials

Experimental materials were selected from three textbooks commonly used by Japanese majors in China (i.e., "ZongHe RiYu (Comprehensive Japanese), Peking University Press, 2007"; "XinBan ZhongRi JiaoLiu BiaoZhun RiBenYu (The New Edition of Standard Japanese for Sino-Japanese Communication), People's Education Press and Mitsumura Tosho Publishing Co., Ltd., 2005"; "XinBian RiYu (Newly Compiled Japanese), Shanghai Foreign Language Education Press, 2009"). Collocations are multiword units that are more complex than words. Although the participants had attained JLPT N1 certificate, they were still unbalanced bilinguals, as indicated by their Japanese language proficiency self-assessment scores in Table 1. Considering the language ability of the participants and to ensure the psychological authenticity of the experimental material, collocations containing words from the JLPT N1 and words above the JLPT level were excluded.

According to the definition mentioned above and the classification of collocations, a total of 48 collocations in four conditions (12 collocations for each condition) were created (see **Supplementary material**). They were (A) collocations with high semantic transparency and congruent translation with L1, (B) collocations with medium semantic transparency and congruent translation with L1, (C) collocations with low semantic transparency and congruent translation with L1, and (D) collocations with low semantic transparency and incongruent translation with L1 (Table 2).

The number of mora of collocation, the frequency of use by the Tsukuba Web Corpus (developed by the University of Tsukuba and the search engine is provided by Japan National Institute for Japanese Language and Linguistics), and the MI score (Mutual Information score, a measure of collocational strength) were calculated. Furthermore, one-way analysis of variance (ANOVA) was conducted for each characteristic. The results showed that no main effects were significant in any condition [number of mora: ($F(3, 44) = 1.21$, $p = 0.316$, $\eta^2 = 0.08$); log-transformed frequency of use: ($F(3, 44) = 1.88$, $p = 0.147$, $\eta^2 = 0.11$); MI score: ($F(3, 44) = 0.48$, $p = 0.699$, $\eta^2 = 0.03$)]. We also asked 279 advanced Chinese learners of Japanese to rate the degree of familiarity with all collocations on a 7-point scale (1: not at all familiar ~ 7: very familiar). One-way ANOVA was conducted on their mean ratings, and the main effect was not significant ($F(3, 44) = 2.00$, $p = 0.128$, $\eta^2 = 0.12$).

Materials of semantic transparency and translational congruency are classified according to the results of research in the field of linguistics (Nattinger and DeCarrico, 1992; Miyoshi, 2007). Moreover, 23 Japanese L1 speakers were asked to rate the semantic transparency of the collocations in each condition on a 7-point scale (1: lowest semantic transparency ~ 7: highest semantic transparency). One-way ANOVA was performed on the mean ratings of the four lists, and the main effect was significant ($F(3, 44) = 46.48$, $p < 0.001$, $\eta^2 = 0.76$). Tukey's HSD Test for multiple comparisons found that A > B > C≒D (A-B: $t(44) = 4.58$, $p < 0.001$, *cohen's* $d = 1.87$; A-C: $t(44) = 9.28$, $p < 0.001$, *cohen's* $d = 3.79$; A-D: $t(44) = 10.58$, $p < 0.001$, *cohen's* $d = 4.32$; B-C: $t(44) = 4.70$, $p < 0.001$, *cohen's* $d = 1.92$; B-D: $t(44) = 6.01$, $p < 0.001$, *cohen's* $d = 2.45$; C-D: $t(44) = 1.31$, $p = 0.565$, *cohen's* $d = 0.53$). Next, to reconfirm the reasonableness of the selection of congruent translational collocations (i.e., A, B, C) and incongruent translational collocations (i.e., D), we asked five advanced Chinese JFL learners to judge the materials selected by the authors based on the judgment of translational congruency, i.e., whether the Chinese and Japanese meanings were congruent. The result showed no objection to the selection of materials based on the judgment of translational congruency.

Finally, 3 lists, (A), (B), and (C), were used for the semantic transparency experiment, and 2 lists, (C) and (D), were used for the translational congruency experiment. In addition, 36 non-collocations (i.e., not exist in Japanese) were created for the lexical judgment task. The audio stimuli were recorded by a woman from the Greater Tokyo Area who had experience teaching Japanese, and the recordings were edited for auditory presentation.

TABLE 2 Summary of characteristics of the test items.

| Type | Log-transformed frequency of use | MI score | Mora | Familiarity | Semantic transparency | Example |
|---|---|---|---|---|---|---|
| A | 3.14 (0.45) | 9.88 (2.28) | 6.92 (1.24) | 6.52 (0.39) | 5.78 (0.25) | 自信を失う (jishin-wo-ushinau, lose confidence) |
| B | 3.28 (0.44) | 10.85 (2.17) | 6.58 (1.08) | 6.23 (0.42) | 4.79 (0.88) | 世話を焼く (sewa-wo-yaku, take care of someone) |
| C | 2.88 (0.72) | 10.81 (2.79) | 6.25 (0.87) | 6.14 (0.36) | 3.78 (0.39) | 耳を疑う (mimi-wo-utagau, be hard to believe) |
| D | 2.81 (0.56) | 10.07 (2.73) | 6.25 (0.75) | 6.14 (0.58) | 3.49 (0.36) | 油を売る (abura-wo-uru, loaf around) |

The types of collocations were (A) collocations with high semantic transparency and congruent translation with L1, (B) collocations with medium semantic transparency and congruent translation with L1, (C) collocations with low semantic transparency and congruent translation with L1, and (D) collocations with low semantic transparency and incongruent translation with L1. Twelve collocations for each type of collocation were prepared as the test items. Frequency of use and MI score are based on Tsukuba Web Corpus. Chinese JFL learners rated familiarity. Japanese native speakers rated semantic transparency. Results are expressed as mean (*SD*).

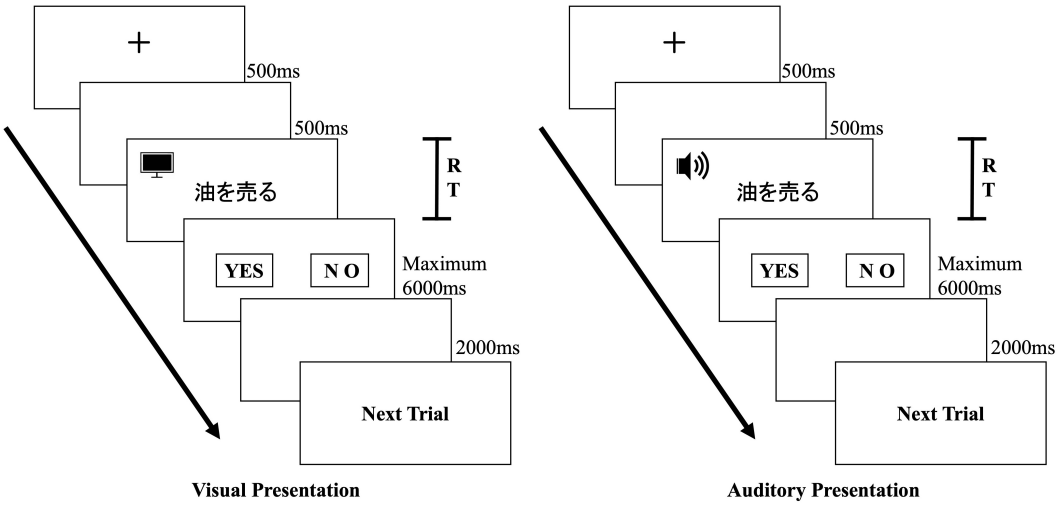



FIGURE 1
Flow of one experiment trial in the visual and auditory presentation conditions.

## 3.4. Apparatus

The computer program used in the experiment was created using E-Prime 2.0 software. In the auditory presentation experiment, the auditory stimuli were presented through headphones. A personal computer and peripherals were used to present the computer program.

## 3.5. Procedure

The experiments were conducted individually. Before starting the experiment, we conducted five practice sets to ensure that the participant understood the experimental procedure. Figure 1 shows the flow of the experiment for one trial.

In the visual presentation experiment, as a cue to present the collocation, the gaze point was presented on the computer screen for 500 ms. Then, after a blank space of 500 ms, the collocation was presented. Participants were instructed to judge whether the collocation presented on the computer screen was a correct Japanese collocation or not as quickly and accurately as possible, and to press the "Yes (*Z* key)" or "No (*M* key)" buttons, respectively. The collocation was presented for a maximum of 6000 ms. During that time, either the participant responded or, if no response was made and 6000 ms had elapsed, a blank screen was presented for 2000 ms, and then the next trial was started. The computer automatically measured the time from the start of the visual presentation until the participant pressed the key as the reaction time for the collocation.

In the auditory presentation experiment, as in the visual presentation experiment, after the gaze point was presented for 500 ms, a blank space of 500 ms was presented. Then, the participant listened to the collocation presented auditorily through headphones. Participants were instructed to judge whether the collocation they had heard was a correct Japanese collocation or not as quickly and accurately as possible, and to press the "Yes (*Z* key)" or "No (*M* key)" buttons, respectively. After the auditory presentation, if the participant had responded or if there was no response and 6000 ms had elapsed, a blank screen was presented for 2000 ms, and then the next trial was started. The computer automatically measured the time from the end of the auditory presentation until the participant pressed the key as the reaction time for the collocation.

In both the visual and auditory presentation experiments, the collocations were randomly presented by the computer program. After the completion of all trials, unknown collocations were checked by the participants. A written questionnaire about the participant's language learning experiences was administered.

# 4. Results and discussion

## 4.1. Data trimming

Only correct responses to Yes trials were included in the analysis. The percentage of incorrect responses was 15.22%, and the percentage of more than 2.5*S Ds* beyond the mean and collocations the participants did not know was 3.82%. The incorrect responses and collocations the participants did not know were excluded from the analysis. To deal with the skewed data, reaction times were log-transformed. Tables 3, 4 show the results of the reaction times of correct responses to Yes trials and the accuracy rates in the semantic transparency and translational congruency experiments.

Data analyses were conducted using the software R (version 4.2.1, R Core Team, 2022). We adopted linear-mixed effects modeling utilizing the lme4 (Bates et al., 2015) and lmerTest (Kuznetsova et al., 2017) packages. The emmeans package (Lenth, 2022) was used to examine interactions. The model with the lowest Akaike information criterion (AIC) was selected as the optimal model for model fitting. Wald *t*-distribution approximation was used to compute *p*-values for reaction times data. Wald *z*-distribution was used to compute *p*-values for the accuracy rates data. Since semantic transparency and translational congruency are related to the type of collocation, analyzing all stimulus items

TABLE 3 Reaction times and accuracy rates in each condition of the semantic transparency experiment.

| Variables | | Reaction times (ms) | Accuracy rates (%) |
|---|---|---|---|
| Visual | Low semantic transparency | 1341.17 (693.26) | 85.19 (10.52) |
| | Medium semantic transparency | 1312.84 (480.59) | 82.48 (11.39) |
| | High semantic transparency | 1203.09 (420.71) | 93.52 (12.31) |
| Auditory | Low semantic transparency | 2232.68 (720.41) | 83.33 (9.48) |
| | Medium semantic transparency | 2231.58 (707.50) | 79.63 (14.64) |
| | High semantic transparency | 2477.98 (829.72) | 85.19 (17.04) |

Results are expressed as mean (*SD*).

TABLE 4 Reaction times and accuracy rates in each condition of the translational congruency experiment.

| Variables | | Reaction times (ms) | Accuracy rates (%) |
|---|---|---|---|
| Visual | Incongruent translation | 1256.99 (459.32) | 79.63 (12.53) |
| | Congruent translation | 1357.44 (628.08) | 85.19 (10.52) |
| Auditory | Incongruent translation | 1839.68 (482.59) | 86.57 (10.36) |
| | Congruent translation | 2180.70 (679.43) | 83.33 (9.48) |

Results are expressed as mean (*SD*).

as random effects can improve the accuracy of the experimental results (Song and Fei, 2022).

## 4.2. Results and discussion of the semantic transparency experiment

Based on the AIC, semantic transparency, presentation modality, and the interaction between semantic transparency and presentation modality were selected as fixed effects and experimental participants and items were selected as random effects in the model. The analysis of results of the semantic transparency experiment are reported in **Figure 2** and **Table 5**.

In the semantic transparency experiment, the main effect of presentation modality was significant. The visual presentation condition was associated with shorter reaction times than the auditory presentation condition ($t(50.26) = 8.88$, $p < 0.001$). The main effect of semantic transparency was not significant ($p > 0.100$). Conversely, since the interaction between presentation modality and semantic transparency was significant, simple main effects were tested. Results indicated that in the auditory presentation condition, the reaction time for collocations with high semantic transparency tended to be longer than that for collocations with medium semantic transparency ($t(33.0) = 1.96$, $p = 0.058$) and was significantly longer than that for collocations with low semantic transparency ($t(32.4) = 2.04$, $p = 0.050$). There was no significant difference in reaction time between medium and low transparency collocations ($t(33.0) = 0.29$, $p = 0.946$). Meanwhile, it was shown that the effect of semantic transparency was not significant in the visual presentation condition [high/medium semantic transparency: ($t(32.5) = 1.23$, $p = 0.228$); high/low semantic transparency: ($t(32.2) = 1.45$, $p = 0.156$); medium/low semantic transparency: ($t(33.5) = 0.22$, $p = 0.831$)]. In addition, under all conditions, the reaction time upon visual presentation was significantly shorter than the reaction time upon auditory presentation [high semantic transparency: ($t(50.6) = 8.88$, $p < 0.001$); medium semantic transparency: ($t(52.2) = 6.67$, $p < 0.001$); low semantic transparency: ($t(51.1) = 6.52$, $p < 0.001$)].

We also analyzed the accuracy rates using the *glmer* function. The results showed that the main effect of presentation modality was significant, and the visual presentation condition was associated with higher accuracy rates than the auditory presentation condition ($z = 2.41$, $p = 0.016$) (**Figure 3** and **Table 6**). The main effect of semantic transparency was significant. The accuracy rate was higher for collocations with high semantic transparency than for collocations with medium semantic transparency ($z = 2.21$, $p = 0.027$). Since the interaction between presentation modality and semantic transparency tended to be significant, simple main effects were tested. Results indicated that in the visual presentation condition, the accuracy rate tended to be higher for collocations with high semantic transparency than for those with medium semantic transparency ($z = 2.22$, $p = 0.068$). In addition, among collocations with high semantic transparency, the accuracy rate was significantly higher in the visual presentation condition than in the auditory presentation condition ($z = 2.41$, $p = 0.016$).

These results indicate that presentation modality and semantic transparency interact in Japanese collocational processing. Based on these results, we will discuss the first research question of this study. The effect of semantic transparency in the auditory and visual presentation conditions differed, and it can be inferred that collocational processing differs depending on presentation modality. In the auditory presentation condition, the reaction time for collocations with high semantic transparency was longer than that for collocations with medium or low semantic transparency. This result is consistent with the results of studies of phonographic languages using the visual presentation condition (Brooks and Cid de Garcia, 2015; Chen et al., 2020).

In contrast, the effect of semantic transparency was less pronounced in the visual presentation condition. In the case of visual presentation, the presence of ideographic characters (Chinese characters) may have led to the superiority of decompositional processing, regardless of the degree of semantic transparency. In the auditory presentation condition, collocations with high semantic transparency were associated with the longest reaction times. This result may have been obtained because collocations with low semantic transparency are more likely to be processed as a whole than collocations with high semantic transparency.

The results about accuracy rates also showed that collocational processing differed depending on presentation modality. In the visual presentation condition, the accuracy rate tended to be higher for collocations with high semantic transparency than for those with medium semantic transparency. Meanwhile, the accuracy rate in the visual presentation condition was higher than that in the auditory presentation condition for collocations with high semantic transparency. These results reconfirmed the strong effect of L1. When collocations with high semantic transparency were

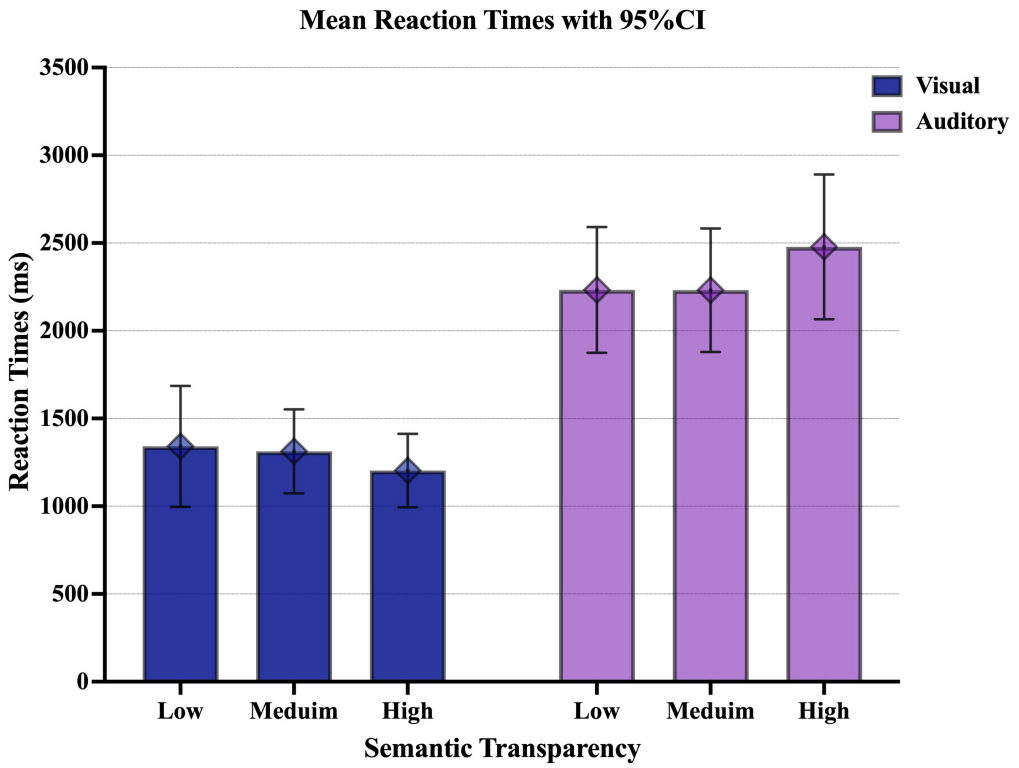



FIGURE 2
Graph of the mean reaction times upon visual or auditory presentation of Japanese collocations with different degrees of semantic transparency. The error bar shows the 95% confidence interval (CI).

TABLE 5 Results of LME model analysis of reaction times in the semantic transparency experiment.

| Variables | Estimate | SE | df | *t*-value | Pr (>\|*t*\|) |
|---|---|---|---|---|---|
| Intercept | 3.07 | 0.03 | 56.04 | 113.56*** | <0.001 |
| modalityAud | 0.32 | 0.04 | 50.26 | 8.88*** | <0.001 |
| conditionB | 0.03 | 0.02 | 31.28 | 1.23 | 0.228 |
| conditionC | 0.03 | 0.02 | 31.01 | 1.45 | 0.156 |
| conditionB: modalityAud | −0.08 | 0.03 | 32.68 | −2.92** | 0.006 |
| conditionC: modalityAud | −0.08 | 0.03 | 31.64 | −3.20** | 0.003 |

**$p < 0.01$, ***$p < 0.001$. Participants = 36. Items = 36. Total observation = 1037. SE, standard error; df, degree of freedom. The optimal model is *lmer*(logrt~modality × condition + (1|participant) + (modality|item), data = datAC).

visually presented to Chinese JFL learners, the learners could maximize the use of their knowledge of Chinese characters (e.g., Fei and Song, 2021).

Based on the above discussion, it can be said that when applying previous research results (e.g., Farrokh, 2012) to vocabulary teaching in order to promote the acquisition of collocations, if the students are Chinese JFL learners, the influences of presentation modality and L1 knowledge should also be taken into account.

## 4.3. Results and discussion of the translational congruency experiment

The results of the translational congruency experiment were analyzed (Figure 4 and Table 7). Based on the AIC, the model with translational congruency, presentation modality, and the interaction between translational congruency and presentation modality as fixed effects and experimental participants and items as random effects was determined to be the optimal model.

The main effect of presentation modality was significant, and the reaction time was significantly longer in the auditory presentation condition than in the visual presentation condition ($t(39.33) = 6.89$, $p < 0.001$). The main effect of translational congruency was not significant, and there was no significant difference in reaction time between the congruent translational collocations and the incongruent translational collocations ($t(39.33) = 0.89$, $p = 0.381$). The interaction between translational congruency and presentation modality was significant. More specifically, in the auditory presentation condition, congruent translational collocations were associated with longer reaction times than incongruent collocations ($t(29.4) = 2.83$, $p = 0.008$).

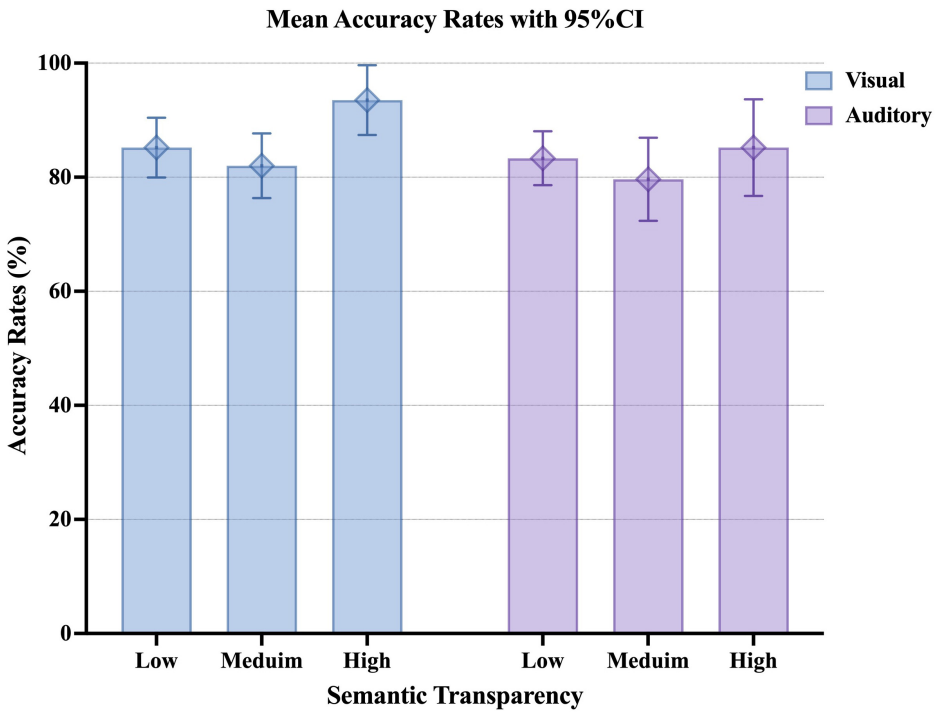



FIGURE 3
Graph of the mean accuracy rates upon visual or auditory presentation of Japanese collocations with different degrees of semantic transparency. The error bar shows the 95% CI.

However, there was no significant difference in reaction time between congruent translational collocations and incongruent collocations in the visual presentation condition ($t(29.9) = 0.89$, $p = 0.382$). We also found that auditory presentation was associated with significantly longer reaction times than visual presentation, whether congruent translational collocations ($t(39.4) = 6.89$, $p < 0.001$) or incongruent translational collocations ($t(39.2) = 5.33$, $p < 0.001$) were presented.

We also analyzed the accuracy rates using the *glmer* function. The results showed that the main effect of the presentation modality ($z = 0.52$, $p = 0.604$) and the main effect of translational congruency ($z = 0.97$, $p = 0.331$) were not significant (Figure 5 and Table 8). However, the interaction between presentation modality and translational congruency tended to be significant. As a result of simple main effects, the auditory presentation condition tended to be associated with higher accuracy rates than the visual presentation condition among incongruent translational collocations ($z = 1.83$, $p = 0.068$).

These results indicate that presentation modality and translational congruency interact in Japanese collocational processing. Based on these results, we will discuss the second research question of this study. In the visual presentation condition, the effect of translational congruency was weak, whereas, in the auditory presentation condition, the effect of translational congruency was strong. In the visual presentation condition, decompositional processing of syntactic analysis may have been dominant, regardless of the difference in translational congruency. This point differs from the results of studies on phonographic languages (Wolter and Gyllstad, 2011; Zhang, 2017). It can be seen that the influence of ideograms, i.e., Chinese characters, is stronger than that of phonetic characters. On the other hand, in the auditory presentation condition, incongruent translational collocations may have been predominantly processed as a whole.

In addition, the accuracy rate in the case of auditory presentation tended to be higher than that in the case of visual presentation only for incongruent translational collocations. Since the incongruent translational collocations all had low semantic transparency, it can be inferred that even collocations with low semantic transparency are affected differently by L1 and the presentation modality. These results further support the hypothesis that Chinese JFL learners process Japanese collocations in different ways in the case of different presentation modalities. Therefore, the

TABLE 6 Results of GLME model analysis of accuracy rates in the semantic transparency experiment.

| Variables | Estimate | SE | *z*-value | Pr (>\|*z*\|) |
|---|---|---|---|---|
| Intercept | 3.11 | 0.44 | 7.10*** | <0.001 |
| modalityAud | −1.00 | 0.41 | −2.41* | 0.016 |
| conditionB | −1.17 | 0.53 | −2.21* | 0.027 |
| conditionC | −0.83 | 0.54 | −1.54 | 0.123 |
| conditionB: modalityAud | 0.80 | 0.43 | 1.84† | 0.065 |
| conditionC: modalityAud | 0.85 | 0.45 | 1.89† | 0.058 |

†$p < 0.10$, *$p < 0.05$, ***$p < 0.001$. Participants = 36. Items = 36. Total observation = 1296. SE, standard error. The optimal model is *glmer* (acc~modality × condition + (1|participant) + (1|item), family = binomial, data = datAC).

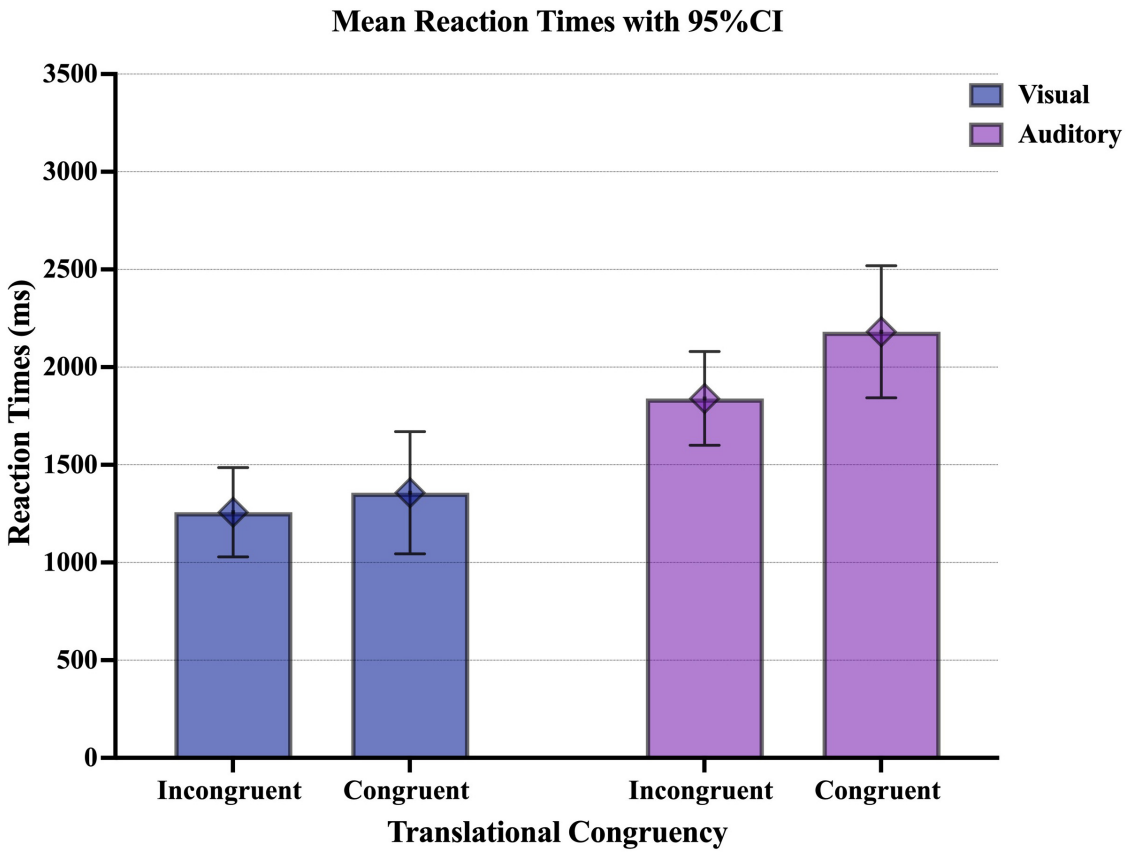



FIGURE 4
Graph of the mean reaction times upon visual or auditory presentation of Japanese collocations with low semantic transparency and different translational congruency. The error bar shows the 95% CI.

TABLE 7 Results of LME model analysis of reaction times in the translational congruency experiment.

| Variables | Estimate | SE | df | *t*-value | Pr (>\|*t*\|) |
|---|---|---|---|---|---|
| Intercept | 3.11 | 0.03 | 54.28 | 118.62*** | <0.001 |
| modalityAud | 0.22 | 0.03 | 39.33 | 6.89*** | <0.001 |
| conditionD | −0.02 | 0.02 | 30.21 | −0.89 | 0.381 |
| conditionD: modalityAud | −0.05 | 0.02 | 629.62 | −2.99** | 0.003 |

**$p < 0.01$, ***$p < 0.001$. Participants = 36. Items = 24. Total observation = 687. SE, standard error; df, degree of freedom. The optimal model is *lmer*(logrt~modality × condition + (1|participant) + (1|item), data = datCD).

presentation modality should be considered when discussing the Japanese collocational processing model in Chinese JFL learners.

## 5. General discussion

When Chinese JFL learners process Japanese *Kanji* words, their L1 knowledge of Chinese characters will be activated simultaneously and have a solid facilitative or inhibitory effect upon different presentation modalities (e.g., Matsumi et al., 2012; Hsieh et al., 2017, 2021; Fei et al., 2022). Therefore, it can be inferred that the Japanese collocational processing in Chinese learners may differ from that in learners of phonographic languages. In this study, we empirically investigated the effects of presentation modality, semantic transparency, and translational congruency on the processing of Japanese collocations by Chinese JFL learners. The experimental results revealed that these three factors are closely related to the processing of collocations, which complements the results of research on the processing of ideographic characters and provides a new empirical basis for the dual-route model.

Based on the results of this study and previous studies (i.e., the revised hierarchical model, Kroll and Stewart, 1994), we propose a Japanese collocational processing model for Chinese JFL learners in the case of different presentation modalities, as shown in **Figure 6** (congruent translational collocations with different semantic transparencies) and **Figure 7** (incongruent translational collocations with low semantic transparency). The thickness of the connecting arrows means the strength of the links between the representations, and the dotted arrows indicate that direct semantic access is impossible. We comprehensively examine the processing

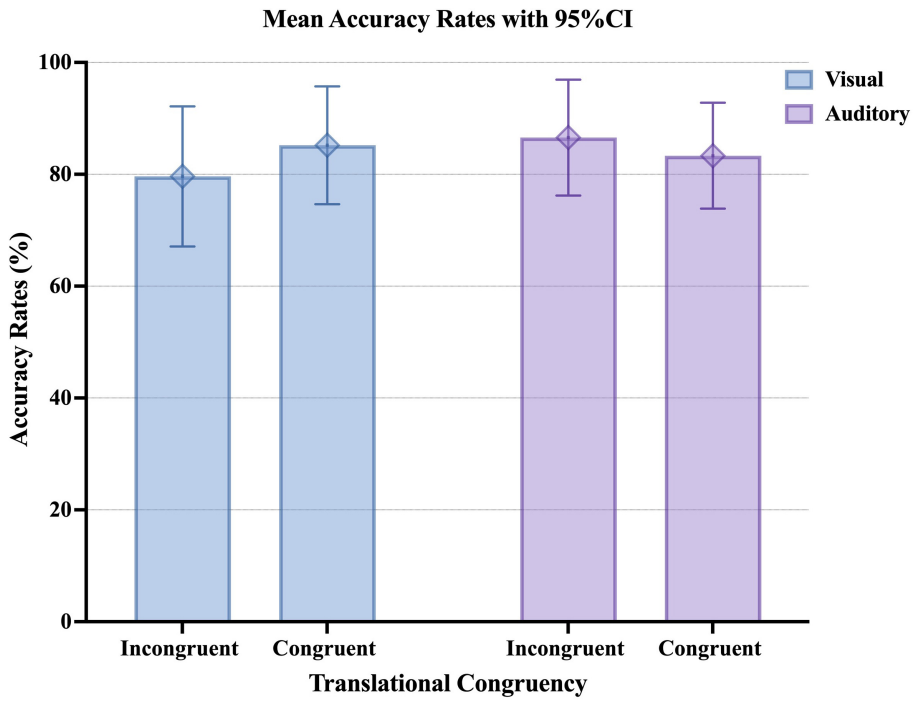



FIGURE 5
Graph of the mean accuracy rates upon visual or auditory presentation of Japanese collocations with low semantic transparency and different translational congruency. The error bar shows the 95% CI.

of collocations by Chinese learners of Japanese from the perspective of the influence of their knowledge of Chinese characters in the two languages and discuss the third research question of this study.

## 5.1. Effect of semantic transparency on collocational processing in the case of different presentation modalities

The effect of semantic transparency was not pronounced in the visual presentation condition. As mentioned above, all semantic transparency experiments in this study used congruent translational collocations. In other words, when the Chinese JFL learners received the visual input of Japanese collocations, the L1 semantics were rapidly activated, and the meaning of the collocation could be inferred from the L1 semantics. As shown in Figure 6, after visual presentation of the Japanese collocation, the L1 lexical representation was quickly activated, suggesting that semantic access *via* L1 was dominant (Figure 6A, ① → ②).

TABLE 8 Results of GLME model analysis of accuracy rates in the translational congruency experiment.

| Variables | Estimate | SE | *z*-value | Pr (>\|z\|) |
|---|---|---|---|---|
| Intercept | 2.27 | 0.42 | 5.40*** | <0.001 |
| modalityAud | −0.17 | 0.33 | −0.52 | 0.604 |
| conditionD | −0.54 | 0.55 | −0.97 | 0.331 |
| conditionD: modalityAud | 0.77 | 0.40 | 1.92† | 0.055 |

†$p < 0.10$, ***$p < 0.001$. Participants = 36. Items = 24. Total observation = 864. SE, standard error. The optimal model is *glmer* (acc~modality × condition + (1|participant) + (1|item), family = binomial, data = datCD).

In contrast, in the auditory presentation condition, the reaction time for collocations with high semantic transparency tended to be longer than the reaction time for collocations with medium semantic transparency and was significantly longer than the reaction time for collocations with low semantic transparency. The degree of activation of the L1 Chinese lexical representations may have been weaker in the auditory presentation condition than in the visual presentation condition. In addition, the degree of activation of L1 may differ depending on the level of semantic transparency. Based on these two reasons, it can be inferred that collocations with medium or low semantic transparency, in which the linkage between the constituent words is more robust than in collocations with high semantic transparency, have dominant semantic access to conceptual representations directly from Japanese phonological information (Figure 6B, line ③).

The results of the present study support the existence of a dual-route model (e.g., Sosa and MacFarlane, 2002; Wolter and Gyllstad, 2011; Fang and Zhang, 2021), as claimed in previous studies. Furthermore, from the perspective of presentation modality, we were able to show a new way of interpreting the dual-route model. That is, in the visual presentation condition, the influence of ideographic Chinese characters was strong and changed the collocation processing pattern. As previously mentioned, studies on visual presentation of collocations in phonographic languages have shown that collocations with low semantic transparency dominate overall processing (Libben et al., 2003; Brooks and Cid de Garcia, 2015; Chen et al., 2020). However, according to the

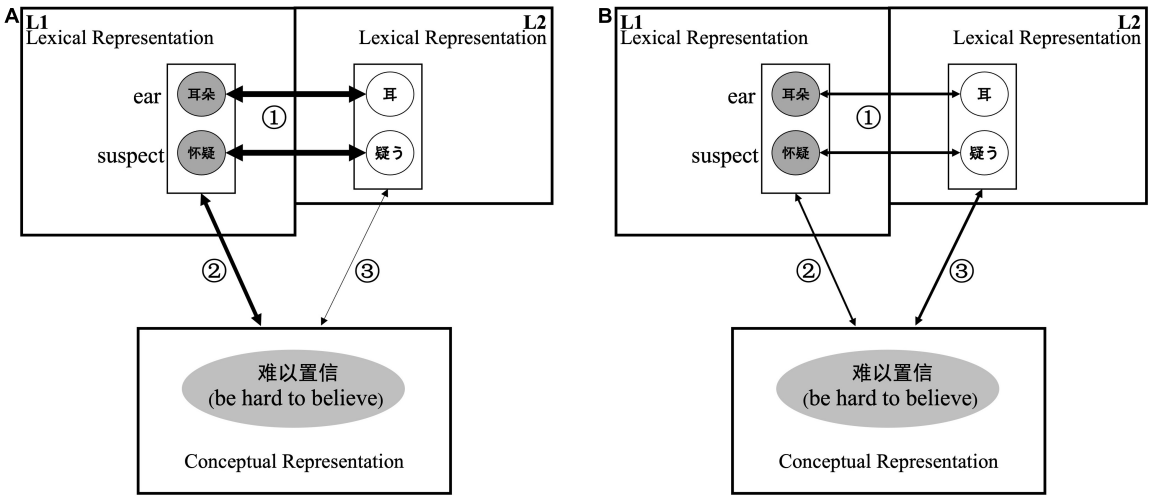



FIGURE 6
Processing of a congruent translational collocation in Chinese JFL learners. The thickness of the arrows indicates the strength of the links between representations. **(A)** Represents the processing of congruent translational collocation upon visual presentation. **(B)** Represents the processing of congruent translational collocation upon auditory presentation.

results of the present study focusing on Japanese collocational processing in Chinese JFL learners, there was no significant difference in reaction time upon visual presentation of collocations with different semantic transparency. Therefore, it can be further speculated that even collocations with low semantic transparency may be analytically processed by syntactic analysis upon visual presentation.

On the other hand, in the case of auditory presentation, the results were consistent with those of phonographic language studies. This indicates that the morphology of Chinese characters has a relatively weak effect on overall semantic processing upon auditory presentation of Japanese collocations. This differs from the results of research on the lexical processing of Japanese *Kanji* words (Fei et al., 2022). We can infer that the word combination changed the influence of Chinese characters on semantic processing to some extent. Therefore, collocations with low semantic transparency are more prone to full-listing processing at the semantic access stage.

## 5.2. Effect of translational congruency on collocational processing in the case of different presentation modalities

In the auditory presentation condition, congruent translational collocations were associated with longer reaction times than incongruent translational collocations. As discussed above, collocations with low semantic transparency have dominant semantic access from the Japanese lexical representation directly to the conceptual representation (**Figure 6B**, line ③). Based on this result, it can be inferred that incongruent translational collocations have more dominant semantic access to the conceptual representation directly from the L2 lexical representation (**Figures 7A, B**, line ③).

However, there was no significant difference in reaction times in the visual presentation condition, regardless of whether or not the translation between Chinese and Japanese was congruent. As mentioned above, conceptual representation access *via* L1 semantic representation was dominant in the visually presented condition, even for collocations with low semantic transparency. Since incongruent collocations cannot be accessed directly from the activated lexical representation in L1 (**Figure 7A**, ① → ②), semantic access to the conceptual representation may be performed directly from the L2 Japanese lexical representation (**Figure 7A**, line ③). Therefore, it can be inferred that the reaction time is lengthened by trying to access L1. Ultimately, there was no significant difference in reaction time for incongruent translational collocations compared with the reaction times for congruent translational collocations.

Similar to the semantic transparency results, it is clear that the influence of translational congruency differed depending on presentation modality. The results of the visual presentation condition were consistent with the results of the study of Fang and Zhang (2019) on Chinese learners of English. However, through the above analysis and discussion, the absence of a significant difference in reaction time was not due to the absence of the effect of L1 but due to the prolonged reaction times resulting from access to conceptual representations through L1 lexical representation.

Nevertheless, the results in the case of auditory presentation in the present study differ from the results reported in learners of phonographic languages in the visual presentation condition, the latter of which showed that the reaction times were shorter when processing congruent translational collocations (Yamashita and Jiang, 2010; Wolter and Gyllstad, 2011; Zhang, 2017). Chinese JFL learners spent longer reaction times when processing congruent translational collocations. Based on the experimental results of semantic transparency and translational congruency, in the auditory presentation condition, Chinese JFL learners were more likely to process Japanese collocations with low semantic transparency and incongruent translation as a whole (**Figure 7B**, line ③).

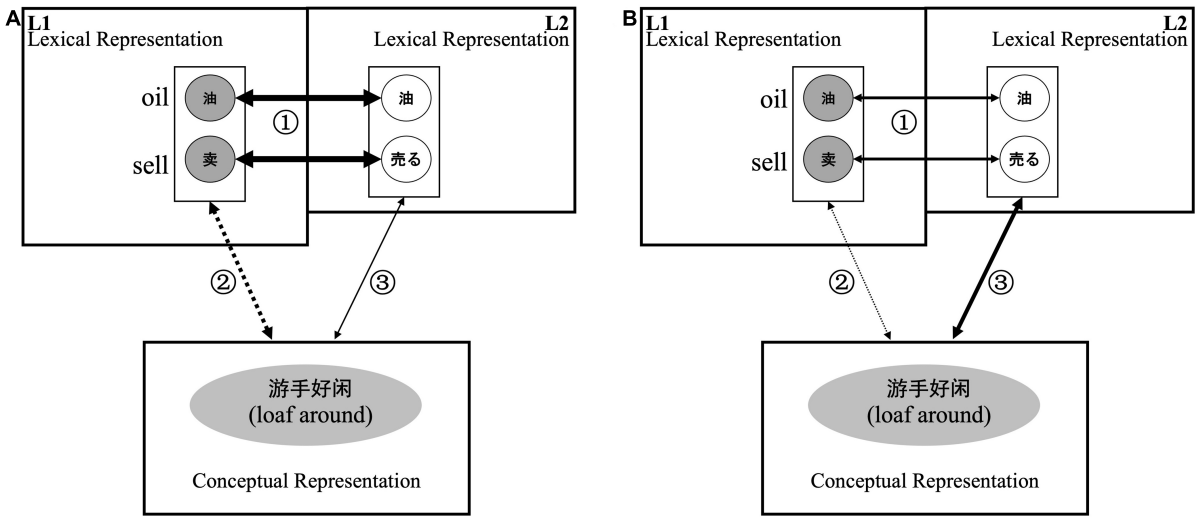



FIGURE 7
Processing of an incongruent translational collocation in Chinese JFL learners. The thickness of the arrows indicates the strength of the links between representations and the dotted arrows indicate that direct semantic access is impossible. **(A)** Represents the processing of incongruent translational collocation upon visual presentation. **(B)** Represents the processing of incongruent translational collocation upon auditory presentation.

## 5.3. Suggestions for teachers on Chinese JFL learners' acquisition of Japanese collocations

Based on the experimental results and the discussion, we hereby put forward some suggestions on Japanese collocational acquisition by Chinese JFL learners.

Firstly, Japanese collocation teaching must pay attention to the influence of presentation modality. Within the scope of descriptive statistics, it can be seen that in the visual presentation modality situation, the collocations with low semantic transparency took longer reaction times, while in the auditory presentation modality situation, the collocations with high semantic transparency took longer reaction times. The influence of semantic transparency on auditory presentation modality is more substantial than that on visual presentation modality. Chinese JFL learners tend to rely on visual information in Japanese language acquisition and therefore need to pay attention to auditory information-obtaining exercises (e.g., Fei et al., 2022), the lack of which may result in slower development of listening comprehension. In this regard, we propose two suggestions. First, teachers must remind learners to do more targeted auditory information-capturing exercises. Visual information is compared with auditory information to strengthen the linkage in terms of orthography, phonology, and semantics. The second suggestion is to intensify collocational practices. Even for collocations with high semantic transparency, teachers still need to list several common collocations, such as "石を投げる (ishi-wo-nageru), throw stones," "石を拾う (ishi-wo-hirou), pick up stones," "石を積む (ishi-wo-tsumu), pile up stones," "石を運ぶ (ishi-wo-hakobu), carry stones," etc. By strengthening visual and auditory exercises, the processing of collocations with high semantic transparency is upgraded to holistic processing.

Secondly, Japanese collocation teaching must pay attention to the influence of L1. In both cases of visual and auditory presentation modalities, the reaction times for collocations with congruent translation were longer. This suggests that L1 is activated and has an impact on the processing of congruent translational collocations. Especially in the case of visual presentation modality, learners can quickly understand the meaning with the aid of L1, while ignoring the integrity of collocations. Moreover, the accuracy rates for collocations with incongruent translation in the case of visual presentation modality were the lowest. For instance, the meaning of "油を売る (abura-wo-uru), loaf around" has no relation to the meanings of "油 (abura), oil" and "売る (uru), sell." Due to the strong visual influence of L1, Chinese learners are more likely to make mistakes in the quick lexical decision of collocations. In addition, some participants made wrong responses to the two collocations of "困難を開く" and "興味を上げる" as fillers, thinking they were correct Japanese collocations, even though they do not exist in Japanese. Based on these results, we infer that this may also be due to the negative transfer of their L1 because there are two collocations of "解开困局 (jiekai-kunju), untie the dilemma" and "提高兴趣 (tigao-xingqu), enhance interest" in Chinese. However, since Chinese learners cannot wholly exclude the influence of L1, teachers' advice to "avoid the influence of L1 as much as possible" may backfire and confuse learners. Therefore, it is recommended that teachers point out the difference between L1 Chinese and L2 Japanese to help learners understand the similarities and differences of the two languages. By understanding the characteristics of Chinese-Japanese bilingual vocabulary, learners can further strengthen the linkage between the orthography, phonology, and semantics of L2 and upgrade the processing of collocations with congruent translation to the level of holistic processing.

The above suggestions for teachers aim to comprehensively improve learners' L2 processing efficiency and understanding. They improve the efficiency of visual information processing and help learners with language use in reading, test taking, and other scenarios. Likewise, they improve the efficiency of auditory information processing and help learners to apply

language knowledge in listening comprehension, communication, and other activities.

## 6. Conclusion

In this study, based on the findings of previous studies, we comprehensively examined the processing of Japanese collocations by Chinese JFL learners. The results revealed that the linguistic characteristics of collocations in Japanese and the relationship between the two languages strongly influence the processing of Japanese collocations. Moreover, the influence of these two factors is closely related to presentation modality, such as visual or auditory presentation.

Learning and memorizing as a whole is generally advocated for the acquisition of L2 collocations. However, according to the results of this study, in the process of acquisition of Japanese collocations by Chinese JFL learners, it seems necessary to find ways to treat the relevance of their knowledge of Chinese characters and Japanese *Kanji* words correctly. We await further empirical studies using eye-tracking devices to investigate these issues, mainly focusing on the relationship between the Chinese and Japanese languages for each constituent word.

## Data availability statement

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

## 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 patients/participants provided their written informed consent to participate in this study.

## Author contributions

QS, XF, and NM contributed to the conception of the work. QS and XF collected the experimental data. QS conducted the analysis and wrote the first manuscript. All authors revised the manuscript and confirmed its final version.

## Funding

This study was supported by the "Excellent Talents Program of Beijing Foreign Studies University 2021".

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1142411/full#supplementary-material

## References

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annu. Rev. Appl. Linguist.* 25, 133–150. doi: 10.1017/S0267190505000073

Brooks, T. L., and Cid de Garcia, D. (2015). Evidence for morphological composition in compound words using MEG. *Front. Hum. Neurosci.* 9:215. doi: 10.3389/fnhum.2015.00215

Cervera, T., and Rosell, V. (2015). The effects of linguistic context on word recognition in noise by elderly listeners using Spanish sentence lists (SSL). *J. Psycholinguist. Res.* 44, 819–829. doi: 10.1007/s10936-014-9321-7

Chai, X., and Bao, J. (2023). Linguistic distances between native languages and Chinese influence acquisition of Chinese character, vocabulary, and grammar. *Front. Psychol.* 13:1083574. doi: 10.3389/fpsyg.2022.1083574

Chen, S., Peng, Y., Wang, S., Ma, Q., and Yang, L. (2020). The influence of familiarity and semantic transparency on the processing of English compound Nouns. *Mod. Foreign Lang.* 43, 94–105.

Conklin, K., and Schmitt, N. (2008). Formulaic sequences: are they processed more quickly than nonformulaic language by native and nonnative speakers? *Appl. Linguist.* 29, 72–89. doi: 10.1093/applin/amm022

Du, X., Afzaal, M., and Al Fadda, H. (2022). Collocation use in EFL learners' writing across multiple language proficiencies: a corpus-driven study. *Front. Psychol.* 13:752134. doi: 10.3389/fpsyg.2022.752134

Fang, N., and Zhang, P. (2019). Effects of translational congruency and L2 lexical frequency on English collocation processing. *Mod. Foreign Lang.* 42, 85–97.

Fang, N., and Zhang, P. (2021). L1 congruency, word frequency, collocational frequency, L2 proficiency, and their combined effects on Chinese–English bilinguals' L2 collocational processing. *Int. J. Bilingualism* 25, 1429–1445. doi: 10.1177/13670069211024747

Farrokh, P. (2012). Raising awareness of collocation in ESL/EFL classrooms. *J. Stud. Educ.* 3, 55–74. doi: 10.5296/jse.v2i3.1615

Fei, X., and Song, Q. (2021). The processing of Japanese collocations — Evidence from NIRS. *Foreign Lang. Teach. Res.* 53, 91–101. doi: 10.19923/j.cnki.fltr.2021.01.008

Fei, X., Zhao, S., and Liu, J. (2022). Auditory recognition of Chinese-Japanese cognates and homographs by Chinese JFL learners. *Psychologia* 64, 1–22. doi: 10.2117/psysoc.2021-A144

Frisson, S., Niswander-Klement, E., and Pollatsek, A. (2008). The role of semantic transparency in the processing of English compound words. *Br. J. Psychol.* 99, 87–107. doi: 10.1348/000712607X181304

Garibyan, A., Schilling, A., Boehm, C., Zankl, A., and Krauss, P. (2022). Neural correlates of linguistic collocations during continuous speech perception. *Front. Psychol.* 13:1076339. doi: 10.3389/fpsyg.2022.1076339

Günther, F., Marelli, M., and Bölte, J. (2020). Semantic transparency effects in German compounds: a large dataset and multiple-task investigation. *Behav. Res. Methods* 52, 1208–1224. doi: 10.3758/s13428-019-01311-4

Hsieh, M. C., Jeong, H., Kawata, K. H. D. S., Sasaki, Y., Lee, H. C., Yokoyama, S., et al. (2017). Neural correlates of bilingual language control during interlingual homographs in a logogram writing system. *Brain Lang.* 174, 72–85. 2017.06.006 doi: 10.1016/j.bandl

Hsieh, M. C., Jeong, H., Sugiura, M., and Kawashima, R. (2021). Neural evidence of language membership control in bilingual word recognition: an fMRI study of cognate processing in Chinese-Japanese Bilinguals. *Front. Psychol.* 12:643211. doi: 10.3389/fpsyg.2021.643211

Jeon, E. H., and Yamashita, J. (2014). L2 reading comprehension and its correlates: a meta-analysis. *Lang. Learn.* 64, 160–212. doi: 10.1111/lang.12034

Jiang, N., and Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Mod. Lang. J.* 91, 433–445. doi: 10.1111/j.1540-4781.2007.00589.x

Jiang, Y. (2022). Effects of L1-L2 congruency, collocation type, and restriction on processing L2 collocations. *Front. Psychol.* 13:947725. doi: 10.3389/fpsyg.2022.947725

Kjellmer, G. (1991). "A mint of phrases," in *English corpus linguistics: studies in honor of Jan Svartvik*, eds K. Aijmer and B. Altenberg (London: Longman), 111–127. doi: 10.2196/16962

Koizumi, R., and In'nami, Y. (2020). Structural equation modeling of vocabulary size and depth using conventional and Bayesian methods. *Front. Psychol.* 11:618. doi: 10.3389/fpsyg.2020.00618

Kroll, J. F., and Stewart, E. (1994). Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *J. Mem. Lang.* 33, 149–174. doi: 10.1006/jmla.1994.1008

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

Laufer, B., and Waldman, T. (2011). Verb-noun collocations in second-language writing: a corpus analysis of learners' English. *Lang. Learn.* 61, 647–672. doi: 10.1111/j.1467-9922.2010.00621.x

Lenth, R. (2022). *_emmeans: estimated Marginal Means, aka Least-Squares Means_. R package version 1.7.5*. Available online at: https://CRAN.R-project.org/package=emmeans (accessed December 18, 2022).

Li, P., Zhang, F., Yu, A., and Zhao, X. (2020). Language history questionnaire (LHQ3): an enhanced tool for assessing multilingual experience. *Bilingualism Lang. Cogn.* 23, 938–944. doi: 10.1017/S1366728918001153

Libben, G., Gibson, M., Yoon, Y. B., and Sandra, D. (2003). Compound fracture: The role of semantic transparency and morphological headedness. *Brain Lang.* 84, 50–64. doi: 10.1016/s0093-934x(02)00520-5

MacGregor, L. J., and Shtyrov, Y. (2013). Multiple routes for compound word processing in the brain: evidence from EEG. *Brain Lang.* 126, 217–229. doi: 10.1016/j.bandl.2013.04.002

Matsumi, N., Fei, X., and Cai, F. (2012). "The lexical processing of Japanese *Kanji* words," in *Second language acquisition research and language education*, eds K. Hatasa, Y. Hatasa, M. Kudara, and T. Shimizu (Tokyo: Kuroshio Shuppan Press), 43–67.

Miyoshi, Y. (2007). An experimental study of the effects of teaching words in collocations. *J. Japanese Lang. Teach.* 134, 80–89.

Nation, I. S. P. (2013). *Learning vocabulary in another language*, Second Edn. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139858656

Nattinger, J. R., and DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Pawley, A., and Syder, F. H. (1983). "Two puzzles for linguistic theory: nativelike selection and nativelike fluency," in *Language & communication*, eds J. C. Richards and R. W. Schmidt (London: Longman), 191–226.

Pawley, A., and Syder, F. H. (2000). "The one-clause-at-a-time hypothesis," in *Perspectives on fluency*, ed. H. Riggenbach (Ann Arbor, MI: University of Michigan Press), 163–199.

R Core Team (2022). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Schmitt, N., and Underwood, G. (2004). "Exploring the processing of formulaic sequences through a self-paced reading task," in *Formulaic sequences: acquisition, processing and use*, ed. N. Schmitt (Philadelphia, PA: John Benjamins), 173–189. doi: 10.1075/lllt.9.10sch

Seidenberg, M. S., and Gonnerman, L. M. (2000). Explaining derivational morphology as the convergence of codes. *Trends Cogn. Sci.* 4, 353–361. doi: 10.1016/S1364-6613(00)01515-1

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. New York, NY: Oxford University Press.

Siyanova, A., Conklin, K., and Van Heuven, W. J. (2011). Seeing a phrase "time and again" matters: the role of phrasal frequency in the processing of multiword sequences. *J. Exp. Psychol. Learn. Mem. Cogn.* 37:776. doi: 10.1037/a0022531

Song, Q., and Fei, X. (2022). The effects of semantic transparency, translational congruency and context on Japanese collocational processing. *Mod. For. Lang.* 54, 659–670.

Sosa, A. V., and MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: collocations involving the word of. *Brain Lang.* 83, 227–236. doi: 10.1016/S0093-934X(02)00032-9

Sprenger, S. A., Levelt, W., and Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *J. Mem. Lang.* 54, 161–184. doi: 10.1016/j.jml.2005.11.001

Wolter, B., and Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Appl. Linguist.* 32, 430–449. doi: 10.1093/applin/amr011

Wolter, B., and Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Stud. Second Lang. Acquisit.* 35, 451–482. doi: 10.1017/S0272263113000107

Wolter, B., and Yamashita, J. (2015). Processing collocations in a second language: a case of first language activation? *Appl. Psycholinguist.* 36, 1193–1221. doi: 10.1017/S0142716414000113

Wolter, B., and Yamashita, J. (2018). Word frequency, collocational frequency, L1 congruency, and proficiency in L2 collocational processing. *Stud. Second Lang. Acquisit.* 40, 395–416. doi: 10.1017/S0272263117000237

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511519772

Xu, Y., and Wang, T. (2015). The effects of formulaic sequence frequency, structures and English proficiency on Chinese EFL learners' processing of formulaic sequences. *Foreign Lang. Teach. Res.* 47, 393–404.

Yamashita, J., and Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Q.* 44, 647–668. doi: 10.5054/tq.2010.235998

Zhang, P. (2017). L1 semantic priming and word type effect: processing of Chinese EFL learners' English collocations. *Foreign Lang. Teach. Res.* 49, 818–832.

Zhang, P., and Fang, N. (2020). Effects of lexical frequency, semantic congruency, and language proficiency on English collocational processing by Chinese EFL learners. *Foreign Lang. Teach. Res.* 52, 532–545. doi: 10.19923/j.cnki.fltr.2020.04.005

Zhang, P., Lu, Y., and Zhang, B. (2021). Effects of constituent position and semantic transparency on the time course of L2 English compound processing. *J. PLA Univ. Foreign Lang.* 44, 1–8.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 04 April 2023
DOI 10.3389/fpsyg.2023.1126194



OPEN ACCESS

EDITED BY
Dingfang Shu,
Shanghai International Studies University, China

REVIEWED BY
Reza Kafipour,
Shiraz University of Medical Sciences, Iran
Fahd Alalwi,
Prince Sattam Bin Abdulaziz University, Saudi Arabia

*CORRESPONDENCE
Qunfeng Wang
felixwang41@163.com

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

RECEIVED 17 December 2022
ACCEPTED 22 February 2023
PUBLISHED 04 April 2023

CITATION
Wang Q (2023) Text memorization: An effective strategy to improve Chinese EFL learners' argumentative writing proficiency. *Front. Psychol.* 14:1126194. doi: 10.3389/fpsyg.2023.1126194



# Text memorization: An effective strategy to improve Chinese EFL learners' argumentative writing proficiency

Qunfeng Wang*

School of Humanities and Foreign Languages, Xi'an University of Technology, Xi'an, China

This study aims to explore the impact of text memorization strategies on Chinese EFL learners' English argumentative writing proficiency, the process of their text memorization, and specific strategies deployed for the enhancement of the memorization effect. Seven text memorization tests, one pre-test, and one post-test were administrated to 33 Chinese English majors to, respectively, examine students' memorization outcomes as well as their English argumentative writing proficiency before and after memorizing seven model English writings. Data were also collected through interviews with the 12 top scorers in text memorization tests. The results showed that text memorization as a foreign language learning strategy significantly impacted the improvement of EFL learners' argumentative writing proficiency. Moreover, in the text memorization process, in which varieties of strategies were employed, it was found that storage was preceded by understanding among the majority of the interviewees. Since text memorization was found to be advantageous to EFL learners' writing proficiency, a new system of text memorization strategies was developed in the current study to provide both scholars and teachers with insight into text memorization strategies associated with the writing skills of EFL learners.

KEYWORDS

**English writing proficiency, text memorization strategy, text memorization process, language learning strategies, EFL learners**

## 1. Introduction

The relationship between language learning strategies (LLSs) and language achievements has long been the subject of research, much of which suggests that LLSs, as an aid, are effective in successful language learning (Griffiths and Soruç, 2020). Since some of the strategies seem to remain inordinately attached to or associated with language skills in specific areas (Oxford, 2017), there is increasing interest in the investigation into the strategies employed in language skill areas, such as reading and writing. Among the four modalities: listening, reading, speaking, and writing in foreign language learning (FLL), writing, at times, frustrates and challenges the majority of foreign language learners (Fareed et al., 2016). The development of writing skills requires English as a foreign language (EFL) learners to present clear ideas in line with their thinking by the application of linguistic knowledge, for instance, the arrangement of words, clauses, and sentences in a coherent manner based on systematic rules (Hyland, 2003). For EFL learners, memorizing this knowledge by using effective memorization strategies is the initial step to not only develop their writing skills but also improve their overall language proficiency. Therefore, memorization is viewed as one of the essential learning strategies for FLL, the

importance of which is highlighted in the definition of learning strategies. It refers to operations that learners deploy for acquisition, storage, retrieval, or use of information (Dansereau, 1985), or intentional behavior and thoughts used by learners during learning to better help them understand, learn, or remember new information (Richards and Platt, 1992).

Given the significance of memorization in language learning, due attention has been given to revealing the relationship between the choice of language learners' memorization strategies and their outcome of vocabulary learning. Previous research suggests a positive correlation between memorization strategy use and vocabulary achievement (Rashidi and Omid, 2011). Since memory strategies are frequently applied to memorize vocabulary and structures during the early phase of language learning, studies on memorization thus far seem to be limited to linking memorization with just the amelioration of learning vocabulary, the smaller units of language. However, in many Asian nations, for example, China, the memorization material has extended from a single word or character to complex texts, the larger units of language, such as sentences, paragraphs, and whole essays. When acquiring L1 from kindergarten to university, Chinese students are instructed and encouraged to memorize varieties of texts. The content and style of these texts vary from Chinese proverbs and poems to full articles by representatives of masters of Chinese literature. As text memorization has sustained and developed into a traditional Chinese literacy learning strategy, it has also been transferred to the process of FLL and become a crucial approach for both Chinese acquisition and English learning.

On account of the wide employment of memorization as a learning strategy, Chinese learners used to be stereotypically characterized in Western educational settings, as passive learners who rely heavily on memorization of material in their learning process (Chan and Rao, 2009), but much more recent research has shed light on the vital roles that memorization plays in FLL (Mouziraji and Mouziraji, 2015; Khamees, 2016; Sonmez, 2018). A body of research into Chinese EFL learners' memorization strategy use indicated that Chinese learners were not rote memorizers but active archivers who applied a series of memorization strategies to facilitate their language learning (Biggs, 1996; Kember, 2000; Li and Cutting, 2011). However, in many previous studies, "text memorization" was not differentiated from "memorization" academically. In other words, the two terms were mostly used interchangeably without consideration of the difference in the length of the material for memorization, which would influence the specific memorization strategy use and memorization process. Based on the differentiation of text memorization from memorization, a few studies attempted to unfold the relationship between text memorization and the language proficiency of EFL learners. It was suggested that text memorization worked effectively to enhance EFL learners' language competence (Dai and Ding, 2010). Few studies have been conducted to relate text memorization as a learning strategy to proficiency in English writing, with a systematic observation and analysis of strategies involved in the text memorization process.

In China, though Chinese EFL learners employ text memorization strategies widely in their English learning processes, many of them use a variety of specific text memorization strategies unsystematically since, until now, there is no system for text memorization strategy available to refer to and help enhance the effects of text memorization. Moreover, in the field of teaching English as a foreign language, the focus of English writing teaching lies in writing strategies, such as planning strategy, while-writing strategy, and revising strategy [as suggested by Petri and Czárl (2003)] or other relevant writing skills, for example, idea development. As such, the influences that the traditional Chinese literacy learning strategy (text memorization), may exert on Chinese ESL learners' English writing outcomes have been ignored. Therefore, with the aim of providing EFL learners and teachers with new insight into the use of text memorization strategies and its impact on EFL learners' English writing proficiency, this study tends to focus on uncovering the relationship between text memorization and English writing proficiency, particularly the proficiency of English argumentative writing, through the exploration of Chinese EFL learners' processes of text memorization and the specific strategies involved.

## 2. Literature review

### 2.1. Classifications of LLSs and memorization strategies

Since language learning requires skillful employment of an array of strategies, efforts have been made in the previous studies to classify LLSs with diverse perspectives to present a system of LLSs. The classification scheme developed by Rubin (1981) includes two general categories. The first group of strategies contributes directly to learning and is subdivided into clarification/verification, monitoring, memorization, guessing/inductive inferencing, deductive reasoning, and practice. The second group of strategies contributes indirectly to learning (e.g., creating opportunities for practice and production tricks). According to O'Malley and Chamot (1990), LLSs are composed of cognitive strategies, metacognitive strategies, social strategies, and affective strategies. However, Oxford (1990) depicted them as direct strategies that include memory strategies, cognitive strategies, and compensation strategies, and others as indirect strategies composed of metacognitive strategies, social strategies, and affective strategies. Through the consideration of the overlap between cognitive strategies and metacognitive strategies, Macaro (2001) presented the categories of strategies along a continuum of subconscious (or less conscious), based on which, LLSs are divided into direct strategies at one end and conscious and indirect strategies at the other. Cohen (2014) classified LLSs in a new light on the basis of the reasons for using the strategies, for example, strategies for language learning vs. language use, strategies by language skill area, and strategies according to function (namely, metacognitive, cognitive, affective, or social). Despite the different classifications, the recognition of systematic LLSs not only enables scholars and teachers to examine the learning strategies but also helps language learners control their learning and become more proficient in language learning.

As well as LLSs having been classified by some, scholars have also pointed out the important roles that memorization strategy plays in FLL. Therefore, attempts have been made to identify the specific memorization strategies and categorize them to present a system of memorization to help EFL learners improve their learning outcomes. Memorization, one of the direct learning strategies that Rubin (1981) classified, includes four subsets: take notes of new items, pronounce out loud, find an association, and use other mechanical devices. Oxford (1990) grouped memory strategy into four categories: creating

mental links, applying images and sound, reviewing well, and employing action, and subdivided them into 10 specific strategies. The strategy for formally committing the materials that are not acquired naturally through exposure to memory is included in the language learning strategies identified by Cohen (2014). Since memorizing or storing the elements of a new foreign language is the foundation for enlarging learners' knowledge and paving the way for FLL, it is recognized as the initial key to FLL. However, more recent studies on the classification of memorization can rarely be found. To enable language learners to be aware of and discover the specific memorization strategies that suit them best in their language learning, more studies on memorization classifications need to be conducted.

## 2.2. Memorization, understanding, and rote learning

Rote learning is traditionally defined as memorization based on repetition without understanding. Therefore, once rote learning or memorization is adopted, learners are described as rote memorizers, passive and unproductive rote learners, or low-level strategy users. In particular, language learners with Asian backgrounds such as Chinese, Japanese, and Korean have long been viewed and labeled as rote memorizers (Mok et al., 2001; Mathias et al., 2013), which means they are inclined to include memorization as a major strategy for language learning. This concept of the connections between memorization and rote learning is common mainly in the West, and it prompted Marton et al. (1996) to put forward the "paradox" that heavy reliance on memorization makes Chinese learners successful language learning achievers, which is later explained by the findings of the studies that classify and redefine memorization. Based on the learners' intention to understand or the lack thereof, Mugler and Landbeck (2000) identified two meanings of memorization: rote learning, which implies a lack of understanding, and memorization, which implies understanding. Moreover, in terms of sequential order between the two processes, three patterns were found: understanding, then memorization; memorization, then understanding; and a combination of both. A similar exploration into memorization revealed three memorization molds associated with understanding (Marton et al., 2005): rote memorization, in which memorization precedes understanding (Hess and Azuma, 1991); meaningful memorization 1, in which memorization succeeds understanding (Kember and Gow, 1990); and meaningful memorization 2, in which memorization and understanding are seen as simultaneous and combined in the learning process (Kember, 1996). Thus far, it has been proven by the previous documents that the employment of a memorization strategy does not necessarily mean that language learners learn only in a mechanical way when understanding is engaged in the learning procedure. Therefore, memorization strategy, as one of the indispensable parts of FLLs, needs to be considered and interpreted in a different light.

## 2.3. Vocabulary memorization and text memorization

Previous research is primarily concerned with two types of memorization strategies according to the length of material to memorize: strategies to memorize smaller linguistic units of language, which is termed "vocabulary memorization," and strategies to memorize larger linguistic units of language (e.g., complex material of a consecutive text, including sentences, paragraphs, and full essays), which is termed "text memorization." Much research is focused on the former, and a substantial body of studies has been conducted to explore a range of strategies employed by EFL learners to memorize vocabulary (Oxford, 1990; Klapper, 2008; Al-Qaysi and Shabdin, 2016). Furthermore, a number of memorization strategies have been designed purposely and suggested for the instruction of EFL learners on vocabulary learning (Abbasi et al., 2018; Badr and Abu-Ayyash, 2019). As many scholars (Schmitt, 1997; Takač, 2008; Sinhaneti and Kyaw, 2012) have found, a large group of subsets of memory strategies identified so far, for example, memory strategies by Oxford (1990), are effective vocabulary memory strategies, which are believed to be closely associated with only vocabulary learning and have become part of vocabulary learning strategies. Therefore, when memorization as a language learning strategy was accounted for in many previous studies, memorization was generally used to refer to strategies applied by language learners to commit vocabulary to memory, with little consideration of text memorization.

Though text memorization is widely accepted by EFL learners with Asian cultural backgrounds as an efficacious language learning strategy, little research has been conducted on it. A few studies have revealed that text memorization facilitates FLL in many respects, such as vocabulary, grammar, structure, and language skills. Based on an interview with three winners of national English-speaking competitions or debate tournaments in China, Ding (2007) reported that the practice of text memorization facilitates FLL through the enhancement of noticing and rehearsal because collocations and sequences can be learned and then borrowed for productive use. Moreover, the habit of tending to details of language in the context of language input was developed. This conclusion was echoed by another study (Dai and Ding, 2010), which found that text memorization exerts a positive influence on EFL learners' language proficiency and writing performance due to the accuracy and variation of formulaic sequences used. Through collecting data from a group of Chinese learners and teachers ($N = 62$) from 15 middle schools and universities, Yu (2013) concluded that the employment of text memorization not only contributes to the improvement of learning of vocabulary, phrases, sentence structures, grammar, and language skills such as writing and speaking but also affords psychological satisfaction, which is built on EFL learners' sense of achievement and confidence. Compared with the interview-based studies that have clarified the manifold benefits that EFL learners gained from text memorization, more recent research by Harris (2015) proposed a systematic pattern for the memorization of a story or dialog, which is advantageous to language learning, known as a top-down and bottom-up pattern. The top-down mold, requiring examination of the overall content of the text, consists of three steps: start by understanding the main idea of the entire text, then break the text down into manageable sections to understand the main idea of each, and finally, analyze each sentence for the general content and main idea. The bottom-up mold involves careful analysis of language elements at the lexical level and how the meaning was created by the combination of words.

In summary, previous literature has suggested that text memorization as an effective foreign language learning strategy has a positive impact on EFL learners' overall English proficiency (i.e., Ding, 2007; Dai and Ding, 2010; Yu, 2013), but few studies have been

conducted to reveal the relationship between text memorization and EFL learners' writing proficiency. Furthermore, it has been revealed in previous studies that understanding, which is involved in the process of memorization, makes EFL learners with Asian cultural backgrounds active learners rather than rote memorizers. However, little is known about how text memorization proceeds when understanding is engaged in the process of text memorization. In addition, the majority of existing research is centered on the strategies of vocabulary memorization, and few studies have attempted to present a system for text memorization strategies that EFL learners apply in their processes of memorizing complex and consecutive text. Therefore, the findings of this study fill the gap by answering three questions:

1. Does text memorization as a foreign language learning strategy affect the improvement of EFL learners' argumentative writing proficiency?
2. When Chinese EFL learners employ text memorization strategies, is understanding involved in their memorization process? If yes, what is the text memorization process of these successful memorizers?
3. What specific strategies do the successful Chinese memorizers who learn English as a foreign language employ to memorize the texts, and how could these strategies be classified to formulate a system for text memorization?

# 3. Method

## 3.1. Design of the study

Mixed-method research was conducted, in which both quantitative and qualitative methods were used. In the majority of educational settings, random assignment of students is hardly possible, thus, quasi-experiments are often conducted to create the comparisons, from which treatment-caused change is inferred (Zoltán, 2007). Therefore, in the present study, quasi-experiments were designed to answer the first question about whether text memorization could help improve Chinese EFL learners' writing proficiency with the application of quantitative research methods. The experimental sample involved one intact university class. A pre-test that required this class group to produce argumentative writing was implemented to assess their writing proficiency before the treatment. Afterward, this group of students was given a text to memorize each week (seven in total) and was required to report the full text in a memorization test in the following week; thus, their text memorization effects were examined by a total of seven memorization tests. Finally, a post-test was also administrated to this group of students, requiring them to produce another piece of argumentative writing. The post-test was targeted at evaluating their writing proficiency after the treatment. Then a comparison between students' writing performances in the pre-test and post-test was made to discover whether students' writing proficiency was improved after the treatment. To answer the second and third questions, the qualitative research method was adopted. After all the tests were completed, semi-structured interviews were conducted with the 12 students who performed best in the text memorization tests to gain information about these successful memorizers' text memorization processes and the specific text memorization strategies involved, based on which, a system for text memorization was attempted to be formulated.

## 3.2. Participants

Convenience sampling was applied to include one intact class group that consisted of 33 students as participants (29 women and four men) who were in their second year of studies at one of the universities in China. The participants were accepted as English majors by the university through the national college entrance examination and were between 19 and 20 years of age. The author had taught them *Intensive Reading* for more than half a year. When the research was conducted, the participants had been learning English as their major for approximately one and a half years by taking a variety of professional English courses designed for college English majors, such as *Intensive Reading*, *Extensive Reading*, and *Grammar*. This meant that the participants were on the way to becoming qualified English majors. All the participants were informed of the purpose of the study and agreed to participate in the study, but they were free to withdraw at any time. At the time of data analysis, all the participants ($n = 33$) were included.

## 3.3. Instruments

### 3.3.1. Text memorization tasks and text memorization tests

Seven argumentative writings were selected as model writings (refer to Appendix A) for students to memorize. Two university English teachers were invited and, together with the author, examined the quality of seven selected writings by discussion and reached a consensus that they could be used as model writings. To select model writings, two major aspects were concerned: the quality of the writings and the topics of the writings. The criteria to ensure the high quality of the model writings lay in such aspects as lexical resources, grammatical structures, accuracy, coherence and cohesion, and idea development. The topics of the writings were to be diverse and closely associated with life experiences that students are familiar with. The topics of the finally selected seven model writings ranged from technology and life (model writing 1), campus life (model writings 2, 5, 7, and 8), and family issues (model writing 3), to environmental issues (model writing 4) and social issues (model writing 6). Students were given 1 week to memorize a model writing, and then a memorization test was conducted. Students' text memorization quality was examined by requiring them to recall and verbatim write out the texts of the model writings without referring to the model writings and without the use of dictionaries. In total, seven memorization tests were administrated within 7 weeks, one for each week.

### 3.3.2. Pre-test and post-test

In the pre-test, the writing task (see Appendix C) from the 2017 Test for English Majors-Band 4 (TEM 4) was used to assess the participants' writing proficiency before the task of memorizing the seven model writings was undertaken. In the post-test, the one from 2019 (refer to Appendix C) was used for the evaluation of students' writing performance after the memorization work was done. There were two reasons why the two writing tasks of TEM 4 were used in the

present study. First, TEM 4 was a written test designed by the National Education Committee of People's Republic of China and used to assess the English proficiency of Chinese English majors in their second-year studies at the university since 1991. The students who participated in the present study were all sophomores majoring in English. Therefore, the previous writing tasks of TEM 4 were appropriate to be used to evaluate the participants' writing proficiency. Second, the topics of the writing tasks of TEM 4 covered education, university life, society, etc., which Chinese college students are familiar with. This would minimize the negative influences that the lack of topic knowledge may have exerted on the students' writing performances.

### 3.3.3. Semi-structured interview

By using both qualitative and quantitative approaches, the best of both paradigms (quantitative and qualitative research) could be deployed (Zoltán, 2007). Many researchers have developed qualitative studies as the support and supplement for quantitative research to explore EFL learners' learning behavior, for instance, the use of LLSs (He and Shi, 2008; Jiang and Smith, 2009; Yu, 2017). In view of research questions 2 and 3, interviews with participants about their experiences of text memorization were used as a logical approach in the present research to triangulate the quantitative data in a broad sense. The top 12 students in the score list for the memorization of seven model English argumentative writing tests were chosen to be interviewed. The interviews were semi-structured with two major interview guide questions (refer to Appendix E) developed by the author with reference to the interview guild questions in the study by Jiang and Smith (2009), which were aimed to detect Chinese learners' strategies used in English learning. The first interview question of the present study was focused on the participants' text memorization processes of the model writings and the specific text memorization strategies involved. The second question was aimed at collecting information about students' attitudes toward text memorization strategies. The reliability of the interview data was enhanced in four ways. First, all the interviewees were students whom the author had taught for more than half a year; therefore, a good relationship established between them enabled the interview to proceed in a more natural and conversational manner. Second, to ensure that participants had enough time to recall their memorization experiences and provide authentic information about their memorization procedures, the students were informed of the interview questions 1 day before the interview. Third, sub-questions such as "*What learning strategies did you use to help you commit the model writings to memory?*" and "*In what aspects do you think text memorization is effective to improve your English argumentative writing?*" were designed to encourage and assist students in providing more details about their memorization process. Fourth, the individual interviews were conducted in Chinese instead of English so that the students could express their ideas more clearly in their native language. The interviews were conducted with flexibility, and a group of interview techniques suggested by Zoltán (2007), such as carry-on feedback and encouraging elaboration, were employed by the author. For instance, when one interviewee reported, "I'm able to read the text fluently," the interviewer raised the probe question or follow-up question, "*Do you mean you need to read it out loud or not?*" to both confirm the intended meaning of the utterance and increase the richness and depth of the response. In addition, the wording of the prepared questions was sometimes geared to better suit the different personalities of the interviewees or make the interviews more natural.

## 3.4. Data collection and analysis procedures

Before the study, all the participants were informed of the purpose of the study, the roles that participants play in data collection, and the confidential and voluntary nature of the research. On the Monday of the first week of the research, a pre-test was conducted, and model writing 1 was given to students to memorize. On the Monday of the second week, memorization test 1 was administrated, and model writing 2 was handed out. On the Monday of the third week, memorization test 2 was administrated, and model writing 3 was handed out. Such a process was repeated so that within 8 weeks, seven text memorization tests had been conducted. In the ninth week, after all the memorization tests had been completed, the post-test was administrated. Two experienced university English teachers were given 2 weeks (weeks 10 and 11) to rate students' performances in all memorization and writing tests. The evaluation criterion (see Appendix B) for memorization tests was code-signed by them to make the rating consistent. The reliability of the two raters was acceptable according to the Pearson correlation coefficient (test 1, $r = 0.91$; test 2, $r = 0.86$; test 3, $r = 0.90$; test 4, $r = 0.88$; test 5, $r = 0.91$; test 6, $r = 0.92$; test 7, $r = 0.88$). Thus, the average score achieved by each student was used as the final score, which indicates each student's memorization effect of seven model argumentative writings. To ensure the reliability of the pre-test and post-test, the scoring criterion (see Appendix D) was also implemented by both two university English teachers. The interrater reliability was acceptable for the present study according to the Pearson correlation coefficient (pre-test, $r = 0.67$; post-test, $r = 0.64$). Therefore, the average score given by the two raters to each student was used as the final score, representing each student's writing proficiency. Then, descriptive statistics, such as means and standard deviations of the participants' writing proficiency before and after completion of the memorization tasks, were reported and interpreted. The paired samples *t*-test was adopted to examine whether participants' writing proficiency was improved after the text memorization assignments were accomplished. Pearson correlation coefficients were used to represent the relationships between participants' performance in the memorization of seven model writings and their writing proficiency after the text memorization tasks were completed.

The semi-structured interviews were conducted in the office of the English department in the ninth week by the author after all the experimental tests had been completed. Each participant agreed to the interview being recorded by smartphone for data analysis for the study. All the interviews were then transcribed and translated into English by the author. The transcribed and translated texts were double-checked, during which the author sought validation from the interviewees, through smartphone conversations, when ambiguity arose. Then the translated texts were analyzed to gain information about the participants' text memorization processes (including whether understanding was involved) and the specific text memorization strategies used to enhance their text memorization effects. More importantly, based on the analysis, varieties of specific text memorization strategies were classified to formulate a system for text memorization. In the process of analysis, to ensure the accurate interpretation of the interviewees' viewpoints, exchanges through smartphone or WeChat voice calls were conducted for confirmation when necessary.

# 4. Results and findings

## 4.1. Comparison of students' writing performances between pre-test and post-test

Descriptive statistics of the pre-test and post-test shown in Table 1 demonstrate that the mean scores of the post-test, M = 12.21, SD = 0.48, are improved compared to those of the pre-test (M = 10.64, SD = 0.54). The results of the paired-samples *T*-test suggest that there is a statistically significant difference between the mean scores of pre-test and post-test, $t(32) = -3.88$, $p = 0.00 < 0.05$, indicating that students' argumentative writing proficiency was greatly improved after the memorization tasks of seven English argumentative writings had been completed.

## 4.2. Relationship between memorization effect of seven model writings and writing performances after the completion of text memorization tasks

The Pearson correlation coefficients in Table 2 show a significant positive relationship between participants' performances in text memorization of seven model writings and their writing proficiency after the completion of the text memorization tasks, r = 0 0.63, $p < 0.01$, which indicates that the better the achievements made in text memorization of seven model writings, the higher the proficiency in English argumentative writing.

## 4.3. Strategies used for memorization of the model writings and students' perspectives

### 4.3.1. Strategies used for understanding that facilitates memorization

All the interviewees except for interviewees 2 and 10 emphasized the importance of understanding before they memorized the texts, but the strategies used to understand were reported to be differential. Two of them (1 and 8) mentioned that translation of the whole text into Chinese was done prior to memorization to help them understand the content and main ideas of the writings. Judging from the tone of their utterances, the claim of such a way for memorization seemed awkward to them, presumably due to it being time-consuming or other unknown reasons. However, it was a personal and satisfying approach that they appreciated.

Another practice that many of the interviewees, including interviewees 3, 4, 5, 8, 11, and 12, developed was to read through the texts to get the main idea of the model writings. Afterward, they would seek assistance from the dictionary if the words and phrases became barriers preventing them from understanding the texts. "On Wednesday morning, when I have no classes from 10 am to 12 pm [*sic*], I will read through the text and look up the words I do not know in the dictionary. The initial things for memorization like that." After recalling her first step, the use of a dictionary for memorization of the model writings, interviewee 4 continued to state that she did not favor learning by rote. Underlining the primary significance of understanding with strong opposition to rote memorization, she said, "It is not good to start to memorize without understanding the texts, for example, mechanical memorization or rote memorization." Similarly, interviewee 5 attempted to understand the text when the memorization task started, but unlike interviewees 1 and 8, the frequent method used for understanding was not translating but repeated reading. "I do not intend to translate the texts. When I read them over and over, I can understand the ideas that the texts convey," she said.

Compared with the majority of the interviewees' notions that understanding precedes memorization, interviewee 10 reported that she adopted mechanical memorization by two mechanical means: storing the words' locations and typing the model writings. She explained:

> "My approaches to memorizing the texts are a bit mechanical. I have a unique method. Unlike others, I often memorize the specific location of a word on the writing handed to me, and it can be stored in my brain. When I forget the content, I can recall where the word is located. There is another good method, which seems a bit time-consuming. I type the whole model writing using the computer. This method helps me memorize it faster, but it is also easier to forget because I don't fully understand it."

### 4.3.2. Memorizing through text analysis

Among the strategies adopted, another frequently used approach was to analyze the structure of complicated sentences to facilitate memorization with the grammatical knowledge acquired. The normal practice that the majority of these successful memorizers deployed, not excluding interviewees 1, 3, 4, 6, 7, 9, and 11, was to identify the subject, the predicate, and the object of a long sentence that was formed with more linguistic elements, such as phrases and clauses. This process is, in fact, the analysis of the hierarchal order of a sentence, which was conducted with intention of breaking down complicated sentences into smaller units to ease the process of memorization. By being aware of the subject (either a noun, a noun phrase, or a nominal clause) and the predicate (a verb or a verb phrase), the efforts made resulted in the linguistic portrayal of a fundamental tree diagram. However, as second-year students, they were not aware of that tree diagram since they had not yet attended any courses relevant to linguistics or syntax. The conspicuous objective

TABLE 1 Paired-samples *T*-test for pre-test and post-test.

| | N | M | SD | T | DF | Sig. |
|---|---|---|---|---|---|---|
| Pre-test | 33 | 10.64 | 0.54 | −3.88 | 32 | 0.00 |
| Post-test | 33 | 12.21 | 0.48 | | | |

TABLE 2 Pearson correlation coefficients for text memorization test of seven model writings and post-test.

| | Memorization test | Post-test |
|---|---|---|
| Memorization test | 1 | |
| Post-test | 0.63** | 1 |

**$p < 0.01$, two-tailed.

of such analysis, as they explained, was to make their memorization tasks easier.

> "I often analyze the main structure of the sentences. A seemingly long sentence would not be complicated anymore if the subject, the predicate, and the object are [sic] detected … The rest part [sic] of a sentence would be modifiers like adjectives. This is the way I do to [sic] simplify a complicated sentence" (interviewee 1).

In addition to sentence structure analysis, another skill that interviewees 6 and 3 mentioned was writing an outline for each selected writing, which was firmly believed to offer intelligible guidance to them to recall the contents and opinions of the writings. Their inceptive attention was attached to the organization of the writing and the ways by which the ideas developed. In other words, when the entire stages of memorization were examined, analysis and storage of the outline always predated that of the language details. Interviewee 3 outlined the given model writing by clarifying the main ideas, analyzing the supporting details, and picking up on the keywords. She acknowledged that such a method worked well for the memorization of the whole essay. The importance of outlining the writings was echoed by interviewee 6, describing how an outline is produced:

> "The awareness of the organization of the whole writing makes the memorization tasks easier. An introduction of the topic discussed is often made in the beginning, which is followed by one argument of the topic. Then the author's thesis statement is put forward. In the main body, the core notion is often analyzed from two respects. Finally, the conclusion is made. My focus is always on the keywords and linking words like however, therefore, etc., so that an outline of the writing can be drawn. I memorize them based on the outline afterward."

#### 4.3.3. Storage strategies for memorization

Recitation, a widely accepted L1 learning strategy in China to increase the effect of memorization, was often employed by the interviewees. Seven of them, interviewees 1, 2, 4, 5, 8, 9, and 11, reported that reciting the texts was one of the necessary channels that assisted them in committing the texts to memory. Two forms of recitation were reported: audibly (loudly or in a low voice) and silently, according to the places where they recited the texts. Choosing this memorization strategy was a personal preference depending on individual learning styles and characteristics.

To be able to memorize the whole essay, interviewees 1 and 2 developed the habits of reading and repeatedly reciting aloud, the value of which was emphasized by interviewee 2, who said, "Compared with reading and reciting silently, the memorization effect through reading and reciting loudly and repeatedly is much better." However, in addition to reading and reciting loudly, reading and reciting softly was also acceptable to interviewee 5 because reading and reciting by heart did not enhance their memorization. Such an audible approach to reading and reciting, either loudly or softly, was shared by interviewee 9, who emphasized that when in the library, she would retrieve the texts from her mind and write them out instead of uttering any sound. Interviewee 4 said that she read and recited silently in the library since audible reading and recitation would disturb others and were not allowed. However, she added, "If the learning environment permits, I often read as loudly as what I have done in my middle school during the time for morning reading," (a period of time, approximately 20–30 min during the morning of each school day for middle school students to read either Chinese or English texts loudly in China).

The depictions of the interviewees' recitation procedures revealed that reading the texts, either aloud or in a low voice, was a commonly used approach that was incorporated into the process of their recitation. Students intended to read out so that they could hear the contents of the texts, thus enhancing the memorization effect. Followed by audible reading, reciting to themselves audibly or silently reinforced their memorization. The unavoidable forgetfulness urged students to refer to the texts either by reading loudly or silently and then the texts would be recited again. Such a reading–reciting cycle would be repeated several times until they were conscious that the texts had been memorized. Therefore, repetition was taken as one of the indispensable strategies for strengthening the memorization effect. Moreover, reviewing on a daily basis or at intervals of several days was verbally reported by interviewees 1, 2, 3, and 5.

#### 4.3.4. Students' perspectives on the impact of text memorization on English writing proficiency

All the students interviewed valued the impact of memorizing the model writings on improving their writing performances. The majority stressed that the benefits gained from it were extremely positive. The enhancements can be illustrated in three aspects of argumentative writing. The most conspicuous effect highlighted by interviewees 1, 2, 3, 4, 5, 7, and 12 was that memorization of the model writings enabled them to acquire the ability to construct a piece of English argumentative writing and develop the ideas in a coherent manner. For instance, interviewee 1 claimed that the regular structure of the argumentative writing and thinking patterns were learned as they accomplished the memorization assignments. Another advancement was that the students learned diversified sentence patterns and extended their vocabulary. For example, interviewees 4 and 12 revealed that their previous argumentative writings were all completed using the translation mold; that is, a Chinese sentence was generated in their minds first and was then translated into a simple English sentence. More often than not, the sentence was organized with improper diction and grammar that caused ambiguity and confusion. With more model writings memorized, this translation mold used for English writing was gradually abandoned on the grounds that native-like sentences could be produced with the application of more advanced words, correct collocations, and complicated sentence patterns.

## 5. Discussion

### 5.1. Impact of text memorization on EFL learners' argumentative writing proficiency

The present study examines the impact of text memorization as a foreign language learning strategy on the improvement of EFL students' argumentative writing proficiency. The results revealed that text memorization of the model essays is significantly effective in enhancing the argumentative writing proficiency of Chinese EFL learners, which is consistent with the previous finding (Dai and Ding,

2010) that the adoption of text memorization in FLL contributes to greater progress in EFL learners' writing ability. At the lexical level, the memorization of full text enables EFL learners to store a string of words that are formed in a fixed sequence, termed "formulaic sequences" (Schmitt, 2004), thus reducing the possibility of EFL learners making grammatical errors (Boers and Lindstromberg, 2008). If the knowledge of FSs is absent, EFL learners depend on their L1 knowledge to possibly combine irrelevant words to generate incorrect collocations in English argumentative writing. As a result, negative transfer occurs with the production of deviant L2 combinations (Ellis, 2008). However, with the accumulation of these prefabricated patterns stored in the mind through text memorization, students are able to retrieve such chunks or multi-word units easily and put them to use, with no need to arrange them through word selection and grammatical sequencing (Tremblay and Baayen, 2010). In terms of the syntactical and textual level, when learners memorize whole texts with the help of a series of text memorization strategies, especially the syntactical and textual analysis strategy, the relevant sentence structures and the arrangements of the argumentative essays can be retrieved and, more importantly, imitated by EFL learners to produce better writings of their own. This finding is supported by previous studies, suggesting that with the recognition of how information in texts is packaged and organized through text analysis, more cohesive and coherent writings can be composed (Tovar, 2016). Text analysis is not intended to encourage copying but rather to promote awareness of style and writing subskills. In short, in the process of text memorization, both small and large linguistic forms are noticed by EFL learners and stored in their short-term memory. When successive text memorization helps to transfer short-term memory into long-term memory, there is a greater likelihood that EFL learners can retrieve the retained information or knowledge related to the lexicon, syntax, and textual organization to generate better-structured English sentences. Therefore, their arguments on a variety of the given topics can be clearly presented for communication in the form of writing. In other words, it is the memory-based system that enables EFL learners to access and deploy chunks of language, on which language fluency depends (Skehan, 1995).

## 5.2. Text memorization process of Chinese EFL learners

When it comes to memorization of larger linguistic units of a foreign language, i.e., full texts, the current study showed that Chinese EFL learners' memorization process is more complex than memorization merely by repetition, as it has previously been described. The memorization process, including memorization strategies, is presented in Figure 1. When memorization starts, mechanical memorization is possibly adopted in the absence of understanding. However, if understanding is involved, such strategies as using a dictionary, translating, and reading are used for comprehension of the model writings first. Then the structures of the whole writing and sentences are analyzed to facilitate memorization. Afterward, audible or silent reading and reciting are repeated by language learners to store the texts. In this process, the reviewing strategy helps learners return to the procedure of understanding to deepen their comprehension, which further strengthens memorization.

In the present study, two types of memorization practices were identified for storing and recalling the texts reliably and retrieving them easily. The less common was memorizing mechanically without the engagement of understanding, whereas the most common, which was also discovered in the studies of Hess and Azuma (1991) and Kember and Gow (1990), was giving priority to understanding, followed by memorization. The third type, the integration or combination of memorization and understanding explored by Mugler and Landbeck (2000) and Marton et al. (2005), was not reported by the participants in this study, though the author assumed that it existed. With regard to understanding, it was found that the most frequent and effective strategy employed by Chinese EFL learners was a top-down model. This mold shares similarities with the top-down approach to facilitating text memorization in the study of Harris (2015), which reported three steps for text memorization: (1) understanding the main idea of the entire text, (2) analyzing the structure of the full text, and (3) analyzing the structure of sentences. Since the top-down approach and the bottom-up approach could complement each other (Davies, 1988), the bottom-up mold based on the careful breakdown of sentences with a focus on lexical and syntactic forms was also found in this study. Following this pattern, EFL learners can be informed of the formal and content schemata for the construction of the essay on the top level, as well as the lexis and syntactic forms that realize it at the sentence level.

## 5.3. Text memorization strategies by Chinese EFL learners

The present study found that in the process of the memorization of the given model writings, all of the successful Chinese memorizers actively employed a range of strategies to improve their memorization effect, which were concluded as the text memorization strategies. These strategies worked together to contribute to the storage, retention, recall, or retrieval of the texts that Chinese EFL learners

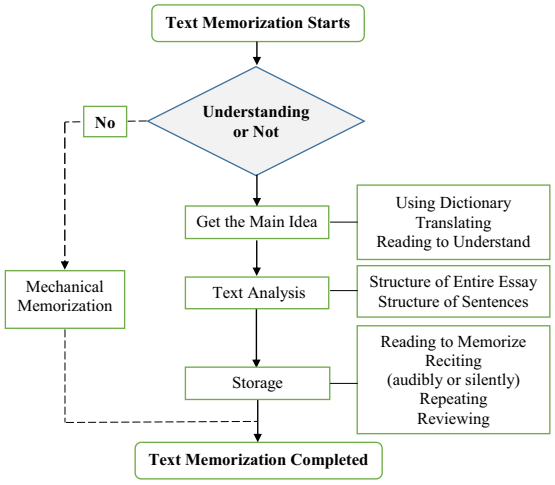



FIGURE 1
Diagram of text memorization process and strategies used by Chinese EFL learners (source: original).

tended to memorize. A new strategy for text memorization used by Chinese EFL learners was developed based on the findings of the present study and is shown in Figure 2.

Considering whether understanding is involved or not, text memorization strategies can be divided into two major classes: mechanical strategies and non-mechanical strategies. The term, mechanical strategies, is preferred to generally describe one of the intentional approaches of text memorization. The evidence obtained in the current study indicates that Chinese EFL learners do not just rely on repeating, they also seek out other strategies, such as keeping in mind the word's location in the writing and typing the text on a computer. Though these strategies are believed to be mechanical, they are different from conventional rote memorization, referring to memorization by mere repetition in the previous study (Lazaric, 2012). Therefore, two groups: storing the word's location and typing the text, were discovered and grouped into mechanical strategies. The non-mechanical strategies consist of understanding strategies for text comprehension that facilitate memorization and storage strategies for remembering and retrieval of the texts. Understanding strategies fall into four categories: using a dictionary, translating, reading to understand, and text analysis including structural analysis of the entire writing and structural analysis of the sentence. Storage strategies are made up of four other categories: reading to memorize, reciting (audibly or silently), repeating, and reviewing. It should be noted that two forms of reading strategies were identified according to the goals that Chinese EFL learners tend to reach. The connotation of "reading to understand" is the same as that of "reading" in the phrase "reading a novel," which means to comprehend the contents. However, "reading to memorize" is defined as the strategy of reading the texts audibly or silently with the intention of retaining the texts in memory.

There is no consensus on how strategies are defined and categorized (Takač, 2008), thus, classification conflicts do exist. This study suggests that translating and analyzing are two strategies for understanding that fall under the text memorization strategy. However, both of these types were considered cognitive strategies in the study of Oxford (1990). Similarly, repeating, a cognitive strategy in Oxford's taxonomy of strategies, is viewed as a set of the text memorization strategy in this study. Such different classifications could be explained in two respects: (1) Previously, memorization strategies were investigated as the major strategies for memorizing vocabulary, the smaller units of language. Therefore, a shift in concern from memorization of vocabulary to memorization of full texts leads to the distinct classifications of specific memorization strategies. (2) The roles that a strategy plays in different learning phases impact the divisions of the strategies. For example, when translating is seen as a tool for manipulation or transformation of the target language by the learner, it is recognized as a cognitive strategy. However, in the present study, translating is identified as a sub-strategy of text memorization because it is deployed for the purpose of storage. It is understandable that when the role of a certain strategy changes in a given task in language learning, it can be sorted into two groups simultaneously, which are meant to overlap (Oxford, 1990). More recently, emphasizing the free and flexible operations of LLSs, Oxford (2017) proposed that while a strategy might be thought of as being in a certain group, other potential functions of that strategy should also be seriously considered. Despite the unclear boundaries between the categories of some strategies, the system of text memorization strategies developed in the current study could still present a wide-ranging framework to both examine text memorization strategy that many Chinese EFL learners accepted as a means to underpin their text memorization outcome and extend our understanding of it.

## 6. Conclusion and limitations

The current study contributes to the identification of the relationship between LLS use and language proficiency, particularly the relationship between the use of text memorization as a foreign language learning strategy and English argumentative writing proficiency by presenting further evidence from a Chinese FLL context. The results suggest a significantly positive relationship between text memorization and Chinese EFL learners' writing outcomes. It is concluded that text

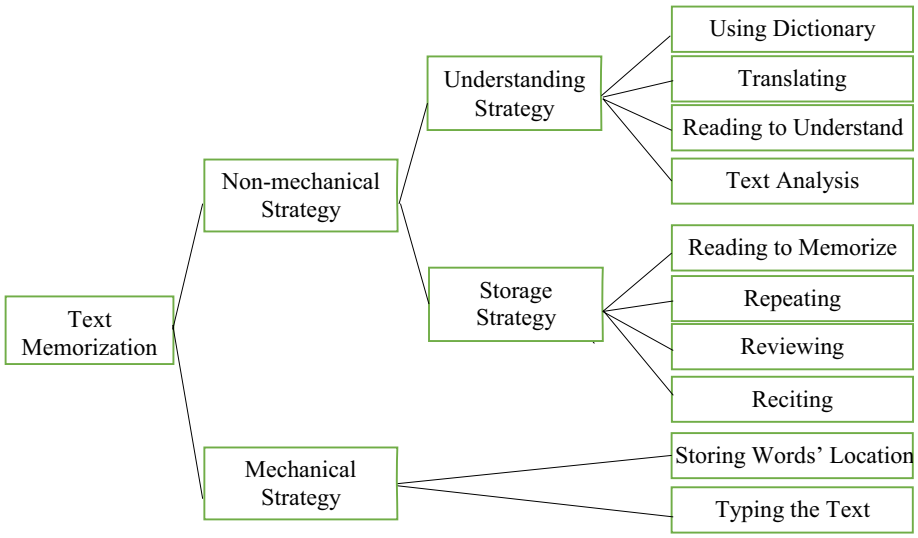



FIGURE 2
Diagram of text memorization strategies used by Chinese EFL learners (source: original).

memorization as a foreign language learning strategy is effective in improving EFL learners' writing proficiency, which previous literature on language learning strategy use in general settings and specific language skill areas has seldom documented. Moreover, this study is the first attempt to explore and develop a new system of text memorization strategies based on the analysis of the text memorization processes and strategies deployed by Chinese EFL learners. The suggestion on the differentiation of memorization of larger units of language, such as the full text of an essay, from memorization of smaller units of language, such as vocabulary, provides new insight into memorization strategy research within the field of LLSs.

Although this study detected a variety of strategies used for text memorization, which have been proven to work effectively in improving Chinese EFL learners' writing proficiency, other strategies to increase the text memorization effect are likely to offer a more comprehensive interpretation of Chinese EFL learners' text memorization strategy use. Future studies should focus on the discovery of these strategies and how these strategies are involved in the process of Chinese EFL learners' text memorization processes. Moreover, a multitude of independent variables potentially impact strategy use (Grainger, 2012), thus leading to differential language performance. Gender, in this study, could have played a role because 29 out of 33 subjects were women. In addition, this study was conducted in China, with Chinese EFL learners as the objects, so when EFL learners' cultural backgrounds or socio-ecological contexts change, whether the strategy used and the impact of text memorization on EFL learners' writing proficiency would differ is anticipated to be uncovered in future studies.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

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

## Funding

This study was supported by the 2019 Social Science Foundation of Shaanxi Province, China (2019M006).

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1126194/full#supplementary-material

## References

Abbasi, A., Hassaskhah, J., and Tahriri, A. (2018). The effect of teaching memory strategies on Iranian EFL learners' vocabulary retention in terms of learners' multiple intelligences. *Int. J. Educ. Lit. Stud.* 6, 1–9. doi: 10.7575/aiac.ijels.v.6n.2p.1

Al-Qaysi, F. H., and Shabdin, A. A. (2016). Vocabulary memorization strategies among Arab postgraduate English foreign language learners. *Adv. Lang. Lit. Stud.* 7, 184–196. doi: 10.7575/aiac.alls.v.7n.5p.184

Badr, H., and Abu-Ayyash, E. S. (2019). Semantic mapping or rote memorisation: which strategy is more effective for students' acquisition and memorization of L2 vocabulary? *J. Educ. Learn.* 8, 158–170. doi: 10.5539/jel.v8n3p158

Biggs, J. B. (1996). "Western misconceptions of the Confucian-heritage learning culture" in *The Chinese Learner: Cultural, Psychological and Contextual Influence*. eds. D. A. Watkins and J. B. Biggs (Hong Kong: Comparative Education Research Centre), 45–67.

Boers, F., and Lindstromberg, S. (2008). "From empirical findings to pedagogical practice" in *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology*. eds. F. Boers and S. Lindstromberg (Berlin: Mouton de Gruyter), 375–393.

Chan, C. K. K., and Rao, N. (2009). "Moving beyond paradoxes: understanding Chinese learners and their teachers" in *Revisiting the Chinese Learner: Changing Contexts, Changing Education*. eds. C. K. K. Chan and N. Rao (Hong Kong: Springer), 3–32.

Cohen, A. D. (2014). *Strategies in Learning and Using a Second Language. 2nd Edn*. London: Routledge.

Dai, Z., and Ding, Y. (2010). "Effectiveness of text memorization in EFL learning of Chinese students" in *Perspectives on Formulaic Language: Acquisition and Communication*. ed. D. Wood (London: Continuum), 71–87.

Dansereau, D. F. (1985). "Learning strategy research" in *Thinking and Learning Skills: Relating Learning to Basic Research*. eds. J. W. Segal, S. F. Chipman and R. Glaser (Hillsdale NJ: Erlbaum), 209–240.

Davies, F. (1988). "Designing a writing syllabus in English for academic purposes: process and product" in *Academic Writing: Process and Product, ELT Documents 129*. ed. P. Robinson (London: Modern English Publications/British Council), 130–142.

Ding, Y. (2007). Text memorization and imitation: the practices of successful Chinese learners of English. *System* 35, 271–280. doi: 10.1016/j.system.2006.12.005

Ellis, R. (2008). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Fareed, M., Ashraf, A., and Bilal, M. (2016). ESL learners' writing skills: problems, factors and suggestions. *J. Educ. Soc. Sci.* 4, 83–94. doi: 10.20547/jess0421604201

Grainger, P. (2012). The impact of cultural background on the choice of language learning strategies in the JFL context. *System* 40, 483–493. doi: 10.1016/j.system.2012.10.011

Griffiths, C., and Soruç, A. (2020). *Individual Differences in Language Learning*. Cham, Switzerland: Springer Nature Switzerland AG.

Harris, T. M. (2015). Text and dialogue memorization in English language learning. Available at: http://id.nii.ac.jp/1338/00001242/ (Accessed March 2021).

He, L., and Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing tests. *Assess. Writ.* 13, 130–149. doi: 10.1016/j.asw.2008.08.001

Hess, R. D., and Azuma, M. (1991). Cultural support for schooling: contrasts between Japan and the United States. *Educ. Res.* 20, 2–9. doi: 10.3102/0013189X020009002

Hyland, K. (2003). *Second Language Writing*. Cambridge: Cambridge University Press.

Jiang, X., and Smith, R. (2009). Chinese learners' strategy use in historical perspective: a cross-generational interview–based study. *System* 37, 286–299. doi: 10.1016/j.system.2008.11.005

Kember, D. (1996). The intention to both memorise and understand: another approach to learning. *High. Educ.* 31, 341–354. doi: 10.1007/BF00128436

Kember, D. (2000). Misconceptions about the learning approaches, motivation and study practices of Asian students. *High. Educ.* 4, 99–121. doi: 10.1023/a:1004036826490

Kember, D., and Gow, L. (1990). Cultural specificity of approaches to study. *Br. J. Educ. Psychol.* 60, 356–363. doi: 10.1111/j.2044-8279.1990.tb00952.x

Khamees, K. S. (2016). An evaluative study of memorization as a strategy for learning English. *Int. J. Engl. Linguist.* 6, 248–269. doi: 10.5539/ijel.v6n4p248

Klapper, J. (2008). "Deliberate and incidental: vocabulary learning strategies in independent second language learning. Language learning strategies in independent settings" in *Language Learning Strategies in Independent Settings*. eds. S. Hurd and T. Lewis (Bristol: Multilingual Matters), 159–178.

Lazaric, N. (2012). "Rote Memorization" in *Encyclopedia of the Sciences of Learning*. ed. N. M. Seel (New York: Springer), 2899–2901.

Li, X., and Cutting, J. (2011). "Rote learning in Chinese culture: reflecting active Confucian-based memory strategies" in *Researching Chinese Learners: Skills, Perceptions and Intercultural Adaptations*. eds. L. Jin and M. Cortazzi (Hampshire: Palgrave Macmillan), 21–42.

Macaro, E. (2001). *Learning Strategies in Foreign and Second Language Classrooms*. London: Continuum.

Marton, F. G., Alba, D., and Tse, L. K. (1996). "Memorizing and understanding: the keys to the paradox?" in *The Chinese Learner: Cultural, Psychological and Contextual Influence*. eds. D. A. Watkins and J. B. Biggs (Hong Kong: Comparative Education Research Centre), 69–83.

Marton, F., Wen, Q. F., and Wong, K. C. (2005). Read a hundred times and the meaning will appear. Changes in Chinese university students' views of the temporal structure of learning. *High. Educ.* 49, 291–318. doi: 10.1007/s10734-004-6667-z

Mathias, J., Bruce, M., and Newton, D. P. (2013). Challenging the Western stereotype: do Chinese international foundation students learn by rote? *Res. Post-Compuls. Educ.* 18, 221–238. doi: 10.1080/13596748.2013.819257

Mok, I., Chik, P. M., Ko, P. Y., Kwan, T., Lo, M., Marton, F., et al. (2001). "Solving the paradox of the Chinese teacher?" in *Teaching the Chinese Learners: Psychological and Pedagogical Perspectives*. eds. D. Watkins and J. Biggs (HKU: Comparative Education Research Centre), 161–181.

Mouziraji, A., and Mouziraji, A. (2015). Memorization makes progress. *Theory Pract. Lang. Stud.* 5, 870–874. doi: 10.17507/tpls.0504.25

Mugler, F., and Landbeck, R. (2000). Learning, memorisation and understanding among distance learners in the South Pacific. *Learn. Instr.* 10, 179–202. doi: 10.1016/S0959-4752(99)00026-2

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

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

Oxford, R. L. (2017). *Teaching and Researching Language Learning Strategies: Self-Regulation in Context. 2nd Edn*. New York: Routledge.

Petri, B., and Czárl, B. (2003). Validating a writing strategy questionnaire. *System* 31, 187–215. doi: 10.1016/S0346-251X(03)00020-4

Rashidi, N., and Omid, A. (2011). A survey on Iranian EL learners' beliefs on the role of rote memorization in learning vocabulary and its effect on vocabulary achievement. *J. Pan-Pac. Assoc. Appl. Linguist.* 15, 139–161.

Richards, J., and Platt, J. (1992). *Longman Dictionary of Applied Linguistics*. London: Longman Group Ltd.

Rubin, J. (1981). Study of cognitive processes in second language learning[1]. *Appl. Linguis.* II, 117–131. doi: 10.1093/applin/II.2.117

Schmitt, N. (1997). *Vocabulary Learning Strategies*. Cambridge: Cambridge University Press.

Schmitt, N. (2004). *Formulaic Sequences: Acquisition, Processing and Use*. Amsterdam: John Benjamins.

Sinhaneti, K., and Kyaw, E. K. (2012). A study of the role of rote learning in vocabulary learning strategies of Burmese students. *US China Educ. Rev.* 2, 987–1005.

Skehan, P. (1995). "Analyzability, accessibility, and ability for use" in *Principles and Practice in Applied Linguistics*. eds. G. Cook and B. Seidlhofer (Oxford: Oxford University Press), 91–106.

Sonmez, H. (2018). A review about the use of the memorization strategy during the learning process by students. *Int. J. Lang. Educ.* 1, 212–230. doi: 10.18298/ijlet.2123

Takač, V. P. (2008). *Vocabulary Learning Strategies and Foreign Language Acquisition*. Clevedon: Multilingual Matters.

Tovar, E. V. (2016). Effectiveness of systemic text analysis in EFL writing instruction. *GIST Educ. Learn. Res. J.* 13, 11–33. doi: 10.26817/16925777.310

Tremblay, A., and Baayen, H. (2010). "Holistic processing of regular four-word sequences: a behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall" in *Perspectives on Formulaic Language: Acquisition and Communication*. ed. D. Wood (London: Continuum), 151–173.

Yu, X. (2013). Learning a foreign language through text and memorisation: the Chinese learners' perceptions. *J. Lang. Teach. Res.* 4, 731–740. doi: 10.4304/jltr.4.4.731-740

Yu, X. (2017). The use of text recitation as a teaching tool in EFL: a cross-educational level investigation in China. *Int. J. Appl. Linguist. Engl. Lit.* 6, 191–197. doi: 10.7575/aiac.ijalel.v.6n.5p.191

Zoltán, D. (2007). *Research methods in applied linguistics[M]*. London: Oxford University Press, 2007.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 17 May 2023
DOI 10.3389/fpsyg.2023.1164262



OPEN ACCESS

EDITED BY
Dingfang Shu,
Shanghai International Studies University, China

REVIEWED BY
Jie Zhou,
Shandong University, China
Bruce W. Lee,
LXPER Inc., Republic of Korea

*CORRESPONDENCE
Zhen Bao
✉ baozhen2020@sjtu.edu.cn

RECEIVED 12 February 2023
ACCEPTED 18 April 2023
PUBLISHED 17 May 2023

CITATION
Peng C and Bao Z (2023) Effects of reasoning demands triggered by genre on Chinese EFL learners' writing performance.
*Front. Psychol.* 14:1164262.
doi: 10.3389/fpsyg.2023.1164262



# Effects of reasoning demands triggered by genre on Chinese EFL learners' writing performance

Cheng Peng and Zhen Bao*

School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China

**Introduction:** Genres, having distinct communicative functions, elicit different levels of reasoning demands in writing tasks. The current study investigated the influence of cognitive complexity triggered by a seldom studied pair of genres (expository writing vs. argumentative writing) on Chinese advanced EFL learners' writing performance.

**Method:** A total of 76 L2 learners participated in two writing tasks: one simpler expository writing task involving fewer reasoning demands and the other more complex argumentative writing task eliciting more reasoning demands. Multiple measure indices were adopted to comprehensively reflect the differences in production dimensions between the two writing tasks, such as lexical complexity, syntactic complexity, accuracy, fluency, and cohesion.

**Results and discussion:** The results showed that cognitive complexity significantly improved lexical complexity, clausal complexity, and cohesion, which generally supported the Cognition Hypothesis. However, phrasal structures and clausal structures within the construct of syntactic complexity displayed a trade-off effect, partially corroborating the Trade-off Hypothesis. Accuracy and fluency were uninfluenced, verifying neither of these hypotheses. Implications for sequencing and designing L2 writing tasks were provided for relevant stakeholders.

KEYWORDS

**task complexity, L2 writing performance, the cognition hypothesis, the trade-off hypothesis, expository writing, argumentative writing**

## 1. Introduction

As second language acquisition theories and task sequencing criteria develop (Xu et al., 2022), the effect of task complexity on L2 writing performance has been examined by many studies, generating conflicting results (e.g., Ong and Zhang, 2010; Rahimi and Zhang, 2018; Zhan et al., 2021). Further research is warranted to deepen our understanding of the conceptualization and operationalization of writing task complexity and to help instructors and assessors appropriately design and sequence writing tasks based on L2 learners' proficiency (Robinson, 2015). Furthermore, it remains to be examined whether L2 writers' attentional resources are sufficient when completing writing tasks of different cognitive complexity.

Revolving around the question, Robinson (2001) proposed the cognition hypothesis that an increase in task complexity could improve L2 production quality. Motivated by the hypothesis, this study was conducted to investigate L2 writing performance across a simple task (expository writing involving a lower level of reasoning) and a complex task (argumentative writing involving a higher level of reasoning). The task complexity was manipulated via reasoning demands elicited by genres, whose effects on lexical complexity, syntactic complexity, accuracy, fluency, and cohesion were investigated.

Two hypotheses concerning task complexity proposed by Skehan and Foster (1997, 1999, 2001) and Robinson (2001, 2015) served as the theoretical foundations for this study. Similar to Kellogg's (1996) L1 writing model, Skehan and Foster (1997) proposed the trade-off hypothesis that attentional resources are limited during the L2 production process. However, Robinson (2001) proposed a contrasting hypothesis, i.e., the cognition hypothesis. In keeping with previous L2 writing studies (e.g., Rahimi, 2019), this study set out to verify the effects of task complexity, manipulated via writing genre, on L2 writing performance, based on the aforementioned two hypotheses.

# 2. Review of the literature

## 2.1. Theoretical background

Writing tasks play an important role in writing improvement and language development. As assumed in the output hypothesis (Swain, 1985), writing tasks could promote L2 learning by making learners realize the "gap" between what they want to write and what they can write, and this "gap" will motivate learners to learn more target language to modify their written output. Furthermore, during the writing process, L2 learners are forced to consider not only the semantic but also the syntactic aspects to generate legitimate and comprehensible language.

Writing tasks could activate and orchestrate various cognitive resources during three sub-processes: formulation, execution, and monitoring, according to Kellogg's (1996) model. Among the three processes, the formulation was theorized to place the most demands on the working memory, followed by monitoring. Formulation entails planning content to be written and translating it into words. While planning, writers deploy attentional resources to generate and organize content coherently. During translation, lexical and syntactic forms are accessed and encoded to express the content into words. Although Kellogg's model was initially proposed for L1 writing tasks, it was also confirmed to be applicable to L2 writing tasks (e.g., Kormos, 2011; Révész et al., 2017).

In task-based language learning, learners are required to allocate their attentional resources to meet tasks' cognitive demands. The two most influential theories concerning the influence of cognitive demands on L2 production are the trade-off hypothesis and the cognition hypothesis. These two theories make contrary predictions about the relationship between task cognitive demands and language production. Although the two hypotheses were originally put forward for L2 oral production, previous studies have confirmed their applicability to L2 writing (e.g., Rahimi and Zhang, 2018; Zhan et al., 2021).

The trade-off hypothesis was put forward by Skehan and Foster (1997, 1999, 2001). They hypothesized that attentional resources and processing capacity are limited, so learners have to prioritize one aspect of language production over others, hence triggering trade-off effects among complexity, accuracy, and fluency. For example, when an increase in cognitive task complexity triggers over-taxation of attentional resources, learners will give priority to meaning and content planning over linguistic forms, leading to increased fluency but decreased complexity and accuracy.

Contrarily, Robinson (2001, 2005) proposed that task complexity was associated with the cognitive demands imposed on learners, but learners could pay attention to multiple facets of language production simultaneously by drawing on multiple attentional resource pools, thus promoting interlanguage learning and development. Task complexity could be manipulated along resource-directing and resource-dispersing dimensions. Increasing task complexity along the resource-directing dimension (by placing cognitive demands on learners) could lead to improved accuracy, improved complexity, and decreased fluency, whereas increasing task complexity along the resource-dispersing dimension (by placing performative demands on learners) could result in lower accuracy, lower complexity, and lower fluency.

Task complexity is assumed to lie in the formulation process, in which content planning and linguistic encoding place great cognitive demands on learners (Robinson, 2005). In the present study, the argumentative writing task requires more reasoning demands, because learners have to conceptualize the content by reasoning, analyzing the controversial issue, and arguing for or against one side with supportive evidence. By contrast, an expository writing task just involves learners presenting information about one campus activity based on their prior knowledge, so the content can be accessed from learners' long-term memory with ease.

## 2.2. Reasoning demands triggered by genre and L2 writing performance

So far, due to the inconsistent operationalizations of task complexity and the use of different language complexity indices, research findings regarding the effects of reasoning demands triggered by genres on L2 writing performance have been conflicting.

A line of research partially supported the trade-off hypothesis. For example, Way et al. (2000) investigated the effects of different genres (i.e., descriptive, narrative, and expository) on L2 learners' writing performance. The results indicated that the syntactic complexity (e.g., T-unit length) was higher in expository essays than that in descriptive or narrative ones. However, an almost reverse trend (descriptive > narrative > expository) was detected in fluency and accuracy measures. Thus, as the reasoning demands increased, the trade-off effect existed among syntactic complexity, accuracy, and fluency, which supported the trade-off hypothesis.

Another line of research partially supported the cognition hypothesis. To further examine the influence of genre, Lu (2011) employed 14 syntactic complexity measures. He found that argumentative essays generally displayed more complex syntactic features than narrative ones. Yoon and Polio (2017) replicated Lu's (2011) study and included other dimensions such as lexical complexity, accuracy, and fluency. Their results revealed that the syntactic complexity in argumentative writing was generally higher than that in narrative writing, as evidenced by the length of production and phrasal complexity. However, no significant effect was detected in clausal complexity, accuracy, and fluency. Based on the above two studies (Lu, 2011; Yoon and Polio, 2017), Zhan et al. (2021) found similar results. Specifically, the argumentative

writing exhibited higher syntactic complexity (length of production and phrasal structures) and fluency than did the narrative writing, but there were no significant differences in lexical complexity or accuracy.

More comprehensively, Yang (2014) examined the effect of task complexity on L2 writing performance across four writing genres (narrative, expository, expo-argumentative, and argumentative). Yang found that accuracy, fluency, lexical diversity, and lexical sophistication were not significantly influenced; the lexical density of expository writing was the highest, while that of narrative writing was the lowest. The syntactic complexity can be ranked as argumentative > expository > narrative (the indices of expo-argumentative writing fluctuated). The general writing performance can be summarized as non-narrative > narrative.

All the above-mentioned four studies (i.e., Lu, 2011; Yang, 2014; Yoon and Polio, 2017; Zhan et al., 2021) revealed that when more reasoning demands were imposed, the produced syntactic complexity, in particular, would increase, partially corroborating the cognition hypothesis.

There still existed some other studies supporting neither of these two hypotheses. Contrary to previous studies' findings, Ruiz-Funes (2013) investigated the effects of task complexity on 24 intermediate L2 learners' writing performance, detecting non-significant differences for all measures across two writing tasks (narrative vs. expository). Again, Ruiz-Funes (2014) examined the influence of task complexity on eight advanced L2 learners' writing production, still finding no significant differences for all indices across two writing tasks (expository vs. argumentative). Although the pair of expository and argumentative writing tasks was studied, the sample size was very small, decreasing its statistical power and generalizability.

In light of different operationalizations of task complexity, it is hardly possible to compare different research results simply based on the broad categories of "less complex writing task" and "more complex writing task." Therefore, we narrow down the concept of task complexity to cognitive complexity brought about by genre in writing tasks. Nevertheless, the findings concerning writing performance influenced by different genres are still not consistent across studies, probably due to such confounding factors as the topic, learner proficiency, and the use of different sets of measure indices. It is, therefore, necessary to control these confounding factors to only focus on the influence of genres and their embedded reasoning demands.

On the other hand, though previous studies did examine genre effects in writing tasks (e.g., narrative vs. non-narrative; argumentative vs. non-argumentative), few studies are setting out to explicitly compare expository writing and argumentative writing among L2 advanced learners. In China, these two major writing genres, whose knowledge has been imparted and constantly applied during secondary and tertiary education, play a pivotal role in writing pedagogy, learning, and assessment. In academic practice, college students frequently need to formally explain concepts/information to others (corresponding to the purpose of expository writing) or to argue for/against someone's viewpoint with supporting evidence to be convincing (corresponding to the purpose of argumentative writing). Yet, the lack of studies explicitly centering on these two writing genres fails to provide pedagogical or assessment implications for these two vital types of writings, though these implications are very practical and essential.

If informed of the differences in writing performance caused by these two genres' distinct reasoning demands, relevant stakeholders will be benefited. For example, L2 instructors will be more expert at arranging or sequencing writing tasks by taking task complexity induced by reasoning demands into consideration, to promote the development of EFL learners' writing ability. Similarly, L2 assessors could be better at anchoring the validity and reliability of writing assessments involving these two genres. Furthermore, EFL learners will more consciously deploy specific linguistic features characteristic of each genre in their writings, to better fulfill the genre-related communicative functions.

To shed light on the role of the reasoning demands elicited by these two genres in writing tasks, this study set out to investigate the influence of task complexity induced by these two genres (i.e., expository writing vs. argumentative writing) on L2 writing performance. Task complexity in our study was manipulated as previous studies did (e.g., Yang, 2014; Zhan et al., 2021), and our study focused on two seldom-examined writing tasks. In addition, participants in our study were advanced Chinese EFL learners, who were seldom investigated in previous studies. Moreover, multidimensional measures of lexical complexity and syntactic complexity were adopted, which is of great importance (Norris and Ortega, 2009; Johnson, 2017). In addition, cohesion indices were also included in this study to assess the macro-level organization of written production and learners' higher-order writing skills so as to address the concern expressed by Kuiken and Vedder (2008) that the improved linguistic dimensions of L2 writing might compete with other higher-order dimensions of writing. Finally, our study classified syntactic complexity into phrasal complexity and clausal complexity, which were less explicitly examined.

Accordingly, the present study answered the following research question: How does the cognitive complexity triggered by genre influence advanced EFL learners' writing performance in terms of lexical complexity, syntactic complexity, accuracy, fluency, and cohesion?

# 3. Method

## 3.1. Participants

A total of 76 undergraduate sophomores majoring in the English language were recruited from a top university in Shanghai, China, using convenience sampling. These students got writing feedback and were awarded bonus credits based on their writing performance. Eleven participants were removed as they failed to follow the researchers' instructions. The essays of 65 students were retained for further analysis and research. Among them, there are 43 female and 22 male students, whose ages ranged from 19 to 21 years ($M = 19.86$, $SD = 9$ months). All of the participants' mother tongues were Mandarin Chinese.

The participants are engaged in a language learning program, which includes courses in linguistics, English literature, English culture, and language skills. They have been learning English as a foreign language in the classroom setting for over 13 years, but

none of them have any overseas study experiences. The Oxford Quick Placement Test (version 2.0) was used to assess students' English proficiency. As a result, all of the students' placement scores fell into the advanced proficiency range (46–57 out of 60), M = 50.95, SD = 2.56.

## 3.2. Writing tasks

Different writing genres will place different cognitive demands on EFL learners (Yoon and Polio, 2017). Expository writing and argumentative writing are two discourse types with distinct communicative purposes. Thus, an expository writing task and an argumentative writing task adapted from the Chinese CET-Band 4 and Chinese TEM-Band 4, respectively, were employed in this study (see writing prompts in Supplementary material). College English Tests (CET) designed for non-English major college students and Test for English Majors (TEM) for English major college students are both standardized English proficiency tests in China, whose reliability and validity have been examined and well-documented (Yan and Huizhong, 2006; Yan and Jinsong, 2011). It is universally accepted in China that TEM-Band 4 is much more difficult than CET-Band 4.

The participants completed the tasks in two English classes (one task each day) under the supervision of their English teacher. Students were asked to work on their own and were not allowed to use cellphones, dictionaries, or reference textbooks. Since the writing tasks were regarded as completely independent of each other, no consideration was taken concerning the practice effect of one task over the other. In this study, the time allotted for each task was 30 min.

Cognitive task complexity is associated with reasoning demands induced by genres (Ruiz-Funes, 2015). The tasks that involve more reasoning demands were thought to be more cognitively complex in EFL writing studies (Rahimi, 2019). In this study, the tasks were designed to elicit different levels of cognitive complexity. Both tasks were created around the theme of activities. In Task 1, students were instructed to introduce an impressive college activity based on their prior knowledge in an expository manner. Task 2 required students to argue for or against the phenomenon (volunteer activities) mentioned in the prompt by giving supporting evidence. In terms of the reasoning demands, Task 2 was considered to impose higher cognitive loads than Task 1.

To further validate the task complexity, five experienced EFL writing instructors and five doctoral postgraduates majoring in linguistics were invited to judge the complexity of these two writing tasks on a Likert scale, which ranged from 1 (extremely simple) to 9 (extremely complex). The expert ratings were in line with the categorization of task complexity in this study.

## 3.3. Dependent variables

The dependent variables consist of lexical complexity, syntactic complexity, accuracy, fluency, and cohesion. Task production quality is typically measured by the indices of complexity, accuracy, and fluency (e.g., Foster and Skehan, 1996; Norris and Ortega, 2009). However, there is no consensus on the measures of writing quality to date (Johnson, 2017). Researchers have adopted different indices to assess learners' writing production. The necessity of using multiple indices of linguistic complexity to investigate L2 writing quality was pointed out by Norris and Ortega (2009). Thus, to avoid the inconsistency of evaluating indices, this study adopted a set of comprehensive indices to evaluate different dimensions of writing production.

### 3.3.1. Lexical complexity

Lexical complexity, a multidimensional construct, can be categorized into three subcategories: lexical diversity, lexical sophistication, and lexical density, but very few studies examined the latter two subcategories (Johnson, 2017). This study examined all three subcategories.

Since the index type/token ratio (TTR) is sensitive to sample size and length (Rahimi, 2019), corrected type/token ratio (CTTR) was employed to measure lexical diversity, countering the influence of sample size and length effect (Ong and Zhang, 2010; Zhan et al., 2021). To comprehensively assess lexical diversity, two other well-validated indices, *D*-value and measure of textual lexical diversity (MTLD), were also adopted, which were employed in previous studies (e.g., Révész et al., 2017; Rahimi, 2019).

In addition, previous studies have found that psycholinguistic values, such as the age of acquisition (AoA) and concreteness, are important indicators of lexical sophistication (De Wilde et al., 2020). The log frequency for content words was shown to be more reliable to indicate lexical sophistication than the raw frequency (Just and Carpenter, 1980). This study adopted these three indices to examine lexical sophistication, which can predict the quality of L2 writing (Zhang et al., 2022). Finally, lexical density (LD) was employed to gain a big picture of learners' lexical complexity.

### 3.3.2. Syntactic complexity

Syntactic complexity, also a multidimensional construct, could be analyzed from different syntactic facets. In line with the previous classification of syntactic complexity (Bulté and Housen, 2014; Kyle and Crossley, 2018), apart from measuring production length, the present study measured syntactic complexity at other two levels: phrasal complexity and clausal complexity.

In this study, multiple indices were adopted to capture the genres' effects on syntactic complexity by following previous studies (e.g., Lu, 2011; Rahimi and Zhang, 2019; Zhan et al., 2021). The first syntactic index (i.e., STS) measures the extent of structural similarity in the text using Coh-Metrix 3.0 (McNamara et al., 2010), which reflects the overall syntactic sophistication. The higher the index is, the less diverse the syntactic structures are. The other 11 syntactic indices obtained from the Syntactic Complexity Analyzer (Lu, 2011) focus on three dimensions: three indices concerning the length of the unit, e.g., mean length of sentence (MLS), mean length of T-unit (MLT), and mean length of clause (MLC); three indices measuring clausal complexity, e.g., clauses per T-unit (C/T), dependent clauses per clause (DC/C), and dependent clauses per T-unit (DC/T); five indices calculating phrasal complexity, e.g., verb phrases per T-unit (VP/T), complex nominals per T-unit (CN/T), complex nominals per clause (CN/C), coordinate phrases per T-unit (CP/T), and coordinate phrases per clause (CP/C).

#### 3.3.3. Accuracy

Accuracy refers to lexical and grammatical correctness in learners' essays, measured by errors made in learners' writing, but the errors are not concerned with punctuation, spelling, or capitalization, which are not typical issues among advanced EFL learners. In this study, the number of errors per T-unit and the number of errors per 100 words were adopted to measure the learner's accuracy, both of which were employed and verified in previous studies (e.g., Ruiz-Funes, 2015). The higher the ratios are, the less accurate learners' essays are.

#### 3.3.4. Fluency

Since the writing time was controlled for all learners, the total number of words was used as one index of fluency (Johnson et al., 2012). In addition, fluency was also assessed by the other index, i.e., the mean number of words per T-unit, which was regarded as a reliable fluency index by Wolfe-Quintero et al. (1998). The two indices combined can better capture the fluency of learners' writing production.

#### 3.3.5. Cohesion

Cohesion proved to be one of the important indicators of L2 writing performance (Crossley et al., 2016; Zhang et al., 2022). Cohesive devices play a vital role in connecting ideas in writing (Halliday and Hasan, 1976). Previous research has validated and confirmed the efficacy of Coh-Metrix indices (i.e., latent semantic analysis, co-reference, and connectives) in evaluating textual cohesion (Foltz et al., 1998; McNamara et al., 2010).

Latent semantic analysis (LSA) is a statistical representation of textual cohesion by evaluating the level of semantic similarity between sentences and paragraphs (Foltz, 1996). In this study, both local and global LSA indices were used to measure the conceptual similarity between sentences and paragraphs, respectively.

Co-reference was measured by stem overlap and content word overlap, both of which are more inclusive compared with noun overlap or argument overlap (Crossley et al., 2016). Stem overlap refers to how often a noun in one sentence shares a common lemma with another content word in another sentence. Content word overlap calculates the number of shared content words between sentences.

Connectives are vital signal words of relations in essays and thus promote discourse cohesion. The appropriate use of connectives can improve textual organization and content unity (Halliday and Hasan, 1976). To reflect the overall use of connectives in learners' essays, we adopted the holistic index to identify all connectives used in learners' writings.

### 3.4. Statistical analyses

First, 11 students' essays were excluded due to not following requirements, so the two writing tasks of 65 students were analyzed. Then, the 130 essays were typed in Microsoft Word documents and coded by researchers via the Lexical Complexity Analyzer (Lu, 2012), Syntactic Complexity Analyzer (Lu, 2011), and Coh-Metrix 3.0 (McNamara et al., 2010) to obtain lexical complexity, syntactical complexity, fluency, and cohesion indices. Paired samples *t*-tests were carried out to check the differences between the two writing tasks' performance indices, with the alpha level set at 0.05 for all tests. Cohen's *d* was adopted to measure the effect size, and the standards were followed: $d = 0.2–0.4$, a small effect size; $d = 0.5–0.7$, a medium effect size; and $d > 0.8$, a large effect size (Cohen, 1988). The program SPSS 21.0 was employed for the abovementioned statistical analyses.

## 4. Results

### 4.1. Effects on learners' lexical complexity

The paired samples *t*-tests detected a series of task effects on lexical complexity (see Table 1). First, the mean differences between the two lexical diversity indices were both statistically significant. For example, as for *D*-value, MD = −12.96, $p = 0.001$, $d = 0.81$; and as for MTLD, MD = −10.85, $p = 0.04$, $d = 0.47$. It meant that EFL learners were more likely to use more diverse lexical forms when completing the complex writing task (i.e., argumentative writing). Furthermore, the complex writing task elicited significantly more abstract words than did the less complex task, MD = 50.30, $p = 0.000$, $d = 2.14$. Task effect was also reflected in AoA, MD = −76.12, $p = 0.000$, $d = 2.72$, which indicated that learners tended to use later acquired words in the more complex task. Moreover, the effect sizes for the above *t*-tests were generally large. These lexical indices indicated that participants tended to deploy diverse and sophisticated vocabulary when dealing with more complex writing tasks. As task complexity increased, the lexical complexity generally showed an upward trend.

In addition, the paired samples *t*-tests failed to reveal the effects of task complexity on other indices of lexical complexity. As for CTTR, log frequency, and LD, none of the mean differences between the two tasks were statistically significant.

### 4.2. Effects on learners' syntactic complexity

As indicated in Table 2, regarding the structural similarity, argumentative writing presented significantly fewer similar syntactic structures than did expository writing, MD = 0.02, $p = 0.004$, $d = 0.70$, which showed that advanced learners were more likely to vary their syntactic structures in more complex writing task. Regarding the length of production, only MLC in Task 1 was significantly higher than that in Task 2, and the effect size was very large (MD = 2.07, $p = 0.000$, $d = 1.06$), which meant that learners produced longer clauses in the simple writing task than in the complex one.

Additionally, increasing task complexity had a significant influence on EFL learners' clausal complexity features. The mean values of three indices in Task 2 were all significantly higher than those in Task 1, with large effect sizes (C/T, MD = −0.38, $p = 0.01$, $d = 0.58$; DC/C, MD = −0.11, $p = 0.001$, $d = 0.87$; DC/T, MD = −0.33, $p = 0.005$, $d = 0.70$), which meant that the more complex writing task elicited more clausal constructions.

Contrarily, advanced EFL learners tended to generate fewer phrasal structures (except VP/T) in complex writing tasks. Instead, they tended to employ more phrasal devices (especially nominals)

TABLE 1 Comparison of lexical complexity in Task 1 (exposition) and Task 2 (argumentation).

| Sub-categories | Indices | Task 1 | | Task 2 | | MD | *p* | *d* |
|---|---|---|---|---|---|---|---|---|
| | | M | SD | M | SD | | | |
| Lexical diversity | CTTR | 6.59 | 1.05 | 6.21 | 0.53 | 0.39 | 0.09 | 0.39 |
| | *D*-value | 79.50 | 15.76 | 92.46 | 16.76 | −12.96 | 0.001 | 0.81 |
| | MTLD | 75.44 | 19.28 | 86.29 | 17.18 | −10.85 | 0.04 | 0.47 |
| Lexical sophistication | Log freq | 3.02 | 0.08 | 3.00 | 0.09 | 0.02 | 0.41 | 0.18 |
| | AoA | 310.64 | 17.78 | 386.76 | 25.04 | −76.12 | 0.000 | 2.72 |
| | Concreteness | 415.43 | 19.17 | 365.13 | 12.10 | 50.30 | 0.000 | 2.14 |
| Lexical density | LD | 0.52 | 0.03 | 0.52 | 0.03 | 0.00 | 0.76 | 0.07 |

CTTR, corrected type/token ratio; D-value, lexical variability based on Malvern and Richards (1997); MTLD, measure of textual lexical density; Log freq, log frequency for content words; AoA, age of acquisition for content words; Concreteness, concreteness for content words; LD, lexical density.

TABLE 2 Comparison of syntactic complexity in Task 1 (exposition) and Task 2 (argumentation).

| Sub-categories | Indices | Task 1 | | Task 2 | | MD | *p* | *d* |
|---|---|---|---|---|---|---|---|---|
| | | M | SD | M | SD | | | |
| Overall sophistication | STS | 0.10 | 0.03 | 0.08 | 0.02 | 0.02 | 0.004 | 0.70 |
| Mean length of unit | MLS | 19.21 | 5.23 | 20.34 | 6.45 | −1.13 | 0.51 | 0.15 |
| | MLT | 17.51 | 4.95 | 18.30 | 6.43 | −0.79 | 0.64 | 0.10 |
| | MLC | 11.88 | 2.69 | 9.81 | 1.44 | 2.07 | 0.000 | 1.06 |
| Clausal complexity | DC/T | 0.48 | 0.24 | 0.81 | 0.39 | −0.33 | 0.005 | 0.70 |
| | DC/C | 0.31 | 0.11 | 0.42 | 0.08 | −0.11 | 0.001 | 0.87 |
| | C/T | 1.48 | 0.28 | 1.86 | 0.56 | −0.38 | 0.01 | 0.58 |
| Phrasal complexity | VP/T | 1.88 | 0.36 | 2.77 | 0.76 | −0.89 | 0.000 | 0.98 |
| | CN/T | 2.30 | 0.76 | 1.90 | 0.61 | 0.40 | 0.050 | 0.46 |
| | CN/C | 1.57 | 0.46 | 1.03 | 0.21 | 0.54 | 0.000 | 1.41 |
| | CP/T | 0.52 | 0.30 | 0.39 | 0.17 | 0.13 | 0.01 | 0.61 |
| | CP/C | 0.36 | 0.23 | 0.22 | 0.10 | 0.14 | 0.002 | 0.77 |

STS, structural similarity; MLS, mean length of sentence; MLT, mean length of T-unit; MLC, mean length of the clause; DC/T, dependent clauses per T-unit; DC/C, dependent clauses per clause; C/T, clauses per T-unit; VP/T, verb phrases per T-unit; CN/T, complex nominals per T-unit; CN/C, complex nominals per T-unit; CP/T, coordinate phrases per T-unit; CP/C, coordinate phrases per clause.

to convey and present information in the simple writing task (i.e., expository writing). All phrasal indices showed significant differences between the two writing tasks with large effect sizes. It can be shown from Figure 1 that there existed a trade-off effect between phrasal complexity and clausal complexity.

In summary, with the increase of reasoning demands in EFL learners' writing tasks, the number of phrasal structures significantly decreased (especially nominals), whereas the number of clausal structures significantly increased.

## 4.3. Effects on learners' writing accuracy and fluency

Task complexity had no significant effect on learners' writing accuracy or fluency, as shown in Table 3. Neither index of accuracy displayed statistically significant differences between the two writing tasks, e.g., MD = −0.10, $p = 0.43$, $d = 0.17$ (Etot/T); MD = 0.02, $p = 0.97$, $d = 0.00$ (NER). Similarly, neither fluency index yielded evidence of significant differences, e.g., MD = −0.80, $p = 0.64$, $d = 0.10$ (W/T); MD = 1.14, $p = 0.98$, $d = 0.01$ (TNW).

## 4.4. Effects on learners' writing cohesion

With the increase in task complexity, the essays generated in the complex task were more cohesive than those in the simple task, as evidenced by both implicit and explicit measures (see Table 4). The implicit measures, i.e., LSA-p, SO-s, and CWO-s, all showed significant differences between the two writing tasks with medium effect sizes. The explicit measure, i.e., ACI, revealed that the difference in the use of connectives between the two writing tasks was statistically significant with a large effect size (MD = −20.41, $p = 0.000$, $d = 1.15$). In short, advanced EFL learners tended to employ more cohesive devices in the more reason-demanding writing task (i.e., argumentative writing).

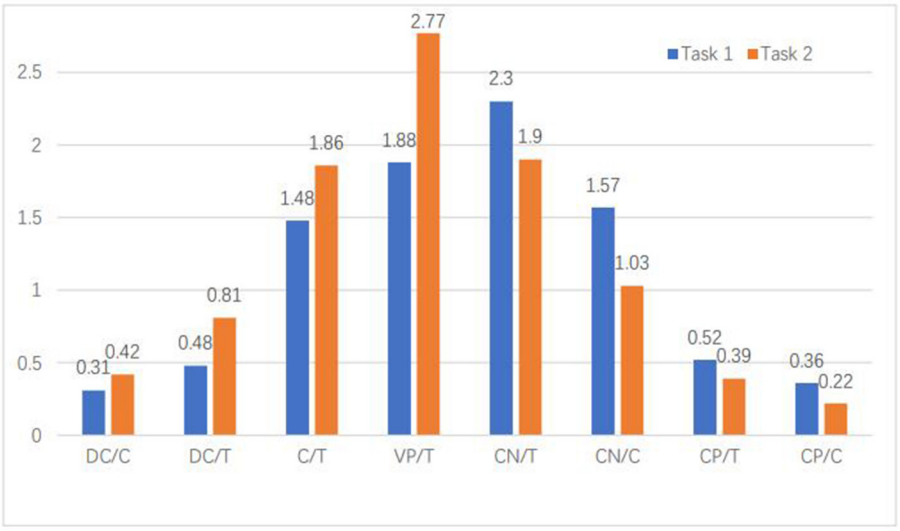



FIGURE 1
Clausal and phrasal complexity indices' comparison.

# 5. Discussion

## 5.1. Effects on lexical complexity

Increasing cognitive complexity had significantly positive influences on lexical diversity (*D*-value and MTLD) and lexical sophistication (AoA and concreteness). With the increase in reasoning demands, advanced EFL learners' lexical complexity in writing also increased, which supported Robinson's cognition hypothesis.

Advanced EFL learners tended to employ more diverse and sophisticated words when dealing with the cognitively demanding task, probably because simple words could not meet the demands of a complex writing task, which involved a deeper level of form-conceptualization mapping. In addition, advanced learners had a good knowledge of L2 vocabulary, and the complex task provided them with a channel of lexical production. Advanced learners may be more skilled at funneling their attentional resources toward the lexical forms while conceptualizing and organizing content to be written (Rahimi, 2019). Learners' lexical density remained constant across two writing tasks, which may indicate that learners' high proficiency pushed their content words' proportion to the limit, reaching the ceiling effect.

On the one hand, some of the results regarding lexical complexity corroborated previous studies. For example, the findings concerning CTTR supported Kuiken and Vedder's (2008) and Zhan et al.'s (2021) studies, which found no significant difference either. Regarding frequency, the findings were consistent with Kormos's (2011) and Johnson et al.'s (2012) studies in which task complexity did not have significant effects on lexical frequency.

On the other hand, our results also refuted previous findings. Concerning *D*-value and concreteness, our results contradicted the study of Kormos (2011), who conducted two narrative writing tasks and found that *D*-value and concreteness decreased significantly with the increase of cognitive complexity. Regarding *D*-value and MTLD, our findings were not in line with Révész et al. (2017), who manipulated task complexity via content support. The inconsistent results may be caused by different manipulations or operationalizations of task complexity.

TABLE 3 Comparison of accuracy and fluency in Task 1 and Task.

| | Task 1 | | Task 2 | | MD | *p* | *d* |
|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | | | |
| Etot/T | 0.68 | 0.25 | 0.78 | 0.56 | −0.10 | 0.43 | 0.17 |
| NER | 4.23 | 2.26 | 4.21 | 1.86 | 0.02 | 0.97 | 0.00 |
| W/T | 17.50 | 4.94 | 18.30 | 6.43 | −0.80 | 0.64 | 0.10 |
| TNW | 386.52 | 177.36 | 385.38 | 60.77 | 1.14 | 0.98 | 0.01 |

Etot/T, number of errors per T-unit; NER, number of errors per 100 words; W/T, words per T-unit; TNW, total number of words.

Our findings regarding general lexical complexity were inconsistent with some previous studies (e.g., Kuiken and Vedder, 2008; Zhan et al., 2021), which may be due in part to the use of different lexical indices. Specifically, relatively few studies adopted sophisticated lexical indices, such as *D*-value and MTLD, which were considered more sensitive to the change of task complexity and less influenced by text length compared with TTR and its transformations (Johnson, 2017). Also, very few studies included multiple lexical complexity indices to reflect the effects of task complexity on all three dimensions of lexical complexity, i.e., lexical diversity, sophistication, and density. Many studies only examined the lexical diversity indices, which might not be sensitive enough to capture the task complexity's effects on lexical complexity, thus resulting in inconsistent findings.

To sum up, the set of comprehensive indices in this study indicated that lexical complexity would increase with the increase of reasoning demands.

## 5.2. Effects on syntactic complexity

The structural similarity index, reflecting the overall syntactic complexity, showed that the syntactic structures in the complex writing task were more varied, partially due to the use of more clauses.

The results also indicated that the clauses in the complex writing task (i.e., argumentative writing) were significantly shorter (as shown by MLC) because more clauses/dependent clauses were embedded in essays produced in the complex task, as revealed by significantly higher DC/T, C/T, and DC/C in argumentative writing. As for the simple writing task, the significantly longer clauses were attributable to the fact that more phrases were embedded in clauses, as reflected by statistically higher CP/C and CN/C in expository writing. EFL learners employed more phrases, instead of clauses or dependent clauses (correspondingly, fewer verb phrases), to pack and condense more information in expository writing (see Table 5).

Phrasal complexity and clausal complexity did not increase simultaneously but competed with each other. As phrasal complexity increased in expository writing, the clausal complexity would fall. On the contrary, when phrasal complexity decreased in argumentative writing, the clausal complexity would rise. Our findings also corroborated the findings in Yang's (2014) and Lei et al.'s (2023) studies, whose clausal complexity exhibited an increasing trend from expository writing to argumentative writing. Contrarily, the nominal phrasal complexity presented a downward trend with the increase in reasoning demands.

Previous researchers also found a trade-off effect between phrasal and clausal complexities in other pairs of writing tasks. For example, Biber et al. (2011) found that informative writing tended to show more phrasal structures and fewer clausal structures, while spoken discourse displayed a reverse trend. The simple writing task in our study also required an informational presentation of certain school activities, thus presenting more phrasal features; by contrast, similar to the communicative function characteristic of conversation, the complex writing task in our study had a persuasive purpose, thus showing more clausal features.

A similar trade-off effect could also be found across narrative and argumentative writing tasks in previous studies (Lu, 2011; Yoon and Polio, 2017; Zhan et al., 2021). With the increase of reasoning demands triggered by genre, the phrasal complexity indices increased, while the clausal complexity indices decreased, probably because learners needed to employ different kinds of syntactic structures to meet distinct reasoning demands induced by genre-related communicative functions.

The findings concerning syntactic complexity can be interpreted in both weak and strong manners. The weak interpretation is that both genres of essays were complex in syntax but in different dimensions. In expository writing, learners were more likely to generate phrasal structures to pack more information into relatively longer clauses. In argumentative writing, which required a higher level of reasoning about others' motivations for doing something controversial, learners preferred to use cognitive state terms accompanied by clausal structures, e.g., *somebody thought that, somebody wondered whether*. No obvious increase of holistic syntactic complexity can be inferred from the weak interpretation, thus not supporting the cognition hypothesis.

The strong explanation for the findings is based on the developmental progression hypothesis (Biber et al., 2011, 2016). In L2 development, clausal structures are acquired at relatively earlier stages and represent a lower level of syntactic complexity. By contrast, complex phrasal embedding is produced in later stages toward adulthood, which is considered to represent a higher level of syntactic complexity. Thus, it can be cautiously concluded that the syntactic complexity decreased (with fewer phrasal features and more clausal features) as task complexity increased. Therefore, the findings were aligned with the trade-off hypothesis.

As noted by Ellis and Yuan (2004), when learners were composing essays, they gave priority to the access of lexical items over the generation of syntactic structures, which meant that the cognitive resources allotted to syntactic complexity were limited. The cognitive loads imposed on limited working memory in a timed condition could easily result in a trade-off effect. It can be inferred that advanced learners in our study would resort to clausal devices to relieve cognitive loads when reasoning demands increased, to funnel more attentional resources to higher-order writing skills, e.g., content conceptualization and textual organization. On the contrary, when a task imposed fewer reasoning demands, learners could allocate more available cognitive resources to retrieve or construct phrasal structures, which is conceived as more challenging.

To sum up, due to the limited attentional resources, participants could only attend to some dimension of syntactic complexity, especially when the reasoning demands involved in

TABLE 4 Comparison of cohesion in Task 1 and Task 2.

| | Task 1 | | Task 2 | | MD | *p* | *d* |
|---|---|---|---|---|---|---|---|
| | M | SD | M | SD | | | |
| LSA-s | 0.18 | 0.07 | 0.16 | 0.05 | 0.01 | 0.42 | 0.18 |
| LSA-p | 0.31 | 0.11 | 0.41 | 0.10 | −0.09 | 0.005 | 0.69 |
| SO-s | 0.33 | 0.14 | 0.40 | 0.19 | −0.07 | 0.05 | 0.44 |
| CWO-s | 0.07 | 0.03 | 0.09 | 0.02 | −0.03 | 0.000 | 0.78 |
| ACI | 80.09 | 16.73 | 100.50 | 15.32 | −20.41 | 0.000 | 1.15 |

LSA-s, Latent Semantic Analysis overlap (at sentence level); LSA-p, Latent Semantic Analysis overlap (at paragraph level); SO-s, Stem overlap (at sentence level); CWO-s, Content word overlap (at sentence level); ACI, All connectives incidence.

TABLE 5 Summary of syntactic complexity across two writing tasks.

| Task type | Genre type | Clause length | Clause density | Phrase density | Phrasal constituents |
|---|---|---|---|---|---|
| Task 1 | Expository | Longer | Lower | Higher | More nominal phrases; fewer verb phrases |
| Task 2 | Argumentative | Shorter | Higher | Lower | Fewer nominal phrases; more verb phrase |

tasks increased. The trade-off effect between clausal complexity and phrasal complexity supported the trade-off hypothesis.

## 5.3. Effects on accuracy and fluency

On the one hand, the two accuracy indices (Etot/T and NER) in our study proved to be uninfluenced by task complexity, supporting previous studies (Kormos, 2011; Ruiz-Funes, 2015; Yoon and Polio, 2017; Zhan et al., 2021). However, our results contradicted the results of Rahimi and Zhang's (2018) and Rahimi's (2019) studies, both of which showed that accuracy significantly decreased when task complexity increased. In addition, our results also refuted Kuiken and Vedder's (2007) and Yang's (2014) studies, which found that accuracy significantly increased when task complexity increased.

On the other hand, the findings for fluency indicated that task complexity did not influence fluency, confirming the findings of Révész et al.'s (2017) and Yoon and Polio's (2017) studies. However, the results were not consistent with some previous research findings (e.g., Yang, 2014; Rahimi and Zhang, 2018; Zhan et al., 2021) that the increase in task complexity had significantly positive impacts on fluency measures. The inconsistent findings regarding accuracy and fluency could be attributed to the adoption of different measure indices and different manipulations of task complexity.

This result did not support the trade-off hypothesis or the cognition hypothesis, which may be caused by learners' insensitivity to respond to different reasoning demands across these two writing tasks or the possibility that the cognitive demands of the two tasks, in the view of the learners, were not different enough to generate distinct performances in accuracy and fluency. Furthermore, that could be because the influences of reasoning demands on accuracy and fluency are insignificant in magnitude compared with that of learners' L2 proficiency, which was well-controlled for in this study (i.e., the participants were homogenous in proficiency). As revealed by Way et al. (2000), L2 proficiency level significantly influenced writing accuracy and fluency across different writing genres. Learners of the same L2 proficiency were expected to consistently produce similar language quality with similar fluency across different writing tasks (Norris and Ortega, 2009; Guo et al., 2013).

## 5.4. Effects on cohesion

The findings in this study indicated that task complexity induced by different reasoning demands had a significant impact on cohesion. In our study, the complex writing essays (i.e., argumentative essays) were found to be more coherent than the simple ones (i.e., expository essays). Specifically, the former featured a higher level of LSA (global), lexical overlap, and connective devices. The results supported Rahimi's (2019) findings that cohesion in essays would improve with the increase of task complexity manipulated via reasoning demands. However, our findings were not consistent with Kormos's (2011) and Révész et al.'s (2017) studies, which found that there were no significant effects of task complexity (manipulated via content support) on cohesion across two writing tasks. The difference in the operationalization of task complexity may lead to inconsistent findings.

The findings concerning cohesion partially supported the cognition hypothesis. As task complexity increased, the cohesive features increased along with lexical complexity and clausal complexity. Although the cognition hypothesis did not explicitly predict the influence of task complexity on textual cohesion, simultaneous improvements in cohesion and linguistic complexity indicated that increasing task complexity triggered by reasoning demands could enhance L2 writing quality, confirming the spirit of the cognition hypothesis.

The reasoning demands in complex tasks prompted L2 learners to utilize multiple resource pools to process different dimensions of L2 production simultaneously. The complex writing task could motivate L2 learners to produce more complex linguistic forms (i.e., micro-level lexical and clausal complexity) and meanwhile to pay attention to higher-order writing skills (i.e., macro-level organizational features), consequently enhancing the overall writing quality. The findings addressed the concern expressed by Kuiken and Vedder's (2008).

# 6. Conclusion and limitations

Considering the paucity of task complexity research into the comparison between expository writing and argumentative writing, our study aimed to investigate the effects of manipulating task complexity (±reasoning demands) on Chinese advanced EFL learners' writing production. The findings showed that increasing task complexity, as manipulated via reasoning demands elicited by genre, generally improved L2 writing performance. Specifically, the essays in the complex writing task (i.e., argumentative writing) exhibited increased lexical complexity and clausal complexity, as well as more cohesive features. However, the increase in task complexity also led to the use of fewer phrasal structures in the complex writing task. Additionally, the accuracy and fluency were not influenced by the increase in task complexity.

Theoretically, these findings overall supported the cognition hypothesis in that increasing reasoning demands led to improvements in lexical complexity, clausal complexity, and textual cohesion. However, within the construct of syntactic complexity, there existed a trade-off effect between phrasal structures and clausal structures, which also supported the trade-off hypothesis. In terms of the constancy of accuracy and fluency measures, neither the cognition hypothesis nor the trade-off hypothesis was corroborated, probably because the two hypotheses are aimed at the influence of task complexity on speaking performance, rather than writing performance.

Methodologically, different coding tools were utilized to measure the same construct to improve measuring reliability, such as the Lexical Complexity Analyzer, Syntactic Complexity Analyzer, and Coh-Metrix. In addition, multi-dimensional/fine-grained operationalizations of one linguistic construct were employed, e.g., seven indices were used to measure lexical complexity. Moreover, the study explicitly classified syntactic complexity into phrasal complexity and clausal complexity, which was frequently proposed

in previous studies (Staples et al., 2016; Yoon and Polio, 2017; Kyle and Crossley, 2018) but was not widely adopted.

Pedagogically, EFL instructors and assessors should consider cognitive demands when assigning writing tasks and designing writing assessments. EFL writing tasks should be sequenced from the simple to the complex in terms of the involved reasoning demands, so as to promote L2 learning and interlanguage development (Robinson, 2015). Tasks requiring fewer reasoning demands should be completed before those requiring more demands. Compared with the process of simply introducing an activity based on prior knowledge, arguing for or against an issue with justifiable evidence consumed more cognitive resources. Since advanced EFL learners were able to generate more complex language in the complex writing task, they should be given more chances to perform complex writing tasks, so as to promote L2 development through output. Moreover, considering that producing phrasal structures and clausal structures simultaneously might overwhelm learners' limited attentional resources, teachers might as well develop and adopt instructional strategies to train learners to use advanced syntactic structures packed with more information (e.g., clausal structures embedded with phrases), which would be helpful for learners to retrieve these advanced structures.

There are some limitations in this study. First, since subjects in this study fell into the advanced L2 proficiency range, the findings cannot be generalized to learners belonging to other proficiency levels. Second, as for fluency measures, although writing time was held constant for all learners, there was a possibility that some learners wrote more quickly and finished ahead of the time limit. Therefore, the findings regarding the effect of task complexity on fluency should be consulted cautiously. Third, we conducted a series of paired-samples t-tests for multiple comparisons without applying the Bonferroni adjustment, which might increase the probability of Type I errors. Considering the concern, we rechecked the statistical analyses using Bonferroni adjusted alphas (e.g., alpha for lexical complexity set at 0.05/7 = 0.0071; alpha for syntactic complexity set at 0.05/12 = 0.0042; alpha for cohesion set at 0.05/5 = 0.01), and found that some indices' test results would become non-significant with alphas above the thresholds (e.g., MTLD, DC/T, CN/T, CP/T). Nevertheless, these indices could still reflect the systematic changes between the two writing tasks' productions. Besides, the overall writing performance discrepancies can be captured through the lens of other alternative indices which measured the same construct in nature (e.g., D-value, DC/C, CN/C, CP/C). In short, we arrived at the same conclusion using either original alphas or adjusted alphas. Future researchers may consider conducting a general linear model (MANOVA) to control for multiple within-participant comparisons. Finally, the trade-off effect between clausal complexity and phrasal complexity calls for more future research to examine it across writing tasks of different cognitive complexity.

## Data availability statement

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

## 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 for participation was not required for this study in accordance with the national legislation and the institutional requirements.

## Author contributions

CP was responsible for conducting the research, writing, and revising the manuscript. ZB was responsible for analyzing the data and revising the manuscript. CP and ZB responded to all the reviewers' comments and edited the final version of the manuscript.

## Acknowledgments

The authors thank the students who participated in this study. The authors also thank two reviewers for their valuable comments.

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1164262/full#supplementary-material

## References

Biber, D., Gray, B., and Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Q.* 45, 5–35. doi: 10.5054/tq.2011.244483

Biber, D., Gray, B., and Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Appl. Linguist.* 37, 639–668. doi: 10.1093/applin/amu059

Bulté, B., and Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *J. Second Lang. Writ.* 26, 42–65. doi: 10.1016/j.jslw.2014.09.005

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

Crossley, S. A., Kyle, K., and McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *J. Second Lang. Writ.* 32, 1–16. doi: 10.1016/j.jslw.2016.01.003

De Wilde, V., Brysbaert, M., and Eyckmans, J. (2020). Learning English through out-of-school exposure: how do word-related variables and proficiency influence receptive vocabulary learning? *Lang. Learn.* 70, 349–381. doi: 10.1111/lang.12380

Ellis, R., and Yuan, F. (2004). The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Stud. Second Lang. Acquis.* 26, 59–84. doi: 10.1017/S0272263104026130

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behav. Res. Methods Instrum. Comput.* 28, 197–202. doi: 10.3758/BF03204765

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discour. Process.* 25, 285–307. doi: 10.1080/01638539809545029

Foster, P. W., and Skehan, P. (1996). The influence of planning and task type on second language performance. *Stud. Second Lang. Acquis.* 18, 299–323. doi: 10.1017/S0272263100015047

Guo, L., Crossley, S. A., and McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: a comparison study. *Assess. Writ.* 18, 218–238. doi: 10.1016/j.asw.2013.05.002

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

Johnson, M. D. (2017). Cognitive task complexity and L2 written syntactic complexity, accuracy, lexical complexity, and fluency: a research synthesis and meta-analysis. *J. Second Lang. Writ.* 37, 13–38. doi: 10.1016/j.jslw.2017.06.001

Johnson, M. D., Mercado, L., and Acevedo, A. (2012). The effect of planning sub-processes on L2 writing fluency, grammatical complexity, and lexical complexity. *J. Second Lang. Writ.* 21, 264–282. doi: 10.1016/j.jslw.2012.05.011

Just, M. A., and Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychol. Rev.* 87, 329–354. doi: 10.1037/0033-295x.87.4.329

Kellogg, R. T. (1996). "A model of working memory in writing," *in The Science of Writing: Theories, Methods, Individual Differences and Applications*, eds C. M. Levy and S. Ransdell (Mahwah, NJ: Lawrence Erlbaum), 57–71.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *J. Second Lang. Writ.* 20, 148–161. doi: 10.1016/j.jslw.2011.02.001

Kuiken, F., and Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *Int. Rev. Appl. Linguist. Lang. Teach.* 45, 261–284. doi: 10.1515/iral.2007.012

Kuiken, F., and Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *J. Second Lang. Writ.* 17, 48–60. doi: 10.1016/j.jslw.2007.08.003

Kyle, K., and Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Lang. J.* 102, 333–349. doi: 10.1111/modl.12468

Lei, L., Wen, J., and Yang, X. (2023). A large-scale longitudinal study of syntactic complexity development in EFL writing: a mixed-effects model approach. *J. Second Lang. Writ.* 59:100962. doi: 10.1016/j.jslw.2022.100962

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college level ESL writers' language development. *TESOL Q.* 45, 36–62. doi: 10.5054/tq.2011.240859

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

Malvern, D. D., and Richards, B. J. (1997). "A new measure of lexical diversity," in *Evolving Models of Language,* eds A. Ryan and A. Wray (Clevedon: Multilingual Matters), 58–71.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., and Graesser, A. C. (2010). Coh-metrix: capturing linguistic features of cohesion. *Discourse Process.* 47, 292–330. doi: 10.1080/01638530902959943

Norris, J. M., and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Appl. Linguist.* 30, 555–578. doi: 10.1093/applin/amp044

Ong, J., and Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *J. Second Lang. Writ.* 19, 218–233. doi: 10.1016/j.jslw.2010.10.003

Rahimi, M. (2019). Effects of increasing the degree of reasoning and the number of elements on L2 argumentative writing. *Lang. Teach. Res.* 23, 633–654. doi: 10.1177/1362168818761465

Rahimi, M., and Zhang, L. J. (2018). Effects of task complexity and planning conditions on L2 argumentative writing production. *Discour. Process.* 55, 726–742. doi: 10.1080/0163853X.2017.1336042

Rahimi, M., and Zhang, L. J. (2019). Writing task complexity, students' motivational beliefs, anxiety and their writing production in English as a second language. *Read. Writ.* 32, 761–786. doi: 10.1007/s11145-018-9887-9

Révész, A., Kourtali, N. E., and Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Lang. Learn.* 67, 208–241. doi: 10.1111/lang.12205

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Appl. Linguist.* 22, 27–57. doi: 10.1093/applin/22.1.27

Robinson, P. (2005). Cognitive complexity and task sequencing: a review of studies in a Componential Framework for second language task design. *Int. Rev. Appl. Linguist. Lang. Teach.* 43, 1–33. doi: 10.1515/iral.2005.43.1.1

Robinson, P. (2015). "The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing," in *Domains and Directions in the Development of TBLT: A Decade of Plenaries from the International Conference*, ed M. Bygate (Amsterdam: John Benjamins), 87–122.

Ruiz-Funes, M. (2013). "The interplay between task complexity in foreign language writing at the intermediate level and measures of syntactic complexity, accuracy and fluency (CAF)," in *Paper presented at the American Council on Teaching of Foreign Languages* (Orlando, FL).

Ruiz-Funes, M. (2014). "Task complexity and linguistic performance in advanced college-level foreign language writing," in *Task-Based Language Learning: Insights from and for L2 Writing*, eds H. Byrnes and R. M. Manchón (Philadelphia, PA; Amsterdam: John Benjamins), 163–191.

Ruiz-Funes, M. (2015). Exploring the potential of second/foreign language writing for language learning: the effects of task factors and learner variables. *J. Second Lang. Writ.* 28, 1–19. doi: 10.1016/j.jslw.2015.02.001

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

Skehan, P., and Foster, P. (1999). The influence of task structure and processing conditions on narrative retelling. *Lang. Learn.* 49, 93–120. doi: 10.1111/1467-9922.00071

Skehan, P., and Foster, P. (2001). "Cognition and tasks," in *Cognition and Second Language Instruction*, ed P. Robinson (Cambridge: Cambridge University Press), 183–205.

Staples, S., Egbert, J., Biber, D., and Gray, B. (2016). Academic writing development at the university level: phrasal and clausal complexity across level of study, discipline, and genre. *Writt. Commun.* 33, 149–183. doi: 10.1177/0741088316631527

Swain, M. (1985). "Communicative competence: some roles of comprehensible input and comprehensible output in its development," in *Input in Second Language Acquisition*, eds S. Gass and C. Madden (Rowley, MA: Newbury House), 235–253.

Way, P., Joiner, E. G., and Seaman, M. (2000). Writing in the secondary foreign language classroom: the effects of prompts and tasks on novice learners of French. *Modern Lang. J.* 84, 171–184. doi: 10.1111/0026-7902.00060

Wolfe-Quintero, K., Inagaki, S., and Kim, H. Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. Honolulu, HI: University of Hawaii, Second Language Teaching and Curriculum Center.

Xu, T. S., Zhang, L. J., and Gaffney, J. S. (2022). Examining the relative effects of task complexity and cognitive demands on students' writing in a second language. *Stud. Second Lang. Acquis.* 44, 483–506 doi: 10.1017/S0272263121000310

Yan, J., and Huizhong, Y. (2006). The English proficiency of college and university students in China: as reflected in the CET. *Lang. Cult. Curric.* 19, 21–36. doi: 10.1080/07908310608668752

Yan, J., and Jinsong, F. (2011). Test for English majors (TEM) in China. *Lang. Test.* 28, 589–596. doi: 10.1177/0265532211414852

Yang, W. (2014). *Mapping the relationships among the cognitive complexity of independent writing tasks, L2 writing quality, and complexity, accuracy and fluency of L2 writing* [doctoral dissertation]. Georgia State University, Atlanta, GA, United States.

Yoon, H. J., and Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Q.* 51, 275–301. doi: 10.1002/tesq.296

Zhan, J., Sun, Q., and Zhang, L. J. (2021). Effects of manipulating writing task complexity on learners' performance in completing vocabulary and syntactic tasks. *Lang. Teach. Res.* 1–22. doi: 10.1177/13621688211024360

Zhang, X., Lu, X., and Li, W. (2022). Beyond differences: assessing effects of shared linguistic features on L2 writing quality of two genres. *Appl. Linguist.* 43, 168–195. doi: 10.1093/applin/amab007

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 12 July 2023
DOI 10.3389/fpsyg.2023.1127718



OPEN ACCESS

EDITED BY
Hui Zhang,
Nanjing Normal University, China

REVIEWED BY
Lin Fan,
Beijing Foreign Studies University, China
Zhi Zhu,
Yunnan Normal University, China

*CORRESPONDENCE
Josje Verhagen
✉ j.verhagen@uva.nl

RECEIVED 19 December 2022
ACCEPTED 05 June 2023
PUBLISHED 12 July 2023

CITATION
Verhagen J and de Bree E (2023) Non-adjacent dependency learning from variable input: investigating the effects of bilingualism, phonological memory, and cognitive control. *Front. Psychol.* 14:1127718. doi: 10.3389/fpsyg.2023.1127718



# Non-adjacent dependency learning from variable input: investigating the effects of bilingualism, phonological memory, and cognitive control

Josje Verhagen[1]* and Elise de Bree[2]

[1]Amsterdam Center for Language and Communication, University of Amsterdam, Amsterdam, Netherlands, [2]Department of Education and Pedagogy, Utrecht University, Utrecht, Netherlands

**Introduction:** One proposed advantage of bilingualism concerns the ability to extract regularities based on frequency information (statistical learning). Specifically, it has been proposed that bilinguals have an advantage in statistical learning that particularly holds in situations of variable input. Empirical evidence on this matter is scarce. An additional question is whether a potential bilingual advantage in statistical learning can be attributed to enhancements in phonological memory and cognitive control. Previous findings on effects of bilingualism on phonological memory and cognitive control are not consistent.

**Method:** In the present study, we compared statistical learning from consistent and variable input in monolingual and bilingual children (Study 1) and adults (Study 2). We also explored whether phonological memory and cognitive control might account for any potential group differences found.

**Results:** The findings suggest that there might be some advantage of bilinguals in statistical learning, but that this advantage is not robust: It largely surfaced only in t-tests against chance for the groups separately, did not surface in the same way for children and adults, and was modulated by experiment order. Furthermore, our results provide no evidence that any enhancement in bilinguals' statistical learning was related to improved phonological memory and cognitive control: bilinguals did not outperform monolinguals on these cognitive measures and performance on these measures did not consistently relate to statistical learning outcomes.

**Discussion:** Taken together, these findings suggest that any potential effects of bilingualism on statistical learning probably do not involve enhanced cognitive abilities associated with bilingualism.

KEYWORDS

**statistical learning, variable input, bilingualism, phonological memory, cognitive control**

## Introduction

The world's growing bilingual population fuels research into the potential advantages of bilingualism. One proposed advantage concerns statistical learning, or the ability to extract regularities based on frequency information in the input. In this study, we evaluate whether there is a bilingual advantage in statistical learning from linguistic input for children (Study 1) and adults (Study 2). Specifically, we compare statistical learning between monolingual and bilingual speakers from input that is consistent and input that is variable. We also investigate whether any effects of bilingualism on statistical learning are due to enhanced phonological memory and cognitive control.

## Bilingualism and statistical learning

Previous work has found that bilinguals fare better at extracting regularities based on frequency information in the input than monolinguals for various age groups, including 7-month-old infants (Kovács and Mehler, 2009), 14-month-old infants (Antovich and Graf Estes, 2017), 24-month-old toddlers (de Bree et al., 2017), 6- to 12-year-old children and adolescents (Bonifacci et al., 2011), and adults (Benitez et al., 2016; Bulgarelli and Weiss, 2016; Poepsel and Weiss, 2016; Potter et al., 2017; Onnis et al., 2018; for reviews, see Hirosh and Degani, 2017; Bulgarelli et al., 2018; Weiss et al., 2020). Such increased performance has been found for statistical learning from different types of auditory input, including tones, syllables, and Morse words.

However, other studies found no evidence for a bilingual advantage. Yim and Rudoy (2013), for instance, found no difference in performance between monolingual and sequential bilingual 5- to 13-year-olds in visual and auditory statistical learning tasks (see also Bogulski, 2013). Furthermore, in some studies, only partial evidence for a bilingual advantage was found. Bartolotti et al. (2011), for example, compared monolingual and sequential bilingual adults' performance on two learning tasks involving Morse words. In the first, in which Morse words were presented for the first time, bilinguals outperformed monolinguals. In the second, in which Morse code words were presented that conflicted with the words from the first task, performance was unaffected by bilingualism.

Vice versa, de Bree et al. (2017) assessed 24-month-old monolingual and native bilinguals' learning of patterns from auditorily presented nonwords in a condition in which the input was fully consistent and one in which the input was variable, containing 14% of "errors" of the pattern. The patterns in this study involved non-adjacent dependencies: relationships between two elements that are separated by another, intervening element. Non-adjacent dependencies have been studied relatively often in statistical learning and are frequently occurring in natural language in constructions of the type "He *is* read*ing*." There were no differences between the monolingual and bilingual toddlers in the consistent input condition, but bilinguals outperformed the monolinguals in the variable condition: the bilingual children showed a stronger sensitivity to the predominant pattern as opposed the other pattern than the monolingual children. Finally, Verhagen and de Bree (2021) found that native bilingual 4- and 5-year-olds fared better than their monolinguals peers on a reaction-time based measure but not an accuracy-based measure in an auditory non-adjacent dependency learning task. There are thus mixed patterns of results regarding effects of bilingualism on statistical learning.

One possible explanation is that bilinguals' enhanced abilities are especially prominent in, or limited to, situations where the input is not uniform, but contains variability (Poepsel and Weiss, 2016). This proposal receives some preliminary support from studies that show that statistical learning from more complex input involving multiple and potentially competing cues is enhanced in bilinguals (Kovács and Mehler, 2009; Bartolotti et al., 2011; Wang and Saffran, 2014; Poepsel and Weiss, 2016; de Bree et al., 2017), but that no such advantage is found in more basic forms of statistical learning. Bilingual learners have to detect the regularities of their two languages on the basis of more limited input than monolingual learners. Furthermore, this input might contain more variation, as the languages might be spoken with more limited linguistic proficiency (Byers-Heinlein and Fennell, 2014). This might mean that bilinguals are better equipped than monolinguals to learn patterns on the basis of more complex and variable cues (Kuo and Anderson, 2012; Kuo et al., 2016). However, to date, no studies have directly compared bilingual children and adults' performance on tasks with variable as opposed to invariable input.

## Cognitive functions

Another open, and potentially related, issue refers to the mechanisms underlying bilinguals' potential advantage (Poepsel and Weiss, 2016; Weiss et al., 2020). Bilinguals' advanced ability could be a direct effect of experience with dual language input, which is typically more complex than single language input as it involves two different language systems (Hirosh and Degani, 2017) with potentially differing quality of input, or an indirect effect, through enhanced cognitive abilities associated with bilingualism (Kovács and Mehler, 2009; Kuo and Kim, 2014; Hirosh and Degani, 2017). These possibilities are not mutually exclusive, as both direct and indirect effects could be at play.

Regarding enhanced cognitive abilities, two functions have been proposed as possible candidates for explaining why bilinguals may learn from variable input more readily than monolinguals: phonological memory and cognitive control (Bartolotti et al., 2011). Phonological memory refers to the ability to store verbal information in short-term memory, and is connected to two processes that are assumed to underlie statistical learning: extraction and integration (Thiessen and Erickson, 2013). Extraction refers to the process of holding statistically congruent clusters in memory (Perruchet and Tillmann, 2010); integration to the process of combining information across these clusters to identify regularities in the input. Verbal working memory, encompassing phonological memory, is considered essential in extraction: participants store exemplars in memory, and integrate information from these prior exemplars (Thiessen, 2017). Features that are consistent across them are strengthened, and features that are inconsistent across them are weakened, leading to knowledge of statistical regularities (Thiessen and Pavlik, 2013). Individual differences in phonological memory have shown positive associations with statistical learning in monolingual children (Kapa and Colombo, 2014) and adults (Karpicke and Pisoni, 2004; Misyak et al., 2010). However, other studies showed no such associations (Kaufman et al., 2010; Siegelman and Frost, 2015; Verhagen and de Bree, 2021).

Cognitive control refers to a set of processes needed to selectively attend to (relevant) stimuli and inhibit or suppress attention to other (less relevant) stimuli. It is typically assessed with tasks in which participants respond to specific targets amongst distracting stimuli or inhibit dominant responses. Cognitive control has been found to predict artificial language learning in monolingual adults and children (Kapa and Colombo, 2014), and

might play an even more important role when input is variable: well-developed cognitive control could allow participants to focus their attention selectively on parts of the input, hold partially conflicting information in memory, and suppress less relevant or conflicting information during training and/or at test. In fact, this association between cognitive control and selectively attending to only relevant properties of the language input has been proposed in earlier research (Kuo and Anderson, 2012).

Two studies have tested the suggestion that bilinguals' advantage in statistical learning may be due to enhanced phonological memory and/or cognitive control. Bartolotti et al. (2011) found that cognitive control was positively related to learning Morse code words when word meanings conflicted with meanings learned previously. However, in their study, enhanced cognitive control did not explain bilinguals' advantage on statistical learning; the bilinguals did not outperform the monolinguals on cognitive control. Similarly, in their investigation of auditory non-adjacent dependency learning in monolingual and bilingual kindergarteners, Verhagen and de Bree (2021) found that the bilingual children showed enhanced performance on (part of the) statistical learning measures, but not on a phonological memory task. Furthermore, this study showed that while phonological memory abilities were correlated with statistical learning performance, they did not account for bilinguals' enhanced performance. What remains unknown from these earlier studies, however, is whether phonological memory and cognitive control relate more strongly to statistical learning tasks in which variable input as opposed to consistent input is presented: Bartolotti et al. (2011) did not compare learning from consistent and variable input, even though they had a condition where meanings contrasted with previously learned meanings, and Verhagen and de Bree (2021) looked at statistical learning from consistent input only.

## The present research

It is currently unknown whether bilinguals have an advantage in statistical learning that particularly holds in situations of variable input and, if so, if this advantage can be attributed to enhancements in phonological memory and cognitive control. In the present study, we compare statistical learning from consistent and variable input in monolingual and bilingual children and adults, and explore whether phonological memory and cognitive control might account for any potential group differences found. Note that this second aim is exploratory, as the literature on cognitive advantages of bilingualism is heavily mixed, and the evidence for effects of bilingualism on phonological memory and cognitive control not at all robust (Paap et al., 2015).

# Study 1: children

In Study 1, monolingual and bilingual children were tested on two auditory statistical learning tasks involving non-adjacent dependency patterns: one in which the input was consistent and one in which the input contained exceptions to a predominant pattern, rendering the input variable. We predicted that both groups would be able to learn non-adjacent dependencies from consistent input, based on earlier results showing that very young children are able to do so (Gómez, 2002; Gómez and Maye, 2005). We did not have a prediction for the variable input condition, in the absence of earlier research using similar tasks (except for toddlers, see de Bree et al., 2017). However, we expected that if an advantage for the bilingual participants was found, it would be most prominent for the variable input task (Poepsel and Weiss, 2016). As to relationships with phonological memory and cognitive control, we did not have a clear prediction: while some studies found that these cognitive skills are implicated in statistical learning in both children and adults (Bartolotti et al., 2011; Kapa and Colombo, 2014; de Bree et al., 2017), others did not find such relationships (Verhagen and de Bree, 2021) or failed to show robust effects of bilingualism on these cognitive skills (Paap et al., 2015). Given these mixed findings, we kept this final question exploratory.

# Method

## Participants

Participants were 53 children with a mean age of 8;5 years ($SD$ = 1;1, min-max = 6;9−10;9). We based our sample size on earlier studies with similar designs and age ranges that found effects of bilingualism on statistical learning. These had similarly sized samples as ours (Bartolotti et al., 2011; Wang and Saffran, 2014) or smaller samples (Poepsel and Weiss, 2016; de Bree et al., 2017). Children between 6 and 10 years were recruited, because earlier work had shown that children of this age range can conduct the non-adjacent dependency task used in our study (Hakvoort, 2009) and because children in this age range are in the same, primary school, phase as the participants in Bonifacci et al. (2011) and Yim and Rudoy (2013). 25 children were monolingual (mean age: 8;4 years, $SD$ = 1;0) and 28 were bilingual (mean age: 8;5 years, $SD$ = 1;2). Age did not differ between the groups [$t_{(1,51)} = 0.061$, $p = 0.952$, $d = 0.017$]. The monolingual group contained 13/26 (50%) boys and the bilingual group contained 11/28 (39%) boys. This difference in gender was not significant ($\chi^2(54) = 0.627$, $p = 0.429$).

Children had been recruited through primary schools offering either bilingual or monolingual education in the Netherlands as well as through personal contacts, and the participant database of the [Utrecht University Babylab]. The monolingual children all came from monolingual Dutch homes and had not been in regular contact with another language than Dutch, as indicated in a parent questionnaire. The bilingual children learned Dutch as well as one out of a diverse set of other languages at home: English ($n$ = 9), Turkish ($n$ = 4), Russian ($n$ = 3), Armenian ($n$ = 3), German ($n$ = 2), Spanish, Italian, Sranan Tongo, French, Limburgian, Bulgarian, Romanian (all $n$ = 1). All children had been exposed to their other language from birth. Exposure to each language varied, as indicated by parents' responses in the questionnaire that were available for 20 out of 28 children: 12 children were mostly exposed to their other language and sometimes to Dutch; 6 children were mostly exposed to Dutch and sometimes to their other language, and 2 children were equally exposed to each language. Two children were multilingual as they spoke two languages other than Dutch at home (one child spoke Dutch, English, and

Romanian; another child spoke Dutch, Italian and Swedish). Dutch receptive vocabulary, as indicated through children's raw scores on the Dutch Peabody Picture Vocabulary Test (PPVT-III-NL, Dunn and Dunn, 2005), [$M = 113.44$, $SD = 11.26$ for monolinguals; $M = 109.82$, $SD = 16.04$ for bilinguals] and controlled for age, did not differ significantly between the groups, $F_{(1,50)} = 1.281$, $p = 0.263$). One monolingual child performed only the consistent input NADL experiment; two children (one monolingual, one bilingual) performed only the variable input NADL experiment, due to illness. Written informed consent was obtained from children's parents.

## Materials

### Non-adjacent dependency learning (NADL) experiments

#### Consistent input NADL

In the consistent input NADL experiment, participants listened to a miniature artificial language. This was modeled after the languages used in previous studies on non-adjacent dependency learning in English children and adults (Gómez, 2002; Gómez and Maye, 2005) and the same as in de Bree et al. (2017). Prior to the experiment, children were told that they were going to listen to a robot that would speak an odd language and informed that they should pay attention to the ordering of the elements in the language. This instruction was included based on previous studies with similar aged groups (Hakvoort, 2009) showing that children otherwise did not understand the task. Children colored a robot while listening to the artificial language (Saffran et al., 1997; Grama et al., 2016).

The language was presented on a laptop computer and through headphones. It consisted of three-item strings that took the form a-X-b or c-X-d. The elements a, b, c, and d represented the novel words *rak*, *toef*, *sot* and *lut*, and X was drawn from a pool of 24 novel words (*wadim*, *kasi*, *poemer*, *kengel*, *domo*, *loga*, *gopem*, *naspu*, *hiftam*, *dieta*, *vami*, *snigger*, *rogges*, *densim*, *fidang*, *rajee*, *seeta*, *noeba*, *plizet*, *banip*, *movig*, *sulep*, *nilbo*, and *wiffel)*. A set size of 24 X-elements was chosen because this yielded the strongest learning effects in previous studies (Gómez and Maye, 2005; Hsu et al., 2014). In the training phase of the experiment, participants were either presented to language 1 that contained the triplets a-X-b and c-X-d (i.e., *rak*-X-*toef*, *sot*-X-*lut*) or to language 2 that contained the triplets a-X-d and c-X-b (i.e., *rak*-X-*lut*, *sot*-X-*toef*). These two language versions were used to rule out any potential effects of the phonological properties of the stimuli or stimuli combinations. There were seven iterations of each of the 48 triplets (2 dependencies per language * 24 X-elements), resulting in a total of 336 triplets per language (see Table 1). The training phase lasted about 15 mins.

Triplets had been created on the basis of novel words that were spliced from triplets recorded from a female native speaker of Dutch, and subsequently, combined into new triplets for both languages. Consequently, the two languages did not differ in pronunciation and there were no speaker-induced differences in individual triplets. A 250-ms inter-stimulus interval occurred in between the three nonwords of each triplet. To ensure that the three nonwords were perceived as one triplet, a 750-ms interval occurred between triplets. In the test phase of the experiment that directly followed the training phase, a forced-choice selection task was presented, in which participants listened to eight pairs of triplets. Each pair contained two spoken sentences that were played in turn: one from the language presented during training (trained triplet) and one from the other language (untrained triplet) (see Appendix Table A1). Participants were asked to indicate for each pair which triplet matched the language they had listened to in the training phase by pressing one of two response buttons on a laptop keyboard. Only the X-elements *wadim, kasi, poemer* and *rogges* were used in the test phase. Pairs were presented in pseudo-randomized lists in which no more than two elements of the same type were presented consecutively and the ordering of triplets within pairs was counterbalanced across participants. Throughout the experiment, stimuli presentation and response logging were controlled through E-prime 2.0 (Psychology Software Tools).

#### Variable input NADL

This experiment was the same as the consistent input NADL experiment, except that a portion of the triplets was inconsistent with the logic of the artificial language. Specifically, 48 out of 336 triplets (14%) contained "errors" in that they came from the other language. For example, if participants were trained on language 1 (i.e., *rak*-X-*toef*, *sot*-X-*lut*), they would hear 48 instances of incorrect *rak*-X-*lut* from language 2 that were intermixed with the correct triplets from language 1. These "errors" were randomly picked from a list and inserted at fixed, pseudo-random positions in the training. Following de Bree et al. (2017), for only one of the two dependencies, an alternative was presented in which the structure had been disrupted. So, in language 1, participants were presented with *rak-X-toef* and *sot-X-lut* as the predominant pattern (86% of the items) and with incorrect **rak-X-lut* in 14% of the items (see Table 1).

As in the consistent input experiment, the forced-choice selection task in the test phase of the experiment contained eight item pairs. Four of these involved a contrast between a triplet from the training phase (trained triplet) and a triplet from the other language (untrained triplet) and thus were identical to the test pairs in the consistent input NADL experiment. The other four involved a contrast between a trained triplet and a "noise triplet," and thus involved a contrast between items that had both been presented during the test phase, but with different frequencies. The first type of item was included to assess learning of the non-adjacent dependency rules, but under more challenging conditions than in the consistent input experiment. The second type was included to see whether participants could identify the more frequent triplet. These items were not included to address directly our research question on whether participants would learn the predominant pattern, but to see if participants could distinguish between the two triplets. This would signal that they were sensitive to the relative frequencies of both types of triplet. As in the consistent input experiment, only four X-elements were used during the test phase, and items were presented in pseudorandomized and counterbalanced lists. Specifically, no more than two trained-untrained items or trained-noise items were presented after one another and the order of triplets within items was counterbalanced across participants.

TABLE 1 Stimuli of the training phase in the consistent and variable input NADL experiments.

| Experiment | Number of triplets | Language1 | | Language2 | |
|---|---|---|---|---|---|
| Consistent input | 336 | a-$X_{(1-24)}$-b | c-$X_{(1-24)}$-d | a-$X_{(1-24)}$-d | c-$X_{(1-24)}$-b |
| | | (rak X toef) | (sot X lut) | (rak X lut) | (sot X toef) |
| Variable input | 288 | a-$X_{(1-24)}$-b | c-$X_{(1-24)}$-d | a-$X_{(1-24)}$-d | c-$X_{(1-24)}$-b |
| | | (rak X toef) | (sot X lut) | (rak X lut) | (sot X toef) |
| | 48 | a-$X_{(1-24)}$-d* | | a-$X_{(1-24)}$-b* | |
| | | (rak X lut)* | | (rak X toef)* | |

*Refers to "noise" triplets. X refers to the different X-items used.

### Nonword repetition

The nonword repetition (NWR) task by Rispens and Baker (2012) was used to assess phonological memory. In this task, children hear a prerecorded nonword over headphones and are then asked to repeat it. The task contains 40 items that range between two and five syllables (ten of each type). Items are pseudo-randomly divided into two blocks of twenty items, with a short break in between. Children's responses were recorded and coded as (in)correct. Cronbach's alpha was.77. Scores were computed as the number of correct responses (maximum score: 40).

### Flanker task

Cognitive control was assessed with a Dutch version of the Flanker task used by Engel de Abreu et al. (2012), in which horizontal rows of five equally spaced yellow fish are presented on a laptop screen. Children have to indicate the direction of the central fish by pressing the corresponding left or right response button on each side of the laptop keyboard as quickly as possible. On congruent items (50% of items), the flanking fishes point in the same direction as the central fish. On incongruent items (50% of items), the flanking fishes point in the opposite direction. Each item starts with a 1-second fixation cross in the middle of the screen, followed by the fish array for five seconds or until a response is made. Responses are followed by a 400-ms blank interval. There are two blocks of 20 items each with randomized presentation of congruent and incongruent items, preceded by eight practice items. Reaction times and accuracy were recorded through E-Prime 2.0 (Psychology Software Tools).

## Procedure

Children were tested individually twice by a research assistant in a quiet room at home or school. The sessions were about 40 mins each, with at least one to two weeks in between. The order of the NADL experiments was counterbalanced across sessions. Twenty-five children performed the consistent input experiment in the first session (11 bilinguals; 14 monolinguals); 28 children performed the variable input experiment in the first session (17 bilinguals; 11 monolinguals). Tasks were presented in a fixed order within sessions: the consistent input experiment preceded the Flanker task, and the variable input experiment preceded the NWR task. Children received a sticker after each task and a small gift at the end of the session.

Written informed consent was obtained from the parents before testing; consent and participation could be retracted at any time. The research was conducted in accordance with American Psychological Association ethical standards as well as The Netherlands Code of Conduct for Scientific Practice issued in 2004 (revised in 2018 by the Association of Universities in The Netherlands).

## Analyses

We first checked whether performance on the NADL experiments was significantly above chance in each group through *t*-tests against the 50% chance level. Then, we ran a generalized linear-mixed effect regression model on participants binary scores (correct vs. incorrect) in the forced-choice selection tasks in each experiment, using the *lme4* package (Bates et al., 2015) in R version 3.4.1 (R Core Team, 2017). As fixed effects, we included "group" (monolingual vs. bilingual), "version" (consistent vs. variable), "experiment order" (consistent input experiment first vs. second), and "age" (in years). Effects of "language" (language 1 vs. language 2) were explored, but not retained, because this factor did not have an effect and yielded a less well-fitting model, as indicated by a higher AIC-value. By-item random intercepts were included, to obtain the maximal random effect structure supported by the data. As a further exploratory analysis, we ran a similar model on the data of the variable input experiment only, to see whether group interacted with item type (trained-untrained pairs vs. trained-noise pairs). This analysis was included to yield a more complete picture of participants' knowledge of the relative frequencies of the dependencies presented in the variable input experiment.

To assess whether individual differences in phonological memory and cognitive control related to participants' learning of the dependency relations in the two groups, we first excluded reaction times below 200 ms and above three standard deviations of children's individual means (<1.8% of all items) for the Flanker task, following Engel de Abreu et al. (2012). Also, following Engel de Abreu et al., accuracy scores were computed, but not analyzed because they were at ceiling (95% correct or higher). Mean reaction times on correct items were calculated for the (in)congruent items separately. Next, we tested for effects of group on participants' scores on the NWR and Flanker task through a *t*-test and a linear model with item type (congruent vs. incongruent) and group as fixed effect factors and by-subject random intercepts,

TABLE 2 Descriptive statistics for the forced-choice selection tasks in the NADL experiments.

| Experiment | Item type | Monolinguals | | Bilinguals | |
|---|---|---|---|---|---|
| | | M | (SD) | M | (SD) |
| Consistent input | Trained—untrained | 0.53 | (0.22) | 0.63 | (0.17) |
| Variable input | Trained—untrained | 0.46 | (0.23) | 0.47 | (0.24) |
| | Trained—noise | 0.45 | (0.25) | 0.43 | (0.24) |

respectively. Subsequently, we calculated bivariate correlations between scores on the NWR and Flanker task and summed accuracy scores in the NADL experiments for the monolinguals and bilinguals separately. Finally, to examine how phonological memory and cognitive control related to statistical learning as well as any effects of bilingualism on statistical learning, we ran the same model as above, with the NWR and Flanker scores as additional fixed effect scores. In all mixed-effect models, orthogonal sum-to-zero contrast coding was applied to our fixed effects "group" (bilinguals: −1/2, monolinguals +1/2), "experiment version" (consistent: −1/2, variable: +1/2) and "experiment order" (consistent input first: −1/2, consistent input last: +1/2) (Schad et al., 2020). Continuous predictors were centered around zero. All data files and scripts can be found at: https://osf.io/b4ps6/?view_only=a18f5b5cb1d04905b6c26f29de2f43b1.

# Results

## Results for NADL experiments

Descriptive statistics for the two NADL experiments are presented in Table 2.

*T*-tests comparing against the 50%-chance level showed that, in the consistent input experiment, the bilingual children performed above chance, but the monolingual children did not [monolinguals: $t_{(1,23)} = 0.636$, $p = 0.531$, $d = 0.130$; bilinguals: $t_{(1,25)} = 3.844$, $p = 0.001$, $d = 0.754$]. In the variable input experiment, neither of the groups performed above chance, neither on the trained-untrained trials [monolinguals: $t_{(1,22)} = -0.890$, $p = 0.383$, $d = -0.189$; bilinguals: $t_{(1,26)} = -0.593$, $p = 0.558$, $d = -0.114$] nor on the trained-noise trials [monolinguals: $t_{(1,23)} = -1.045$, $p = 0.307$, $d = -0.213$, bilinguals: $t_{(1,26)} = -1.615$, $p = 0.118$, $d = -0.331$].

A generalized linear mixed-effect model with "group," "experiment version," and "experiment order" as fixed effects, and "age" as a fixed effect control factor, showed no main effect of group ($\beta = -0.250$, SE = 0.179, $z = -1.400$, $p = 0.162$) or experiment order ($\beta = -0.066$, SE = 0.179, $z = -0.367$, $p = 0.714$). A main effect of experiment version indicated that children performed better on the consistent than variable input experiment ($\beta = -0.412$, SE = 0.184, $z = -2.245$, $p = 0.025$). There also was an interaction effect between group, experiment version and experiment order ($\beta = 1.624$, SE = 0.717, $z = 2.267$, $p = 0.023$), which indicated that the difference in performance across the two experiment versions was larger for the bilinguals than monolinguals and interacted with experiment order: for the bilinguals, the difference was largest when they performed the variable input experiment first, while for the monolinguals it was largest when they performed the consistent input experiment first. For descriptives per experiment plotted by experiment order (see Appendix Figure B1). The other effects and interactions were not significant (see Appendix Table B1). A model on children's scores on the scores on the variable input experiment only with "item type" (trained-untrained vs. trained-noise) showed no effects (see Appendix Table B2).

## Statistical learning and relationships with phonological memory and cognitive control

Descriptive statistics for the NWR and Flanker tasks are presented in Table 3. Data were available for all children, except one monolingual child.

The numerically slightly higher NWR scores of the bilingual children than the monolingual children were not significantly different [$t_{(1,50)} = 1.529$, $p = 0.133$, $d = 0.419$]. Regarding the Flanker scores, a linear model with item type (congruent vs. incongruent) and group as fixed effect factors and by-subject random intercepts showed an effect of item type ($\beta = 98.44$, SE = 17.41, $t = 5.654$, $p < 0.001$), indicating that children responded more slowly to the incongruent than congruent items, but no effect of group ($\beta = -154.60$, SE = 100.69, $t = -1.535$, $p = 0.131$) or interaction between item type and group ($\beta = -11.80$, SE = 34.83, $t = -0.339$, $p = 0.736$).

In the absence of effects of group, it was unlikely that differences in phonological memory and cognitive control could account for the slight advantage of the bilingual children in NADL—which was only observed in the *t*-tests across chance level in the consistent input experiment. Yet, to rule out this possibility, we calculated partial (age-controlled) correlations between the scores in the NADL experiments and the NWR and Flanker scores. Correlations between NWR and NADL were weak and non-significant. Monolinguals' performance on the Flanker task (incongruent trials and difference score) correlated negatively with performance in the variable input experiment, indicating that children who performed relatively well on the Flanker task had relatively good performance in this experiment. For the full correlation matrix (see Appendix Table B3).

On the basis of these data, it seems unlikely that phonological memory and cognitive control played a major role in the bilinguals' higher performance in the variable input experiment. Indeed, adding the NWR and Flanker scores as additional fixed-effect factors to our previous mixed-effect model yielded no effects of NWR ($\beta = 0.006$, SE = 0.009, $z = 0.655$, $p = 0.512$) or

TABLE 3 Descriptive statistics for the nonword repetition task and flanker task.

| | Monolinguals | | Bilinguals | |
|---|---|---|---|---|
| | *M* | *(SD)* | *M* | *(SD)* |
| *Nonword repetition task* | | | | |
| Number correct (maximum = 40) | 26.32 | (5.48) | 28.71 | (5.93) |
| *Flanker task* | | | | |
| RT congruent items | 860.12 | (301.76) | 1008.81 | (353.41) |
| RT incongruent items | 952.66 | (355.67) | 1113.16 | (435.31) |
| RT difference score | 92.54 | (116.87) | 104.34 | (131.86) |

Flanker scores ($\beta = 0.071$, SE = 0.075, $z = 0.956$, $p = 0.339$). The previously found effect of version remained, indicating that children obtained higher scores in the consistent input than variable input experiment, irrespective of group ($\beta = -0.410$, SE = 0.185, $z = -2.219$, $p = 0.027$), as did the interaction between group, experiment version and experiment order ($\beta = 1.635$, SE = 0.722, $z = 2.265$, $p = 0.024$). For the full results (see Appendix Table B4).

## Summary study 1

We investigated whether bilingual children showed enhanced statistical learning, particularly in learning from variable input. Our results (see an overview in Table 6) suggested better performance for the bilinguals only in the consistent input experiment, but only through *t*-tests. In a mixed-effect regression analysis, there were no effects of group and no interaction between group and experiment. Instead, a complex interaction between group, experiment version and experiment order was found that we will turn to in the Discussion.

We also assessed whether a potential statistical learning advantage in learning from variable input was due to potentially better performance on phonological memory (NWR) and cognitive control (Flanker) in the bilingual group. However, the bilingual children did not show better performance on the NWR and the Flanker task than the monolingual group. Furthermore, there was no association between statistical learning and the cognitive abilities (NWR and Flanker).

## Study 2: adults

In Study 2, we investigated the same questions as in Study 1, in adults. We predicted that both monolingual and bilingual adults would be able to learn non-adjacent dependencies from consistent input, as evidenced by above-chance performance, based on earlier results for English-speaking adults (Gómez, 2002; Gómez and Maye, 2005) and Dutch-speaking adults (Grama et al., 2016). Furthermore, we initially predicted that any advantage for the bilinguals would be most prominent for the variable input task (Poepsel and Weiss, 2016). Given that this prediction was not borne out for the children in Study 1, we were not sure what to expect for the adults. Regarding phonological memory and cognitive control, we had no clear predictions either, given that our already tentative prediction in Study 1 was not supported by the data.

## Method

### Participants

Participants were 54 adults with a mean age of 26;0 years (*SD* = 0;6, min-max = 19–37). Of these, 26 were monolingual Dutch and 28 were bilingual (Dutch + other language). As in Study 1, sample size was based on earlier studies with similar test designs that attested effects of bilingualism on statistical learning and had similarly sized samples (Bartolotti et al., 2011; Wang and Saffran, 2014) or smaller samples (Poepsel and Weiss, 2016; de Bree et al., 2017). Participants were recruited via research assistants' friends, acquaintances, and families. They were classified as monolingual if they used only Dutch at home and did not speak another language than Dutch with friends or families regularly. Participants were classified as bilingual if they spoke Dutch and another language(s) at home on a daily basis, with friends/families or at work. The bilingual participants spoke one out of a set of the following languages, next to Dutch: Armenian ($n = 16$), English ($n = 4$), German ($n = 4$), Arabic ($n = 1$), Spanish ($n = 1$), French ($n = 1$), Hebrew ($n = 1$). Participants reported high proficiency levels in Dutch, as rated on a five-point scale ranging from 1 "zero proficiency" to 5 "fluent," with an average of 4.33 ($SD = 0.80$) for questions assessing speaking and listening.

For the bilinguals, self-reported proficiency in the other language was also generally high ($M = 4.80$, $SD = 0.57$). Three participants reported higher proficiency in Dutch than their other language; the remaining participants reported equally high proficiency or higher proficiency in their other language. Twenty of the bilingual participants had acquired their other language prior to Dutch, six had acquired Dutch first, and two bilinguals had acquired both languages simultaneously from birth. For the 26 bilinguals who had learned their languages successively, sixteen had acquired their second language before the age of twelve. Four participants used more than two languages on a daily basis at home ($n = 2$ Dutch/Armenian/Russian, $n = 2$ Dutch/Armenian/Arabic). Although these latter participants were thus multilingual rather than bilingual, we refer to them as bilinguals in this study.

Mean ages [monolingual: 25;10 years ($SD = 0;5$); bilingual: 26;3 years ($SD = 0;7$)] did not differ significantly between the

groups [$t_{(52)} = 0.262$, $p = 0.794$, $d = 0.072$]. The distribution of sex (monolingual: 8/26 (31%) males; bilingual:14/28 (50%) males) did not differ either ($\chi^2(54) = 2.065$, $p = 0.151$). Participants' mean highest level of education, as established on a scale with "1" (primary school) to "6" (university) as its scale points, did not differ significantly between groups (monolingual: $M = 5.00$ ($SD = 1.35$), available for 25/26 monolinguals; bilingual: $M = 4.68$ ($SD = 1.36$), available for 28/28 participants)—[$t_{(1,51)} = 0.860$, $p = 0.384$, $d = 0.237$]. One monolingual participant and two bilingual participants performed only the variable input experiment, due to illness.

## Materials

### NADL experiments

#### Consistent input NADL

This experiment was the same as the consistent input experiment in Study 1, except for the instructions; participants were told that they were going to listen to an odd language, and informed that they had to answer some questions about the language later on. During listening, participants colored a mandala.

#### Variable input NADL

This experiment was the same as the variable input experiment used in Study 1, except for the instructions, which were the same as in the consistent input experiment for the adults.

### Phonological memory and cognitive control

#### Nonword repetition

The Dutch Nonword Repetition Test (NRT) was administered to assess participants' phonological memory ability (De Jong, 1998). In this test, participants repeat pre-recorded nonwords. The test contains two practice items and 48 nonwords that vary in length from two to five syllables (twelve nonwords of each type). The audio files were implemented in the experimental software E-prime 2.0 and administered to participants on a laptop via headphones. Responses were coded as (in)correct, and scores were computed as the number of correct responses (maximum = 48). Cronbach's alpha was 0.88.

#### Trail making test

The Trail Making Test (TMT) was used to assess cognitive control (Reitan, 1956). In part A of this test, subjects are asked to draw lines to connect 25 circles containing numbers (1–25) distributed over a sheet of paper in ascending order. In part B, the circles contain both numbers (1–13) and letters (A-L), and subjects are asked to connect the circles in ascending order, while alternating between numbers and letters (1-A-2-B-3-C, etc.). Participants are instructed to do this as fast as possible, without lifting their pen from the paper. Scores are: (i) the time in seconds it takes participant to connect the "trail" in part A, (ii) the time in seconds that it takes to connect the "trail" in part B, and (ii) the difference between the scores for parts B and A. Part A mainly assesses visuo-perceptual processing, part B primarily working memory and secondarily task switching, and the B-A difference score cognitive control (Sánchez-Cubillo et al., 2009).

## Procedure

Participants were assessed individually at a quiet place at their home or university in two sessions that were one and two weeks apart. Administration of the consistent and variable input experiments was counterbalanced across sessions. Thirty participants performed the consistent input experiment first (15 monolinguals; 15 bilinguals); 24 participants performed the variable input experiment first (13 bilinguals; 11 monolinguals). Task order was fixed within sessions: the consistent input experiment preceded the TMT, and the variable input experiment preceded the NRT.

Written informed consent was obtained from the participants before testing; consent and participation could be retracted at any time. The research was conducted in accordance with American Psychological Association ethical standards as well as The Netherlands Code of Conduct for Scientific Practice issued in 2004 (revised in 2018 by the Association of Universities in The Netherlands).

## Analyses

To assess the two groups' performance on the NADL experiments, the same analyses were performed as in Study 1: We first conducted *t*-tests against the 50% chance level on the two NADL tasks separately. We then ran a generalized linear-mixed effect regression model on participants accuracy scores (correct vs. incorrect) in the forced-choice selection tasks with "group" (monolingual vs. bilingual), "experiment" (consistent input vs. variable input), and "experiment order" (consistent input first vs. consistent input last) as fixed effects. By-subject and by-item random intercepts were included. We also ran a similar model on the data of the variable input experiment only, to test for effects of group on participants' performance on the trained-untrained as opposed to the trained-noise items.

To address our second question of how individual differences in phonological memory and cognitive control related to participants' learning scores, we performed the same analysis as in Study 1 (i.e., testing for group effects on NWR and TMT; bivariate correlations between NWR/TMT, and NADL scores; mixed-effect regression with NWR and TMT scores as fixed effect factors). All data files and scripts can be found at: https://osf.io/b4ps6/?view_only=a18f5b5cb1d04905b6c26f29de2f43b1.

# Results

## Results for NADL experiments

Descriptive statistics for both NADL experiments are presented in Table 4.

*T*-tests comparing performance against the 50% chance level showed that, in the consistent input experiment, both groups

TABLE 4 Descriptive statistics (proportions correct) for the NADL experiments.

| Experiment | Item type | Monolinguals | | Bilinguals | |
|---|---|---|---|---|---|
| | | M | (SD) | M | (SD) |
| Consistent input | Trained-untrained | 0.64 | (0.26) | 0.63 | (0.23) |
| Variable input | Trained-untrained | 0.56 | (0.28) | 0.63 | (0.28) |
| | Trained-noise | 0.51 | (0.26) | 0.53 | (0.28) |

TABLE 5 Descriptive statistics for the nonword repetition test and trail making test.

| | Monolinguals | | Bilinguals | |
|---|---|---|---|---|
| | M | (SD) | M | (SD) |
| **Nonword repetition test (NRT)** | | | | |
| Number correct (max = 48) | 33.84 | (5.74) | 34.08 | (7.75) |
| **Trail making test (TMT)** | | | | |
| Part A (time in sec.) | 18.96 | (7.10) | 26.89 | (8.81) |
| Part B (time in sec.) | 41.23 | (19.54) | 53.70 | (19.58) |
| B—A difference score | 22.27 | (15.61) | 26.81 | (14.37) |

performed above chance [monolinguals: $t_{(1,23)} = 2.781$, $p = 0.012$, $d = 0.555$; bilinguals: $t_{(1,24)} = 2.783$, $p = 0.010$, $d = 0.557$]. In the variable input experiment, the monolinguals did not perform above chance when presented with trained and untrained triplets [$t_{(1,25)} = 1.100$, $p = 0.282$, $d = 0.216$], but the bilinguals did [$t_{(1,25)} = 2.409$, $p = 0.024$, $d = 0.472$]. On the items in the variable input experiment that involved a contrast between a trained and a noise triplet, neither of the two groups performed above chance [monolinguals: $t_{(1,25)} = 0.189$, $p = 0.852$, $d = 0.037$, bilinguals: $t_{(1,25)} = -0.531$, $p = 0.600$, $d = 0.107$]. A generalized linear mixed-effect model on the correct/incorrect scores on the "trained-untrained" items with "group," "experiment version" and "experiment order" as fixed effects showed no main effects of group ($\beta = -0.020$, SE = 0.310, $z = -0.065$, $p = 0.948$), experiment version ($\beta = -0.311$, SE = 0.212, $z = -1.466$, $p = 0.143$) or experiment order ($\beta = -0.272$, SE = 0.309, $z = -0.879$, $p = 0.379$). The interaction between group and experiment version was not significant either ($\beta = -0.745$, SE = 0.401, $z = -1.858$, $p = 0.063$). The only effect found was a three-way interaction between group, experiment version and experiment order ($\beta = -1.880$, SE = 0.803, $z = -2.341$, $p = 0.019$). Monolinguals showed a larger difference in performance between the consistent and variable input experiment than the bilinguals and this interaction was related to experiment order: for the monolinguals, the difference in performance across versions was largest when they performed the variable input experiment first, whereas order did not matter for the bilinguals (see Appendix Figure C1). For the full results of the model (see Appendix Table C1). An analysis on the variable input experiment only in which "item type" was included and "experiment version" and "experiment order" were left out showed no main effects of group ($\beta = -0.226$, SE = 0.231, $z = -0.977$, $p = 0.328$), item type ($\beta = -0.347$, SE = 0.215, $z = -1.613$, $p = 0.107$), and no interaction between group and item type ($\beta = 0.233$, SE = 0.407, $z = 0.571$, $p = 0.568$) (see Appendix Table C2).

## Statistical learning and relationships with phonological memory and cognitive control

Descriptive statistics for the NRT and TMT are presented in Table 5. NRT scores were missing for four participants due to illness ($n = 2$) or experiment error ($n = 2$); TMT scores were missing for one (Hebrew-Russian-Dutch-speaking) participant who had not automatized the alphabet and therefore had trouble completing part B of the task in which numbers and letters had to be connected in alternating order (1-A-2-B-3-C-4-D etc.).

The NRT scores did not differ between groups [$t_{(1,49)} = 0.125$, $p = 0.902$, $d = 0.035$]. On the TMT, the monolinguals were significantly *faster* than the bilinguals on both parts of the test [part A: $t_{(1,51)} = 3.614$, $p < 0.001$, $d = 0.991$, part B: $t_{(1,51)} = 2.231$, $p = 0.024$, $d = 0.637$]. They also obtained slightly lower TMT difference scores than the bilinguals, but this difference was not significant [$t_{(1,51)} = 1.102$, $p = 0.276$, $d = 0.303$].

These outcomes rendered it unlikely that bilinguals' slightly enhanced performance on the variable input experiment, visible only through above-chance performance, could be attributed to differences in phonological memory and cognitive control. However, to see how these cognitive skills related to statistical learning performance, we explored the bivariate correlations between the NRT and TMT scores and the scores on the NADL experiments, and added the NRT and TMT scores to the regression model above. The correlation matrix showed a significant moderate correlation between the TMT difference scores and performance on NADL consistent; participants with a smaller TMT difference score (indicating better cognitive control) tended to perform better on NADL consistent input. For the variable input experiment, no significant correlations were found in either group. For the full correlation matrix (see Appendix Table C3). When the NRT and TMT scores were added to the mixed-effect model presented

TABLE 6 Overview of statistically significant effects in study 1 and study 2.

| Task | Items | Study 1: children Results | Study 2: adults Results |
|---|---|---|---|
| NADL consistent | Trained—untrained | Bilinguals performed above chance. | Monolinguals and bilinguals performed above chance. |
| NADL variable | Trained—untrained | – | Bilinguals performed above chance. |
| | Trained—noise | – | – |
| | | Main effect of version: Participants generally scored higher on consistent than variable input experiment. Group * Version * Order Larger difference in performance between the experiment versions for bilinguals than monolinguals: for bilinguals, the difference was largest when variable input was first; for monolinguals when consistent input was first. | Group * Version * Order: Larger difference in performance between the experiment versions for monolinguals than bilinguals: for monolinguals, the difference was largest when variable input was first; for bilinguals, there was no difference in performance depending on experiment order. |
| Phonological memory (NWR) | | – | – |
| Cognitive control (TMT/Flanker task) | | – | Monolinguals outperformed bilinguals. |
| | | – | Higher cognitive control was positively associated with better statistical learning. |

above, as additional fixed-effect factors, there were still no effects of "group" ($\beta = -0.301$, SE $= 0.303$, $z = -0.995$, $p = 0.320$) or experiment version ($\beta = -0.363$, SE $= 0.219$, $z = -1.657$, $p = 0.098$). However, there was an effect of experiment order ($\beta = -0.673$, SE $= 0.307$, $z = -2.193$, $p = 0.028$), indicating that accuracy was higher when the inconsistent input experiment was presented first. The interaction effect between "group" and "experiment version" was now also significant ($\beta = -0.892$, SE $=$ $0.416$, $z = -2.156$, $p = 0.031$), indicating that the effect of group was largest for the variable input experiment. The above-found three-way interaction between "group," "experiment version" and "experiment order" remained ($\beta = -1.733$, SE $= 0.831$, $z =$ $-2.085$, $p = 0.037$). There was no effect of NRT on NADL performance ($\beta = -0.016$, SE $= 0.015$, $z = -1.070$, $p = 0.285$). There was a negative effect of the TMT difference score on NADL performance ($\beta = -0.237$, SE $= 0.106$, $z = -2.232$, $p =$ $0.026$), indicating that participants with well-developed cognitive control were generally better in distinguishing between trained and untrained items in the NADL experiments. For the full model results (see Appendix Table C4).

## Summary study 2

The main findings of Study 2 are summarized in Table 6. The first aim of this study was to assess whether bilingual adults show an advantage in statistical learning particularly in learning from variable input. The results we found were mixed: while only the bilingual group showed above-chance performance in the variable input experiment, the interaction between group and experiment version did not surpass the 0.05 alpha level in a regression analysis, unless scores on phonological memory (NRT) and cognitive control (TMT) were added to the analysis. Furthermore, an interaction between group, experiment version and experiment order was obtained, that we will return to in the Discussion.

The second aim of this study was to evaluate whether a potential statistical learning advantage in learning from variable input was due to enhanced phonological memory and cognitive control in the bilingual group. However, the bilingual group did not outperform the monolingual group on either NRT or TMT and there was only a positive association between TMT and NADL performance in the consistent input experiment. Adding the NRT and TMT scores to the regression model did not change the above results, except that the trend for an interaction between group and version in the model without NRT and TMT scores became significant. It also showed that participants with better cognitive control were more likely to perform better on the NADL experiments, regardless of group and type of input.

## General result summary

The main findings of Study 1 and Study 2 are summarized in Table 6.

## Discussion

We conducted two studies, one with children and one with adults, to target two questions. The first was whether bilinguals would display enhanced statistical learning, specifically in learning from variable input. The second was whether better statistical learning (from variable input) would be related to improved phonological memory and cognitive control. In both studies, statistical learning was assessed through a non-adjacent dependency learning task.

## Statistical learning in monolinguals and bilinguals

Our results for the children (Study 1) showed that only the bilingual children performed significantly above chance in the consistent input condition. Neither group (monolingual, bilingual) performed above chance in the variable input condition. Furthermore, there was no main effect of group and interaction between input (consistent/variable) and group, in a generalized linear mixed-effect analysis. Our results for the adults (Study 2) showed that both adult groups performed above chance on the consistent input condition, but only the bilingual group performed above chance in the variable input experiment. Although these findings seem to align with Poepsel and Weiss (2016) proposal that a bilingual advantage is especially prominent in situations where the input contains variability, the interaction between group and experiment version did not reach significance in a mixed-effect regression analysis ($p = 0.063$) without including phonological memory and cognitive control outcomes. Together, these findings speak to the previously reported mixed findings on non-adjacent dependency learning in bilingual compared to monolingual children (Yim and Rudoy, 2013; Verhagen and de Bree, 2021) and statistical learning from variable input (de Bree et al., 2017).

The absence of a robust bilingual advantage in the current study needs to be interpreted in light of some methodological issues. The finding that *t*-tests showed effects for one of the experiment versions only, but the regression analysis showed no clear interaction between group and consistent/variable input suggests that a limitation of our study is that there may have been insufficient power to find an effect. We had not conducted a power analysis prior to conducting this study. Instead, we based our sample size on previous studies into bilingualism and statistical learning. Bartolotti et al. (2011), for instance, collected data from 24 bilinguals, and Wang and Saffran (2014) report data of 24 bilinguals and 24 monolinguals. Our sample sizes were higher than in some earlier studies that used the same designs and reported effects: Poepsel and Weiss (2016) included 17 monolinguals, 17 Chinese-English, and 17 English-Spanish bilinguals and de Bree et al. (2017) included 24 monolinguals and only 14 bilinguals. Yim and Rudoy (2013)'s study consisted of a larger sample (63 monolinguals and 49 bilinguals). However, their sample size might be this large due to the considerable age range of their participants (5 to 13 years). Indeed, age was an important and significant predictor of auditory statistical learning in both groups in their study. Onnis et al. (2018) appears to be the only study on bilingualism and statistical learning in which a power analysis was conducted beforehand to establish that a sample of 55 bilingual (undergraduate) participants was necessary. There are challenges in conducting a priori power analyses, in terms of generalizability across designs and assumptions on which to base the analyses. However, reliance on power analyses is needed in future studies on effects of bilingualism on statistical learning to be more confident about the interpretations.

Our results included a three-way interaction in both studies between group, experiment version (consistent/variable input) and experiment order (consistent input experiment first or second). This interaction indicated that participants were influenced by prior experience with the stimuli in the experiment, even if they completed that experiment one to two weeks earlier, and that this influence differed across monolingual and bilingual groups. Moreover, the direction of the interaction order was different for children and adults, yielding a complex pattern of results. One tentative conclusion is that the bilinguals seemed to be less affected by experiment order than the monolinguals, at least in Study 2. The other interactions with experiment order were both hard to interpret and to relate to previous studies on non-adjacent dependency learning in monolingual and bilingual children: Earlier work using similar tasks with the exact same stimuli (but a different task design) only assessed learning from consistent (and not variable) input (Verhagen and de Bree, 2021), or kept experiment order constant (de Bree et al., 2017), such that all children completed the consistent before the variable input experiment.

The composition of our bilingual groups might also have influenced our results: participants in both studies constituted a group of participants speaking Dutch and a myriad of possible other languages. Furthermore, variability was likely present in language usage and proficiency across participants. Onnis et al. (2018) found that bilingual adults with more balanced proficiency in their two languages learned statistical patterns in two miniature grammars better than bilinguals who were dominant in one of their languages. Since we did not take into account individual differences in the bilingual participants' language dominance and proficiency, a limitation of our study is that we cannot draw any conclusions about the potential effects of these factors. Other sources of variation in our sample (as well as in many of the earlier studies) were the languages spoken by the bilinguals and how typologically similar a bilingual's two languages are. It is not unlikely that the similarity between participants' languages and the language that the artificial language is based on determine the ease with which the artificial language is learned. Also, and more speculatively, it is possible that the typological similarity between a bilingual's two languages mediates learning, such that bilinguals who speak two typologically very different languages develop better metalinguistic abilities (or improved "structural sensitivity," cf. Kuo and Anderson, 2012; Kuo et al., 2016) that help them extract linguistic structure in statistical learning tasks. However, previous non-adjacent dependency learning studies with mono- and bilingual children that did find effects of bilingualism also contained heterogeneous samples of bilinguals that varied not only in the languages spoken, but also in language use, language proficiency, and age of onset (de Bree et al., 2017; Verhagen and de Bree, 2021), making it unclear to which extent this heterogeneity affected our findings. Future work should take these factors into account.

There is a real possibility that there is no robust across-the-board bilingual advantage in statistical learning, similar to other areas of research on statistical learning (Schmalz et al., 2017; West et al., 2021). Future work might investigate how presentation order of experiments influences the results. Another avenue is to explore whether a bilingual advantage surfaces solely or more prominently when multiple statistical patterns have to be tracked rather than one pattern. Bilingual learners encounter different languages and may therefore encounter more different patterns than monolinguals (depending on their language experiences). While there has been research on bilinguals' tracking of multiple statistical regularities

(for an overview, see Weiss et al., 2020), to the best of our knowledge, studies have not yet compared single and multiple pattern tracking within the same participants, while taking prior language experience into account (see also Weiss et al., 2020).

### Statistical learning and phonological memory and cognitive control

Bilinguals did not outperform monolinguals on tasks of phonological memory and cognitive control. In fact, for the adults, we found that the bilinguals performed *less* well on the cognitive control task than the monolinguals. Furthermore, there was no clear association between phonological memory and cognitive control abilities and statistical learning from variable input, once these factors were entered in the regressions. For the adults, cognitive control was positively related to NADL irrespective of whether the input was variable. For the children, no effects of phonological memory or cognitive control emerged.

Overall, our findings are in line with those in earlier work, showing no strong evidence for effects of bilingualism on phonological memory and cognitive control (Paap et al., 2015; van den Noort et al., 2019). Possible explanations relate to the tasks and participants at stake: some studies found that the advantage is mainly seen in complex tasks, and does not show in young adults, the current age group, who are at their peak of EF development (Bialystok et al., 2004). Furthermore, the type of bilingual speakers might have played a role. Earlier work has suggested that the advantage is related to bilinguals who use their languages in specific dual-language contexts, for example, with interlocutors who do not speak both languages, such that switching between languages is required to maintain mutual understanding (Green and Abutalebi, 2013).

Thus, while there are some indications from the current study as well as earlier work (Bartolotti et al., 2011) that bilingualism positively affects statistical learning and that this might be due to enhanced phonological memory and cognitive control, the current results as well as earlier mixed findings (Bartolotti et al., 2011; Verhagen and de Bree, 2021) make it unlikely that enhanced phonological memory and cognitive control impact strongly on statistical learning. Both in our study and in earlier studies, correlations between cognitive abilities and statistical learning were found for only some of the tasks or experiments. This may suggest that correlations are not robust and modulated by specifics of the tasks used, such as the stimuli used, task modality, and presumably also the order in which tasks are administered.

## Conclusion

The present results suggest that there might be some advantage of bilinguals in statistical learning, but this advantage is not robust. It largely surfaced only in *t*-tests against chance for the groups separately, did not surface in the same way for children (where it was found for consistent input) and adults (where it was found for variable input), and was modulated by experiment order. As such, the current results add to the mixed findings in earlier work that indicate that there is no broad, overall effect of bilingualism in statistical learning. They raise the suggestion that future assessment of statistical learning should also take variation within bilingual samples into account. Furthermore, our results provide no evidence that any enhancement in bilinguals' statistical learning was related to improved phonological memory and cognitive control: bilinguals did not outperform monolinguals on these cognitive measures and performance on these measures did not consistently relate to statistical learning outcomes. Taken together, these findings suggest that any potential effects of bilingualism on statistical learning probably do not involve enhanced cognitive abilities associated with bilingualism (Kovács and Mehler, 2009; Kuo and Kim, 2014; Hirosh and Degani, 2017). Future work might explore further to what extent effects, if found, are due to individual differences in bilingual language use, proficiency, and exposure.

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://osf.io/b4ps6/?view_only=a18f5b5cb1d04905b6c26f29de2f43b1.

## Ethics statement

The research was conducted in accordance with The Netherlands Code of Conduct for Scientific Practice, as well as with the guidelines of the Ethics Committees of the University of Amsterdam and Utrecht University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

## Author contributions

JV: conceptualized the study, set up experiments, analyzed the data, wrote study, and revised the draft. EB: conceptualized the study, set up experiments, wrote parts of study, and revised draft. All authors contributed to the article and approved the submitted version.

## Funding

This research was partly financed with a VENI grant (Number 275-89-01) from the Netherlands Organization for Scientific Research awarded to the first author.

## Conflict of interest

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

## Publisher's note

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

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1127718/full#supplementary-material

## References


Antovich, D. M., and Graf Estes, K. (2017). Learning across languages: bilingual experience supports dual language statistical word segmentation. *Develop. Sci.* 21, 125–148. doi: 10.1111/desc.12548

Bartolotti, J., Marian, V., Schroeder, S. R., and Shook, A. (2011). Bilingualism and inhibitory control influence statistical learning of novel word forms. *Front. Psychol.* 2, 1–10. doi: 10.3389/fpsyg.2011.00324

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effect models using lme4. *J. Statistic. Softw.* 67, 1-48. doi: 10.18637/jss.v067.i01

Benitez, V. L., Yurovsky, D., and Smith, L. B. (2016). Competition between multiple words for a referent in cross-situational word learning. *J. Memory Lang.* 90, 31–48. doi: 10.1016/j.jml.2016.03.004

Bialystok, E., Craik, F. I. M., Klein, R., and Viswanathan, M. (2004). Bilingualism, aging and cognitive control: evidence from the Simon task. *Psychol. Aging* 19, 290–303. doi: 10.1037/0882-7974.19.2.290

Bogulski, C. (2013). *Are bilinguals better learners? A neurocognitive investigation of the bilingual advantage. Unpublished doctoral dissertation*.

Bonifacci, P., Giombini, L., Bellocchi, S., and Contento, S. (2011). Speed of processing, anticipation, inhibition and working memory in bilinguals. *Develop. Sci.* 14, 256–267. doi: 10.1111/j.1467-7687.2010.00974.x

Bulgarelli, F., Lebkuecher, A. L., and Weiss, D. J. (2018). Statistical learning in bilingualism. *Lang. Speech Hear. Serv. Schools* 49, 740–753. doi: 10.1044/2018_LSHSS-STLT1-17-0139

Bulgarelli, F., and Weiss, D. J. (2016). Anchors aweigh: the impact of overlearning on entrenchment effects in statistical learning. *J. Experiment. Psychol. Learn. Memory Cogn.* 42, 1621–1631. doi: 10.1037/xlm0000263

Byers-Heinlein, K., and Fennell, C. T. (2014). Perceptual narrowing in the context of increased s variation: insights from bilingual infants. *Develop. Psychobiol.* 56, 274–291. doi: 10.1002/dev.21167

de Bree, E. H., Verhagen, J., Kerkhoff, A. O., Doedens, W. J., and Unsworth, S. (2017). Language learning from inconsistent input: bilingual and monolingual toddlers compared. *Infant Child Develop.* 26, e1996. doi: 10.1002/icd.1996

De Jong, P. F. (1998). Working memory deficits of reading disabled children. *J. Experiment. Psychol.*70, 75–96. doi: 10.1006/jecp.1998.2451

Dunn, L., and Dunn,. L. M. (2005). *Peabody Picture Vocabulary Test-III-NL. Nederlandse versie door Liesbeth Schlichting [Dutch version by Liesbeth Schlichting]. Amsterdam: Harcourt Assessment B.V.*

Engel de Abreu, P. M., Cruz-Santos, A., Tourinho, C. J., Martin, R., and Bialystok, E. (2012). Bilingualism enriches the poor: Enhanced cognitive control in low-income minority children. *Psychologic. Sci.* 23, 1364–1371. doi: 10.1177/0956797612443836

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychologic. Sci.* 13, 431–436. doi: 10.1111/1467-9280.00476

Gómez, R. L., and Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy* 7, 183–206. doi: 10.1207/s15327078in0702_4

Grama, I. C., Kerkhoff, A. M., and Wijnen, F. N. K. (2016). Gleaning structure from sound: the role of prosodic contrast in learning non-adjacent dependencies. *J. Psycholinguistic Res.* 45, 1427–1449. doi: 10.1007/s10936-016-9412-8

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

Hakvoort, B. (2009). *Non-adjacent dependency learning across domains in typically developing children and poor readers. Unpublished MA thesis. Utrecht University*.

Hirosh, Z., and Degani, T. (2017). Direct and indirect effects of multilingualism on novel language learning: an integrative review. *Psychonomic Bull. Rev.* 25, 892–916. doi: 10.3758/s13423-017-1315-7

Hsu, H. J., Tomblin, J. B., and Christiansen, M. H. (2014). Impaired statistical learning of non-adjacent dependencies in adolescents with specific language impairment. *Front. Psychol.* 6, 5–175. doi: 10.3389/fpsyg.2014.00175

Kapa, L. L., and Colombo, J. (2014). Executive function predicts artificial language learning. *J. Memory Lang.* 76, 237–252. doi: 10.1016/j.jml.2014.07.004

Karpicke, J. D., and Pisoni, D. B. (2004). Using immediate memory span to measure implicit learning. *Memory Cogn.* 32, 956–966. doi: 10.3758/BF03196873

Kaufman, S., DeYoung, C., Gray, J., Jiménez, L., Brown, J., and Mackintosh, N. (2010). Implicit learning as an ability. *Cognition* 116, 321–340. doi: 10.1016/j.cognition.2010.05.011

Kovács, A., and Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *Science* 325, 611–612. doi: 10.1126/science.1173947

Kuo, L. J., and Anderson, R. C. (2012). Effects of early bilingualism on learning phonological regularities in a new language. *J. Experiment. Child Psychol.* 111, 455–467. doi: 10.1016/j.jecp.2011.08.013

Kuo, L. J., and Kim, T. J. (2014). "Effect of early bilingualism on metalinguistic development and language processing: evidence from Chinese-speaking bilingual children," in *Reading Development and Difficulties in Monolingual and Bilingual Chinese Children*, eds X. Chen- Bumgardner, Q. Wang, and Y. Luo (New York, N.Y: Springer). p. 171–190.

Kuo, L. J., Uchikoshi, Y., Kim, T. J., and Yang, X. (2016). Bilingualism and phonological awareness: re-examining theories of cross-language transfer and structural sensitivity. *Contemp. Educ. Psychol.* 46, 1–9. doi: 10.1016/j.cedpsych.2016.03.002

Misyak, J. B., Christiansen, M. H., and Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Front. Psychol.* 1, 31. doi: 10.3389/fpsyg.2010.00031

Onnis, L., Chun, W. E., and Lou-Magnuson, M. (2018). Improved statistical learning abilities in adult bilinguals. *Biling. Lang. Cogn.* 21, 427–433. doi: 10.1017/S1366728917000529

Paap, K. R., Johnson, H. A., and Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex* 69, 265–278. doi: 10.1016/j.cortex.2015.04.014

Perruchet, P., and Tillmann, B. (2010). Exploiting multiple sources of information in learning an artificial language: human data and modeling. *Cogn. Sci.* 34, 255–285. doi: 10.1111/j.1551-6709.2009.01074.x

Poepsel, T. J., and Weiss, D. (2016). The influence of bilingualism on statistical word learning. *Cognition* 152, 9–19. doi: 10.1016/j.cognition.2016.03.001

Potter, C. E., Wang, T., and Saffran, J. R. (2017). Second language experience facilitates statistical learning of novel linguistic materials. *Cogn. Sci.* 41, 913–927. doi: 10.1111/cogs,0.12473

R Core Team (2017). *R: A language and environment for statistical computing*. R *Foundation for Statistical Computing, Vienna, Austria*. Available online at: https://www.R-project.org/

Reitan, R. M. (1956). *Trail Making Test. Manual for Administration, Scoring, and Interpretation*. Indianapolis, IN: Indiana University Press.

Rispens, J., and Baker, A. (2012). Nonword repetition: the relative contributions of phonological short-term memory and phonological representations in children with language and hearing impairment. *J. Speech Lang. Hearing Res.* 55, 683–694. doi: 10.1044/1092-4388(2011/10-0263)

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., and Barrueco, S. (1997). Incidental language learning: listening (and learning) out of the corner of your ear. *Psychologic. Sci.* 8, 101–105. doi: 10.1111/j.1467-9280.1997.tb00690.x

Sánchez-Cubillo, I., Periáñez, J. A., Adrover-Roig, D., Rodríguez-Sánchez, J. M., Ríos-Lago, M., Tirapu, J., et al. (2009). Construct validity of the trail making test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychologic. Soc.* 15, 438–450. doi: 10.1017/S1355617709090626

Schad, D. J., Vasishth, S., Hohenstein, S., and Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: a tutorial. *J. Memory Lang.* 110, 104038. doi: 10.1016/j.jml.2019.104038

Schmalz, X., Altoè, G., and Mulatti, C. (2017). Statistical learning and dyslexia: a systematic review. *Annal. Dyslexia* 67, 147–162. doi: 10.1007/s11881-016-0136-0

Siegelman, N., and Frost, R. (2015). Statistical learning as an individual ability: theoretical perspectives and empirical evidence. *J. Memory Lang.* 81, 105–120. doi: 10.1016/j.jml.2015.02.001

Thiessen, E. D. (2017). What's statistical about learning? insights from modelling statistical learning as a set of memory processes. *Philosophic. Transact. Royal Soc. B* 372, 1711. doi: 10.1098/rstb.2016.0056

Thiessen, E. D., and Erickson, L. C. (2013). Discovering words in fluent speech: the contribution of two kinds of statistical information. *Front. Psychol.* 3, 1–10. doi: 10.3389/fpsyg.2012.00590

Thiessen, E. D., and Pavlik, P. I. (2013). iMinerva: a mathematical model of distributional statistical learning. *Cogn. Sci.* 37, 310–343. doi: 10.1111/cogs.12011

van den Noort, M., Struys, E., Bosch, P., Jaswetz, L., Perriard, B., Yeo, S., et al. (2019). Does the bilingual advantage in cognitive control exist and if so, what are its modulating factors? a systematic review. *Behav. Sci.* 13, 27. doi: 10.3390/bs9030027

Verhagen, J., and de Bree, E. H. (2021). Effects of bilingualism on statistical learning in preschoolers. *Linguist. Approaches Biling.* 11, 611–639. doi: 10.1075/lab.18102.ver

Wang, T., and Saffran, J. R. (2014). Statistical learning of a tonal language: the influence of bilingualism and previous linguistic experience. *Front. Psychol.* 5, 953. doi: 10.3389/fpsyg.2014.00953

Weiss, D. J., Schwob, N., and Lebkuecher, A. L. (2020). Bilingualism and statistical learning: lessons from studies using artificial languages. *Bilingual. Lang. Cogn.* 23, 92–97. doi: 10.1017/S1366728919000579

West, G., Hulme, C., and Melby-Lervåg, M. (2021). Is a procedural learning deficit a causal risk factor for developmental language disorder or dyslexia? a meta-analytic review. *Develop. Psychol.* 57, 749–770. doi: 10.1037/dev0001172

Yim, D., and Rudoy, J. (2013). Implicit statistical learning and language skills in bilingual children. *J. Speech Lang. Hear. Res.* 56, 310–323. doi: 10.1044/1092-4388(2012/11-0243)

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 14 July 2023
DOI 10.3389/fpsyg.2023.1177094



OPEN ACCESS

EDITED BY
Jinfen Xu,
Huazhong University of Science and Technology, China

REVIEWED BY
Yaru Meng,
Xi'an Jiaotong University, China
Dayu Jiang,
Wuhan University, China
Seriaznita Mat Said,
University of Technology Malaysia, Malaysia
Xiaoyu Sophia Zhang,
academyEX - The Mind Lab, New Zealand

*CORRESPONDENCE
Ying Gao
✉ gaoy687@nenu.edu.cn

RECEIVED 01 March 2023
ACCEPTED 26 June 2023
PUBLISHED 14 July 2023

CITATION
He W and Gao Y (2023) Explicating peer feedback quality and its impact on feedback implementation in EFL writing.
*Front. Psychol.* 14:1177094.
doi: 10.3389/fpsyg.2023.1177094



# Explicating peer feedback quality and its impact on feedback implementation in EFL writing

Wenjing He[1] and Ying Gao[2]*

[1]School of Foreign Languages, Northeast Normal University, Changchun, Jilin, China, [2]Ministry of Education (MOE) Language Training Center and School of Foreign Languages, Northeast Normal University, Changchun, Jilin, China

**Introduction:** Although it is commonly acknowledged that peer feedback quality is crucial to the success of peer review, there is a lack of consensus on how it could be determined. More importantly, how feedback quality interacts with other factors like feedback features and focus, and ultimately influences peer feedback implementation remains insufficiently investigated.

**Methods:** The present study examined peer feedback quality and its impact on Chinese students' feedback implementation in two argumentative writing tasks. Peer feedback quality was measured according to a self-designed two-dimensional measurement scale: accuracy and revision potential.

**Results:** Quantitative analyses of 5,606 implementable idea units of feedback and 440 writing drafts by 110 students revealed that feedback accuracy was at a medium level and revision potential was at a low level, with accuracy demonstrating stronger predictive power on implementation; the predictive strengths of feedback accuracy and revision potential were strongest when feedback features and focus were considered; the overall peer feedback quality was low and medium-quality feedback was implemented most frequently; feedback quality significantly and most strongly predicted implementation in combination with feedback features and focus.

**Discussion:** The study highlights the importance of future instructions in training students to provide and implement high-quality feedback with good accuracy and high revision potential.

KEYWORDS

**peer feedback quality, accuracy, revision potential, implementation, EFL writing**

## Introduction

Despite the proliferation of studies on peer feedback over the past three decades (Tsui and Ng, 2000; Wu, 2019; Cui et al., 2021; Payant and Zuniga, 2022), doubts about the effectiveness of peer feedback remain constant "as students are not experts in a subject area, peer feedback is susceptible to variation" (Strijbos et al., 2010, p. 291). In particular, although the large class size in EFL contexts like China has necessitated the use of peer feedback as a complement to teacher feedback in writing courses (Hu and Zhang, 2014; Wu et al., 2022), this skepticism on feedback quality (Nilson, 2003; Gielen et al., 2010) has hindered the application of this instructional activity in such contexts.

The importance of peer feedback quality has been widely acknowledged (Hattie and Timperley, 2007; Walker, 2015; Rotsaert et al., 2018), but it has not been defined consistently in the literature (Rosalia and Llosa, 2009; Gielen et al., 2010; Hovardas et al., 2014). Generally, conceptualizations of feedback quality have shifted from a comment-centric perspective concentrating on the features, amount, and length of feedback (e.g., Sluijsmans et al., 2002; Prins

et al., 2006; Patchan et al., 2018; Rotsaert et al., 2018) to a text-centric perspective which takes alignment and accuracy of peer feedback to text problems as central to feedback quality (e.g., van Steendam et al., 2010; Hovardas et al., 2014; Gao et al., 2019). More recently, feedback quality has been defined functionally by integrating problem detection accuracy with the usefulness of suggested solutions based on whether a comment could improve essay quality measurably on rubrics (Wu and Schunn, 2020a).

As success in peer review mostly relies on the use of feedback in subsequent revision/feedback implementation (van der Pol et al., 2008; Dressler et al., 2019), this integrative definition well highlights the significance of peer feedback in promoting the writing improvement. However, to capture more dimensions of peer feedback quality, a more detailed measurement is needed. Previous measurement scales from a comment-centric perspective generally focused on feedback types, features, and whether the feedback met certain assessment criteria (Prins et al., 2006; Walker, 2015), but they did not empirically test the contribution of the identified feedback characteristics on performance (Gielen et al., 2010). Thus, it is necessary to develop a comprehensive measurement scale that speaks specifically to the potential effect of peer feedback. Additionally, although feedback quality has been reported to determine implementation (van der Pol et al., 2008; Wu and Schunn, 2020a), it is still difficult to assume that this relationship would be similar in L2 writing classrooms, considering that language and culture may provide expected challenges (Carson and Nelson, 1994; Ramanathan and Atkinson, 1999; Lundstrom and Baker, 2009).

Therefore, in this study with Chinese EFL learners writing English argumentative essays, we intended to explicate peer feedback quality in measurable ways and test its impact on feedback implementation since using feedback to revise is central to peer review. Unlike most measurements concentrating on the characteristics of peer feedback (e.g., Sluijsmans et al., 2002; Prins et al., 2006), the current study focused on feedback content which can critically influence its effectiveness (Anson and Anson, 2017). We also considered two factors crucial to implementation: feedback features and focus, given that the revision process based on peer feedback is complex and feedback implementation is influenced by many factors (Wu and Schunn, 2020a). Specifically, we investigated whether considering these two factors would change the existing effect of feedback quality on feedback implementation.

# Literature review

## Peer feedback quality and its measurement

Although previous studies have shed light on peer feedback quality in the L1 context, understandings of feedback quality have been rather inconsistent (Rosalia and Llosa, 2009; Gielen et al., 2010; Wu and Schunn, 2020a). Generally, research on peer feedback quality mainly falls into three perspectives: comment-centric, text-centric, and integrative functional. Peer feedback quality from a comment-centric perspective was defined by the number and length of comments which could determine the sufficiency of persuasion (Patchan et al., 2018; Zong et al., 2020) or by the inclusion of helpful feedback features like identifying the problem, suggesting a solution, or explaining the problem (Sluijsmans et al., 2002; Prins et al., 2006; Li et al., 2010; Denton, 2018). Following this line of definition, Patchan et al. (2018) examined feedback quality by the amount of peer feedback using three different indicators: the number of words across comments provided, the overall number of comments, and the number of long comments. Along the same line, evaluating feedback quality with a measurement scale is the most commonly used measurement (Sluijsmans et al., 2002; Prins et al., 2006; Gielen et al., 2010). Generally, the measurement scales of feedback quality examine whether students' feedback contains certain features helpful for students' writing improvement. Frequencies or percentages of coded feedback features are usually compared within and across dimensions (Huang, 2018). However, it is problematic to define and measure feedback quality from a comment-centric perspective because there is no guarantee that the comments would orient toward text problems which mostly need to be solved in revision (Wu and Schunn, 2020a).

Peer feedback quality from a text-centric perspective focuses on the accuracy of comments in terms of both correctness and alignment. In the research by van Steendam et al. (2010), participants were asked to point out the problems in a text with 10–20 flaws and suggest revisions. The quality of feedback was determined by considering whether the problems were addressed in the correct ways in terms of "the correctness, exhaustiveness, and explicitness of student comments" (van Steendam et al., 2010, p. 321). Along the same line, Gao et al. (2019) examined the alignment of written peer feedback with text problems by coding each substance and high-prose text problem, and they found that the alignment between feedback and text problems significantly determined revision improvement. However, to exhaustively identify text problems is hard and often impossible, and the effect of accurate feedback can range from correcting the writing mechanics to substantially improving the essay content.

Hovardas et al. (2014) defined feedback quality by measuring both feedback accuracy and feedback features. This hybrid method combining the comment-centric perspective and the text-centric perspective sheds light on conceptualizing feedback quality from more than one dimension. Adopting an integrative functional approach, Wu and Schunn (2020a) defined feedback quality as the accuracy of problems detected and the usefulness of suggested solutions. This definition significantly highlighted the mediating role of peer feedback in guiding students to reflect on the original text and improve the revised text. Feedback quality was rated and categorized into high, medium, and low levels based on the likelihood that a particular comment would lead to essay improvement in measurable or non-measurable ways on a 7-point Likert scale. Wu and Schunn (2023) further investigated the effects of assessor writing performance on feedback quality by examining feedback problem identification accuracy and helpfulness of feedback. Though defining feedback quality from an integrative perspective has the strength of making holistic judgments, labeling feedback quality into high, medium, and low categories fails to reveal the finer grain sizes of feedback quality. For example, what are the specific criteria for determining accuracy levels? And how are the specific comment aspects leading to a meaningful improvement weighted in the rating?

Informed by the text-centric perspective and the integrative functional approach, this study proposed to define and measure peer feedback quality on two dimensions: accuracy and revision potential of peer feedback. The two-dimensional peer feedback quality speaks directly to the core of what teachers and students concern most: is the feedback accurate and has the potential to lead to writing improvement?

Specifically, accuracy refers to both alignment of feedback to a text problem and its correctness in addressing the problem. Aligned and correct feedback is crucial for peers to improve their writing (Hovardas et al., 2014; Gao et al., 2019) because feedback that aligns with a text problem can be either correctly or wrongly given whereas if a seemingly correct/reasonable comment is not aligned with the text problem, it is useless for text improvement. Revision potential refers to the potential of peer feedback in leading to text improvement, which is explicated in detail by rating the extent to which peer feedback could lead to writing improvement at different levels because the revision potential of feedback may vary from improving a minor mechanic issue of writing to significantly improving the gist or the logic of the essay. Unlike Wu and Schunn's (2020a) study which did not examine low-level writing issues (such as spelling and punctuation) due to a lack of statistical power, the present study investigated the feedback quality of the content issues and high-level writing issues (i.e., theme, text organization, and clarity of writing) as well as the low-level writing issues (i.e., grammar and mechanics) with different weights. In particular, content/high-level feedback was rated with higher revision potential in the measurement scale as it deserved more weighting in facilitating writing improvement. In argumentative writing, solid argumentation and reasoning are more challenging to students because the critical analysis of the facts and evidence imposes a heavy cognitive load on them (Noroozi, 2018; Latifi et al., 2021). Similarly, it may also be challenging to conduct a fair and objective assessment of complex content or high-level writing issues and comments may be limited to the surface-level issues without explanations for developing critical thinking (Noroozi et al., 2016; Latifi et al., 2021). Although peer feedback is guided by the review rubric and related to the original text, the choice of feedback focus on simple or complicated issues is made by the students.

## Impact of peer feedback quality on implementation

Peer feedback implementation refers to students' incorporation of peer feedback in revising their written text (Dressler et al., 2019), which is the linchpin of peer review. However, there are still uncertainties over whether or why students implement peer feedback in revisions (van der Pol et al., 2008; Walker, 2015; Wu and Schunn, 2020b). Generally, student writers are more likely to use more elaborated feedback (Noroozi et al., 2016), feedback with concrete suggestions (van der Pol et al., 2008), feedback which aligns with the text problems (Gao et al., 2019) and feedback helpful to writing improvement (Wu and Schunn, 2020a).

The effectiveness of peer feedback in terms of successful implementation hinges at least partly on the quality of the feedback that students provide (van der Pol et al., 2008, p. 1805). Hovardas et al. (2014) reported that students selectively used accurate feedback because they validated the effectiveness of feedback by cross-checking peer feedback and teacher feedback. Gao et al. (2019) found that whether the feedback aligned with the actual text problem or not could pose an impact on students' revision improvement as the revision was found to be consistent with the feedback received. By judging whether the peer feedback had enough potential to generate a meaningful improvement in the text being reviewed, Wu and Schunn (2020a) found that students were more likely to implement feedback when both feedback quality and frequency increased.

Studies have shown that feedback quality is essential to students' use of feedback, but the size of the effect is not clear and the specific contributions of accuracy and revision potential remain unexplored. Practically, with the increasing use of peer feedback among Chinese student writers (e.g., Gao et al., 2019; Li and Zhang, 2021), it is crucial to comprehend how feedback quality influences Chinese students' feedback implementation in order to improve the suggestions offered to students on how to provide constructive feedback. Additionally, to comprehensively explicate feedback quality and its impact on implementation, we also investigated other variables that may contribute to the dynamics and variation of the impact of feedback quality on feedback implementation, namely, feedback features, focus, gender, and comment length.

## Peer feedback features

In addition to feedback quality, feedback implementation could be influenced by other factors such as students' perceptions (van der Pol et al., 2008; Kaufman and Schunn, 2011), feedback focus (Shi, 2021), and individual differences (Winstone et al., 2017). One of the most important factors influencing feedback implementation is feedback features which refer to the structural components of feedback, such as whether they explicitly describe a problem or give praise (Wu and Schunn, 2020b). A large number of categorization systems have been utilized to investigate feedback features (e.g., Nelson and Schunn, 2009; Gielen and de Wever, 2015; Elizondo-Garcia et al., 2019). Psychologically, feedback features can be both cognitive (i.e., summarization, suggestion, explanation) and affective in nature (i.e., praise, mitigating praise) (Nelson and Schunn, 2009).

The impact of feedback features has been reported to be rather complicated. Some implementable features targeting the text problems (i.e., identification of the problem, solutions to address the problem) can be helpful to peers as they can arouse thinking, reflections, critical thinking (Filius et al., 2018), and implementation. Identification of problem (Lu and Law, 2012), suggestion (Nelson and Schunn, 2009; Leijen, 2017), solution (Wu and Schunn, 2020b), and explanation (Gielen et al., 2010; Wu and Schunn, 2020a) have been reported to pose a positive effect on feedback implementation in some studies. By contrast, other studies have reported that there is a negative impact of solution (Patchan et al., 2016) and explanation (Tseng and Tsai, 2007; Nelson and Schunn, 2009) on feedback implementation. However, peer feedback quality has not been considered when determining which feedback features are crucial to feedback implementation (Wu and Schunn, 2020a), which might be one explanation for the inconsistent earlier findings. Possibly, the effect of peer feedback is determined by feedback quality in the first place as inaccurate feedback might not be used no matter how many useful features it contains. Conversely, it is also possible that containing more helpful features (i.e., explanation of the problem) would increase the possibility of implementation even if the feedback does not fully address the text problem. For example, praise in a critical comment may persuade a peer to act upon it even if it is inaccurate (Wu and Schunn, 2020a).

Although feedback features are not the central focus of the current study, feedback features must also be carefully controlled because how peer feedback is structured would influence students' judgment about its persuasiveness and usefulness. Therefore, this study attempts to

extend the current knowledge of feedback quality by considering feedback features when examining what contributes to feedback implementation.

## Peer feedback focus

Another important variable that especially relates to feedback implementation is peer feedback focus. It refers to the topic of the issue described in feedback such as grammar, thesis, and sufficiency of the examples (Patchan et al., 2016). Broadly, peer feedback can be divided into content focus and writing focus (Patchan et al., 2016; Gao et al., 2019). The content focus of feedback is concerned with meaning issues such as missing content, whereas the writing focus involves both high-level and low-level writing issues such as clarity and transitions of the ideas (Patchan et al., 2016; Gao et al., 2019). Content and high-level feedback focuses on aspects like argumentation, flow, and organization whereas low-level feedback covers aspects like mechanics, formatting, tense, and plurals (Allen and Mills, 2014).

Feedback focusing on meaning/content, or high-level and low-level writing issues varies both in cognitive load and feedback implementation rate of feedback, as well as in the effect to improve revision quality (e.g., Baker, 2016; Patchan et al., 2016). Patchan et al. (2016) reported that a writer tended to improve revision quality by implementing high-level feedback. Although implementing high-level feedback is more beneficial to learning cognitively, it usually requires more learner effort (Ene and Upton, 2014; Baker, 2016). Additionally, learners tend to implement more form focus or low-level feedback and less high-level feedback (e.g., Tsui and Ng, 2000; Allen and Mills, 2014; Gao et al., 2019). Gao et al. (2019) found that students repaired a larger number of less challenging problems while ignoring the more demanding content and high-level writing problems, indicating that complex content feedback or high-level feedback sometimes might be beyond learner means.

Feedback focus may have an impact on the relationship between feedback implementation and feedback quality. High-quality feedback that may lead to a meaningful text improvement might not be implemented if it requires major revision on the writing content or the overall writing organization and logic because the revision is cognitively demanding and requires more learner effort. Thus, when investigating the impact of feedback quality on students' implementation, feedback focus should be considered.

## Additional variables

In addition to feedback features and focus, other variables like gender, comment length, and first draft quality may also influence feedback implementation and therefore should be statistically controlled (Noroozi et al., 2020; Wu and Schunn, 2020b). Gender has been found relevant to peer review as students of different gender might respond to peer feedback differently (Prinsen et al., 2009; Noroozi et al., 2020; Wu and Schunn, 2020b; Wu and Schunn, 2021). Noroozi et al. (2020) found that gender could influence feedback quality, essay quality, and students' learning of writing content. Prinsen et al. (2009) found that males disagreed with their learning partners more frequently than females and males expanded on their messages less than women.

Comment length might influence student writers' perceived feedback quality and thus influence feedback implementation (Patchan and Schunn, 2015; Patchan et al., 2018). Students are more likely to reflect on the long and detailed feedback received and perceive a stronger need to make any revisions (Zong et al., 2020).

First draft quality may influence the feedback amount, type, and the likelihood of implementation (Hovardas et al., 2014; Patchan et al., 2016; Wu and Schunn, 2023). For instance, the author may receive less implementable feedback simply because the draft is of good quality and has fewer text problems. Thus, when examining the effect of feedback quality, it is essential to control the first draft quality.

Although much is now known about the influencing factors of feedback implementation, less is known about the role of feedback quality. More importantly, there is not enough work that combines the two dimensions of accuracy and revision potential in explicating feedback quality and its effect on implementation. Further, even less is known about whether, and if so how, the effect of feedback quality changes when other interacting factors are considered.

# Research questions

The current study examined the impact of peer feedback quality on feedback implementation by taking both feedback features and focus into consideration. Specifically, the following two research questions were addressed:

1. What is the relative contribution of feedback accuracy and revision potential to feedback implementation with the consideration of feedback features and focus?
2. What is the relationship between the two-dimensional feedback quality and implementation with the consideration of feedback features and focus?

# Methods

## Participants and settings

This study was conducted in a compulsory course called "Comprehensive English" at a research-intensive university in Northeast China. The course was offered at Fall semesters to first-year graduate students majoring in computer science and communication once a week for three class periods. The course aimed to cultivate students' comprehensive language skills, with a particular focus on reading and writing. An asynchronous online peer review platform (*Peerceptiv*) was used in organizing writing peer review activities. *Peerceptiv*[1] is a research-validated and data-driven peer learning tool to assist students in demonstrating disciplinary knowledge through writing feedback practices (Li, 2023). It was developed over a decade of peer learning research at the Learning Research & Development

1 https://peerceptiv.com

Center at University of Pittsburgh. It is used to implement peer learning in North America and around the world in the sciences, English language arts, business and almost every other subject matter. To guarantee objective review and active engagement of the students, the drafts were randomly and anonymously distributed among peers in a double-blinded manner.

The 116 students were a convenience sample of enrollees in the course in two intact classes. Six students were excluded because they failed to submit drafts or review peers' essays, leaving 110 in the study (60 in Class A and 50 in Class B). Students' age ranged from 21 to 29 ($M = 23.65$). All the students passed the national English graduate record examination (NEGRE, with a possible total of 100 points) ($M = 65.13$, $SD = 6.40$). In general, the L2 proficiency of the students was approximately between 72 and 100 on the Test of English as a Foreign Language (TOEFL), which corresponds to the intermediate level. Results from the independent samples *t*-test revealed that students in the two classes had no significant difference in English proficiency based on their test scores in NEGRE (Class A: $M = 65.74$, $SD = 6.34$; Class B: $M = 64.50$, $SD = 7.84$) ($t = -1.84$, $df = 73.09$, $p > 0.05$). All students were taught by the same teacher and they all agreed that their data could be used for research.

## Procedures

### Training procedures

Peer review training is important for students to define clear objectives and remove misconceptions about the reviewing rubric. Consequently, peer review training activities were carried out to assist students to understand the processes of peer review, get familiarized with *Peerceptiv*, and motivate students to engage in peer review.

Students were trained as a group in class. Training procedures consisted of four steps: watching a short video introduction to *Peerceptiv*; teacher modeling through analyzing sample essays and components of high-quality feedback; teacher lectures on the benefits and ways of being a good reviewer and teacher-guided discussion on implementing feedback to improve writing. Additionally, consistent help was provided after class to help students with difficulties in the reviewing process. Supplementary Appendix B summarizes the training steps.

### Writing and reviewing procedures

Participants completed three main tasks. They submitted the first draft to the *Peerceptiv* platform, then provided feedback for three peers' essays, and finally revised their own draft based on peer feedback. Writing and reviewing activities on two writing tasks were conducted in this study. The two writing topics were: (1) "Some working parents believe childcare centers can provide the best care for their children, others believe that family members like grandparents can do a better job. Which do you prefer?" (Week 3); (2) "Do you agree or disagree with the following statement? One should never judge a person by external appearance" (Week 8). For each topic, students were asked to write a five-paragraph argumentative essay in 250–300 words in English. The essay was expected to include an introduction of the topic, solid evidence and examples, possible counterarguments and rebuttals, and a concise conclusion.

Writing and reviewing activity for each writing task lasted for 4 weeks. After writing and submitting draft one to *Peerceptiv* in the first week, students were given 2 weeks to read and review three peers' texts in English based on a four-dimension reviewing rubric which includes the thesis statement, organization, argument, and grammar and vocabulary (Supplementary Appendix A). The reviewing rubric was developed and adapted by following the previous reviewing rubric in Gao et al. (2019), Wu and Schunn (2020a), and Li and Zhang (2021). A minimum of three comments was required in each dimension. In the fourth week, students revised their own drafts before submitting the revised draft to the platform. Consequently, each student completed 4 writing drafts (2 for each topic) and 2 rounds of peer review (1 for each topic) in an 18-week semester.

## Measures

### Feedback coding

To precisely examine feedback quality, implementation, and other variables, all feedback comments were first segmented into idea units because a reviewer may provide several revision ideas in a single dimension (Wu and Schunn, 2021). An independent idea unit was defined as raising and/or solving one problem on one dimension (Wu and Schunn, 2020b). The comments were segmented by two research assistants who discussed with the authors the precision of segmentation constantly and solved all the disagreements. In total, the comments were divided into 8,107 idea units, among which 5,606 were implementable feedback. Implementable feedback could lead to revisions while non-implementable feedback could not (i.e., feedback including only praise). Since this study focused on feedback implementation, only implementable feedback comments were further analyzed and therefore praise and summary were excluded. The same two research assistants double-coded all the implementable feedback by following the rating and coding schemes (Tables 1, 2), and disagreement was resolved through discussions together with the authors. *Kappa* values for each of the coding categories ranged from 0.70 to 0.90, indicating high inter-rater reliability.

### Feedback quality

Based on our proposed definition, a two-dimensional measurement scale was developed (Table 1). Each idea unit was checked to see whether it aligned with the text problem, whether it correctly addressed the problem, and whether it had the potential to lead to text improvement.

To quantify feedback quality, both the accuracy and the revision potential of feedback were rated on a 0–3 scale, each with a description and an example in Table 1. The best feedback which accurately addressed a problem and could lead to significant improvement of writing through solving a holistic content or high-level writing issue would get 6 points in rating while the worst feedback would get 0 points.

For instance, the feedback "*The writer did not make full explanation of the examples in the second paragraph because he failed to give the reasons why childcare centers make kids more independent than peers. He could write that childcare centers could train kids to get dressed by themselves.*" was rated as 5-point quality feedback as in accuracy it got 3 points for accurately addressing the problem and 2 points in revision potential for leading to writing improvement by solving a singular content problem. In another idea unit "*In the second sentence, 'There is a discussion about whether children should be sent to childcare center or be looked by their grandparents at home'. 'looked by' should be 'looked at by'.*" was only assigned 1 points. The idea unit

TABLE 1 Measurement scale of peer feedback quality.

| Dimension | Score | Description | Example |
|---|---|---|---|
| Accuracy ($K = 0.75$) | 0 | Feedback that is not aligned with the text problem | "In this day and age, childcare centers are becoming more and more professional," "professional" should be in noun form. |
| | 1 | Feedback that is aligned with the text problem but incorrectly addresses it | In the second sentence, "There is a discussion about whether children should be sent to childcare center or be looked by their grandparents at home." "looked by" should be "looked at by." |
| | 2 | Feedback that is aligned with the text problem but only correctly addresses part of it | There aren't topic sentences in the three body paragraphs. I think you should add "The childcare center can enhance children's communication ability" in the beginning of the second paragraph. But I do not know how to revise your third and fourth paragraph. |
| | 3 | Feedback that is aligned with the text problem and correctly addresses it | You may add some counter-arguments and rebuttals to support your position, which means, instead of talking about the benefits childcare centers have, you can list some defects when grandparents take care of children. |
| Revision potential ($K = 0.70$) | 0 | Feedback that has no potential of leading to any writing improvement or has the potential of leading to negative changes | The word "traveled" should be changed to "travelled." |
| | 1 | Feedback that has the potential of leading to minor writing improvement through solving a singular low-level writing problem | I think in the first paragraph, the word "today" should be capitalized. |
| | 2 | Feedback that has the potential of leading to writing improvement through solving a common low-level problem or a singular content/high-level writing problem | The writer did not make full explanation of the examples in the second paragraph because he failed to give the reasons why childcare centers make kids more independent than peers. He could write that childcare centers could train kids to get dressed by themselves. |
| | 3 | Feedback that has the potential of leading to significant improvement of writing through solving a holistic content or high-level writing issue | This article lacks two paragraphs. The second paragraph should be divided into three paragraphs. You can talk about the professionalism of childcare centers in para 2, how childcare centers help children develop their abilities in para 3 and add some counter-arguments in para 4. |

aligned with a grammatical error in the essay (word collocation of "*look*"), but it incorrectly addressed the text problem (The correct form should be "*looked after by*"). Therefore, the idea unit only got 1 points in accuracy and 0 points in revision potential since the sentence was still wrong if the feedback was implemented.

### Feedback features

Feedback was coded for the presence/absence of five feedback features, namely, identification, suggestion, solution, explanation, and mitigating praise (Table 2 for definitions and examples). We coded "1" for the presence and "0" for the absence.

### Feedback focus

Feedback was coded as meaning-level (content and high-level) feedback if it focused on the thesis, argument, evidence for claims, conclusion, and organization. Feedback on word choice, grammar, cohesion, sentence variety, and conventions was labeled surface-level (low-level) feedback. Since each feedback either focused on meaning-level or surface-level issues, it was binary-coded, "1" for meaning-level and "0" for surface-level.

### Feedback implementation

Feedback implementation was coded for whether the feedback was implemented in the revised drafts. The changes between the first and the revised draft were located using MS Word's Compare Document function. If a text change was made in response to the feedback, the feedback was coded as implemented. The feedback was labeled "Not Implemented" if it did not seem to lead to any revisions.

### Text quality

Students' first draft writings were rated and calculated by the mean value of ratings from the same two assistants who coded the feedback. Following ESL Composition Profile (Jacobs et al., 1981), the scoring rubric covered content, organization, vocabulary, language use, and mechanics which were in good alignment with the review prompt questions provided to students. The *Kappa* values of the two raters for the first drafts in the two tasks were 0.77 and 0.82, respectively.

### Comment length

Comment length refers to the number of words in each piece of feedback (Patchan and Schunn, 2015), calculated by the function of MS Excel automatically. The average feedback length was 18.16 words.

## Data collection and analysis

The writing drafts and peer feedback were downloaded from *Peerceptiv*. In total, we examined 440 writing drafts from 110 students in two tasks and 5,606 implementable feedback. Variables and their descriptions were summarized in Table 3.

TABLE 2 Coding scheme of feedback features, focus, and implementation.

| | Definition | Examples |
|---|---|---|
| **Feedback features** | | |
| Identification ($K = 0.90$) | Feedback identifying a text problem | The first paragraph is too long. |
| Suggestion ($K = 0.83$) | Feedback giving general advice for revision | You should pay attention to the punctuation. |
| Solution ($K = 0.79$) | Feedback providing a specific solution for revision | In the second para, "," should be changed to "." |
| Explanation ($K = 0.82$) | Feedback containing an explanation of an issue | The word "external" in the first paragraph can be removed because the word appearance itself has the meaning of external. |
| Mitigating praise ($K = 0.71$) | Feedback on a text problem containing a praise | It is great to associate this topic with the mental health of teenagers and value formation. But the argument process still needs to be strengthened. |
| **Feedback focus** | | |
| Meaning-level/surface-level ($K = 0.89$) | Feedback on thesis, evidence, argument, organization, or conclusion/Feedback on convention, grammar, sentence variety, word choice, cohesion, and reference | M: The conclusion is a little short.<br>S: In the first para, "matters" should be changed to "matter." |
| **Implementation ($K$=0.77)** | | |
| Implemented | Feedback that is incorporated in the revision | Add a title.<br>(The author added a title in draft 2). |
| Not implemented | Feedback that is not incorporated in the revision | There is no thesis statement.<br>(The author did not add the thesis statement in draft 2). |

To address the first research question, a basic description of data such as peer feedback quality, features, focus, and implementation was presented (Table 4) and SPSS 26.0 was used to conduct statistical analysis. Since the feedback data (i.e., features, quality) was nested within authors, two-level hierarchical modeling was conducted with Stata 15. Logistic regression was used because the dependent variable (peer feedback implementation) was a binary outcome variable. The first set of regression was conducted to analyze how the two dimensions of feedback quality predicted feedback implementation. To answer our second research question, the second group of logistic regression was conducted to explore how the overall peer feedback quality predicted feedback implementation.

Since logistic regression was used, the results of the models were presented as odds ratios (*OR*). An odds ratio (*OR*) is a measure of the association between an exposure and an outcome. The exponential function of the regression coefficient is the odds ratio associated with a one-unit increase in exposure. Feedback features and focus were also considered in both sets of regressions to test the interactive strength of prediction on feedback implementation.

# Results

In this section, we first reported the levels of accuracy and revision potential of feedback, as well as the descriptive data of feedback features, focus, implementation and other control variables. We then reported the correlations among different variables and finally reported the relative contribution of feedback accuracy and revision potential to feedback implementation. We reported the findings of the second research question by following similar procedures.

## Relative contribution of feedback accuracy and revision potential to feedback implementation

According to the two-dimensional measurement scale, feedback quality was measured on both accuracy and revision potential of feedback toward text problems. It was found that average feedback accuracy ($M = 2.07$, $SD = 1.24$) was at a medium level (approaching 70% of the total rating). Specifically, 62.5% of the feedback ($N = 3{,}505$) aligned with and accurately addressed the text problems (rated as 3), and 20.0% ($N = 1{,}041$) of the feedback was not aligned with text problems (rated as 0). Revision potential of feedback ($M = 1.29$, $SD = 0.94$) was at a low level (getting about 41% of the total rating). Only 12.5% ($N = 705$, rated as 3) had the potential for significant improvement in writing, and 21.7% of feedback ($N = 1{,}221$, rated as 0) would not lead to text improvement. In particular, for feedback with 3 points in accuracy ($N = 3{,}505$), only 16% ($N = 577$) got 3 points and about 50% ($N = 1{,}910$) got 1 point in revision potential. This big inconsistency between accuracy and revision potential suggests that accurate feedback may not lead to big text improvement due to limited revision potential.

Table 4 presents a summary of the descriptive data averagely on each author. With feedback quality, we reported the average rating. With feedback features, focus, and feedback implementation, the average amount of feedback by the authors was reported.

Among the feedback features, identification ($M = 26.46$) and solution ($M = 19.80$) were the most common, while mitigating praise was the least frequent ($M = 0.53$). Moreover, students received significantly less meaning-level feedback than surface-level feedback according to paired samples $t$-test ($t = -4.47$, $df = 109$, $p < 0.01$). Of the 5,606 implementable feedback analyzed, 2,633 (47%) was implemented. Each author averagely incorporated 23.94 feedback.

TABLE 3 Types of coding and measures of variables in the study.

| Variable | Type | Description |
|---|---|---|
| **Dependent variable** | | |
| Implementation | Binary | Whether the feedback is used in revisions or not |
| **Independent variables** | | |
| **Feedback quality** | | |
| Accuracy | Continuous | Whether the feedback accurately addresses the text problem |
| Revision potential | Continuous | To what extent the feedback could lead to writing improvement |
| **Feedback features** | | |
| Identification | Binary | Whether the feedback identifies the text problem or not |
| Suggestion | Binary | Whether the feedback provides general advice for revision or not |
| Solution | Binary | Whether the feedback provides a specific solution for revision or not |
| Explanation | Binary | Whether the feedback contains an explanation or not |
| Mitigating praise | Binary | Whether the feedback on a text problem includes praise or not |
| **Feedback focus** | | |
| Meaning-level/ surface-level | Binary | Whether the feedback is about meaning-level issues or surface-level issues |
| **Control variables** | | |
| Gender | Binary | Whether the student is female or not |
| Comment length | Continuous | Number of words in an idea unit |
| Draft 1 quality | Continuous | Mean ratings across two writing experts |

TABLE 4 Means and standard deviations of feedback quality, features, focus, and implementation.

| Measure | M | SD | Min | Max |
|---|---|---|---|---|
| **Peer feedback quality** | **3.36** | **1.93** | **2.36** | **4.32** |
| Accuracy (0–3) | 2.07 | 1.24 | 1.36 | 2.78 |
| Revision potential (0–3) | 1.29 | 0.94 | 0.91 | 2.01 |
| **Peer feedback features** | | | | |
| Identification | 26.46 | 11.21 | 5 | 55 |
| Suggestion | 14.49 | 5.40 | 5 | 28 |
| Solution | 19.80 | 7.79 | 6 | 43 |
| Explanation | 1.56 | 1.76 | 0 | 11 |
| Mitigating praise | 0.53 | 0.85 | 0 | 4 |
| **Peer feedback focus** | | | | |
| Meaning-level | 23.91 | 8.20 | 6 | 41 |
| Surface-level | 27.05 | 9.00 | 8 | 49 |
| Peer feedback implementation | 23.94 | 9.29 | 4 | 46 |

Before running the regression tests, a correlation analysis was conducted among the variables (Table 5). Both accuracy and revision potential were significantly related to feedback implementation ($r_{accuracy} = 0.38^{**}$; $r_{revision\ potential} = 0.21^{**}$). Additionally, suggestion ($r = -0.08^{**}$), solution ($r = 0.11^{**}$), feedback focus ($r = -0.11^{**}$), and first draft quality ($r = -0.02^{*}$) significantly correlated with implementation.

To further explore the predictive strength of peer feedback accuracy and revision potential as well as the other variables, the first set of logistic regression test was run (Table 6). Model 1 included accuracy, revision potential, and control variables. In Model 2, feedback features were added on the basis of Model 1. In Model 3, feedback focus was added on the basis of Model 1. Model 4 was the full model examining the effects of all the variables on implementation, and it provided a better fit than the previous three models: $\chi^2$ (11) = 850.0, $p < 0.001$.

The effect of accuracy was significant and constant across models. Students tended to implement more feedback when it accurately addressed the text problem. The effect of accuracy increased when feedback features were added in Model 2 ($B = 0.77$, $SE = 0.04$, $p < 0.001$), and the effect was weakest when only feedback focus was included in Model 3 ($B = 0.64$, $SE = 0.03$, $p < 0.001$). In the full model, among all the factors predicting feedback implementation, the effect of accuracy was the largest among all the variables ($B = 0.78$, $SE = 0.04$, $p < 0.001$). The *OR* value of accuracy reached 2.18 in Model 4, suggesting that feedback with an extra point in accuracy was 2.18 times more likely to be implemented than feedback with a point less. Revision potential was not significant in Model 1 ($B = 0.02$, $SE = 0.04$, $p > 0.05$), but its effect became significant when feedback features and focus were included in Models 2–4. This indicated that revision potential did not predict feedback implementation together with the control variables, but when the effects of feedback features and focus were taken into consideration, revision potential became a significant predictor.

In the full model, among feedback features, identification positively contributed to feedback implementation ($B = 0.22$, $SE = 0.09$, $p < 0.05$). Suggestion was negatively significant ($B = -0.96$, $SE = 0.10$, $p < 0.001$). Surprisingly, the effects of solution, explanation, and mitigating praise were not significant in either of the two models that involved feedback features (Model 2 and Model 4).

Compared with surface-level peer feedback, meaning-level peer feedback significantly led to less implementation in this study (Model 3 and Model 4). Among the control variables, only comment length negatively predicted implementation ($B = -0.01$, $SE = 0.00$, $p < 0.01$). Gender and first draft quality were not significant predictors.

## Relationship between the two-dimensional feedback quality and implementation

In general, average peer feedback quality ($M = 3.36$, $SD = 1.93$) was unsatisfactorily at a low level (getting about 56% of the total rating), with a big variation between high and low quality feedback ($Max = 4.32$, $Min = 2.36$). Of a total of 6 points, 28.5% of feedback ($N = 1{,}600$) was at the assigned 6 or 5 points range, 43.2% ($N = 2{,}421$) got 4 or 3 points, and 28.3% ($N = 1{,}585$) got 2 points or less. This indicated that nearly 30% of feedback was very poor in quality which was either not aligned with/incorrectly addressed text problems or had low potential for writing improvement, or both.

Different from common expectations and previous research findings (Wu and Schunn, 2020a), implementation rates were found to be highest (over 60%) for middle-range quality feedback (4–3 points) and lowest (17.7%) for low-quality feedback (2–0 points) in

TABLE 5 Correlations among two dimensions of feedback quality (accuracy and revision potential), features, focus, and implementation.

| | Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Accuracy | | | | | | | | | | | |
| 2 | Revision potential | 0.56** | | | | | | | | | | |
| 3 | Identification | −0.15** | 0.16** | | | | | | | | | |
| 4 | Suggestion | 0.24** | 0.30** | −0.21** | | | | | | | | |
| 5 | Solution | 0.09** | −0.41** | −0.45** | −0.49** | | | | | | | |
| 6 | Explanation | 0.08** | 0.03** | −0.04** | −0.01 | −0.01 | | | | | | |
| 7 | Mitigating praise | 0.01 | 0.06** | −0.02 | 0.07** | −0.06** | 0.01 | | | | | |
| 8 | Meaning-level (reference: surface-level) | 0.02 | 0.41** | 0.29** | 0.32** | −0.56** | 0.04** | 0.07** | | | | |
| 9 | Gender | 0.00 | −0.03** | 0.02 | −0.01 | 0.01 | −0.03** | −0.02 | −0.02 | | | |
| 10 | Comment length | 0.11** | 0.02 | 0.08** | 0.13** | 0.07** | 0.22** | 0.09** | 0.15** | −0.03* | | |
| 11 | Draft 1 quality | −0.05** | −0.09** | −0.02 | 0.02 | 0.01 | −0.01 | −0.02 | −0.02* | 0.20** | −0.04** | |
| 12 | Implementation | 0.38** | 0.21** | −0.01 | −0.08** | 0.11** | 0.02 | −0.01 | −0.11** | 0.02 | −0.01 | −0.02* |

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

this study (Figure 1). Apparently, the students were able to screen out and discard most of the low-quality feedback in their text revision. However, they also ignored a large proportion (52.3%) of high-quality feedback (accurate feedback with high revision potential). Ideally, high-quality feedback deals with more complex issues of writing and therefore is more helpful to writing improvement if implemented.

In order to identify potential confounds and multicollinearity problems among the variables, Pearson correlation analysis was carried out (Table 7). Peer feedback quality was significantly related to feedback implementation ($r = 0.35$**).

To answer the second research question, we conducted the second set of logistic regressions (Table 8). Model 5 tested the relationship between feedback quality and implementation together with control variables. Model 6 and Model 7 tested the relationship when feedback features alone or focus alone was included. Model 8 included all the variables and provided a better fit to the data: $\chi^2(10) = 812.11$, $p < 0.001$.

In Model 5, feedback quality significantly predicted implementation ($B = 0.42$, $SE = 0.02$, $p < 0.001$). When feedback quality increased by one point, the feedback was 1.52 times ($OR = 1.52$) as likely to be implemented than feedback with one point less. Adding feedback features or focus to the models (Model 6 and 7) did not change the estimated relationships between feedback quality and implementation. When feedback quality, features, and focus were all included (Model 8), feedback quality remained to be a significant predictor with the largest effect ($B = 0.58$, $SE = 0.02$, $p < 0.001$). Specifically, when feedback quality increased by one point, it was 1.78 times ($OR = 1.78$) as likely to be implemented than feedback quality with one point less (Model 8).

In terms of feedback features, identification positively predicted feedback implementation in the full model ($B = 0.22$, $SE = 0.09$, $p < 0.05$). On the contrary, there was a negative relationship between suggestion and feedback implementation (Model 6 and Model 8). Solution, explanation and mitigating praise were not significantly related to implementation in the full model. Similar to that in the first set of regression (Model 3 and Model 4), surface-level peer feedback more significantly predicted implementation (Model 7 and Model 8) and among the control variables, only comment length negatively predicted implementation.

## Discussion

In line with Wu and Schunn's (2020a) study, the current study also deems that the collaborative peer review activities benefit learning in nature (Hovardas et al., 2014; Wu and Schunn, 2020a). The collaboration in peer review acts as a social process in which students work together to handle a writing task that no single hand could reach the intended achievement. The developmental changes experienced by individual ESL learners first occur between peers and then internally within the individual. To better understand and improve the effectiveness of this interactive peer feedback process, this study further explored the issue of peer feedback quality and its impact on feedback implementation.

As the purpose of peer review is to improve writing by involving students actively in providing and receiving feedback, we argue that peer feedback quality should be measured in terms of its degree of helpfulness for text improvement. Inspired by previous studies (van Steendam et al., 2010; Gao et al., 2019; Wu and Schunn, 2020a), the current study ventured further to explicate the nature of peer feedback quality by examining quantitatively what instructors and students care most in peer review: the accuracy (both feedback alignment with original text problems and correct addressing of text problem) and helpfulness (the potential of leading to meaningful revision) of peer feedback for writing improvement. Informed particularly by Wu and Schunn's (2020a) study, this conceptualization of feedback quality highlights the potential function of peer feedback in facilitating revision and writing improvement. The combination of feedback accuracy and revision potential may be closest to the sense of effectiveness that teachers and students value most as a measure of peer feedback's effectiveness. Different from using an overall judgment as in Wu and Schunn's (2020a) study, the designed measurement scale in the current study provides a more detailed measurement and specifies the process of evaluating feedback quality using a four-level rating scale (0–3) for each dimension of peer feedback quality. Practically, the measurement scale serves as a useful tool for teachers when assessing students' feedback quality.

TABLE 6 Logistic regression analysis of the effect of the two dimensions of feedback quality (accuracy and revision potential), features, and focus on implementation.

| Variable | Two dimensions of feedback quality (Model 1) | | | Two dimensions of feedback quality + features (Model 2) | | | Two dimensions of feedback quality + focus (Model 3) | | | Two dimensions of feedback quality + features + focus (Model 4) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *B* | *SE* | *OR* | *B* | *SE* | *OR* | *B* | *SE* | *OR* | *B* | *SE* | *OR* |
| Accuracy | 0.72 | 0.03 | 2.05*** | 0.77 | 0.04 | 2.17*** | 0.64 | 0.03 | 1.89*** | 0.78 | 0.04 | 2.18*** |
| Revision potential | 0.02 | 0.04 | 1.02 | 0.16 | 0.05 | 1.17*** | 0.26 | 0.05 | 1.13*** | 0.28 | 0.05 | 1.32*** |
| **Feedback features** | | | | | | | | | | | | |
| Identification | – | – | – | 0.19 | 0.09 | 1.20* | – | – | – | 0.22 | 0.09 | 1.24* |
| Suggestion | – | – | – | −0.99 | 0.10 | 0.37*** | – | – | – | −0.96 | 0.10 | 0.38*** |
| Solution | – | – | – | 0.10 | 0.11 | 1.10 | – | – | – | −0.13 | 0.12 | 0.88 |
| Explanation | – | – | – | −0.17 | 0.18 | 0.84 | – | – | – | −0.17 | 0.18 | 0.85 |
| Mitigating praise | – | – | – | 0.06 | 0.30 | 1.06 | – | – | – | 0.10 | 0.30 | 1.11 |
| **Feedback focus** | | | | | | | | | | | | |
| Meaning-level (reference: surface-level) | – | – | – | – | – | – | −0.75 | 0.07 | 0.47*** | −0.63 | 0.08 | 0.53*** |
| **Control variables** | | | | | | | | | | | | |
| Gender | 0.12 | 0.11 | 1.13 | 0.11 | 0.11 | 1.12 | 0.13 | 0.11 | 1.13 | 0.12 | 0.11 | 1.12 |
| Comment length | −0.01 | 0.00 | 0.99*** | −0.01 | 0.00 | 0.99*** | −0.01 | 0.00 | 0.99** | −0.01 | 0.00 | 0.99* |
| Draft 1 quality | 0.00 | 0.01 | 1.00 | 0.00 | 0.01 | 1.00 | 0.00 | 0.01 | 1.00 | 0.00 | 0.01 | 1.00 |
| **Model fit statistics** | | | | | | | | | | | | |
| AIC | 6789.20 | | | 6571.34 | | | 6681.57 | | | 6509.94 | | |
| BIC | 6835.62 | | | 6650.92 | | | 6734.62 | | | 6595.15 | | |

$N = 5{,}606$. "–" means that the variable was not included in the model.
$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

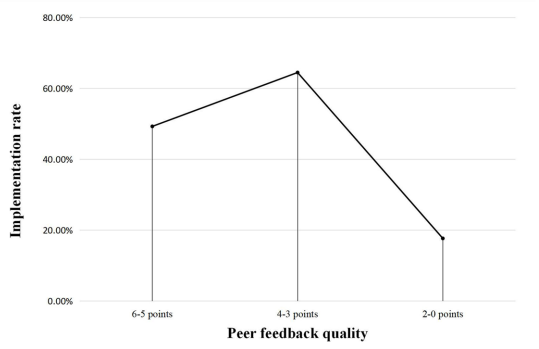



FIGURE 1
Mean feedback implementation rate of high, middle-range, and low-quality feedback ($N$=5,606).

The overall low peer feedback quality (medium in accuracy and low in revision potential) indicated that peer feedback was sometimes not satisfactory (Carson and Nelson, 1996; Tsui and Ng, 2000; Walvoord et al., 2007; Misiejuk and Wasson, 2021). Similarly, Hovardas et al. (2014) also found that the majority of peer feedback were scientifically accurate, but insufficient with suggestions and explanations for changes and improvement of writing skills. Therefore, although feedback accuracy was of medium level, the low level of revision potential suggested that students tended to receive feedback with the potential of leading to only minor writing improvement (Allen and Mills, 2014; Gao et al., 2019). Additionally, the inconsistency between accuracy and revision potential suggested that accuracy or the revision potential alone may not fully reflect the helpfulness of feedback on revision improvement. Accurate feedback with limited revision potential may have limited strength to improve revision quality and vice versa. Therefore, measuring feedback quality using either one of these two dimensions only reveals one side of the coin, which further suggests that the proposed two-dimensional measurement scale is a valid means of describing and reporting feedback quality, at least in the EFL context of the current research.

The predictive strength of peer feedback quality on implementation reveals two significant findings. First and foremost, when examining the predictive power of accuracy and revision potential, the largest *OR* values of accuracy (Model 1–4) suggested that feedback accuracy was the central predictor of feedback implementation and hence it should be of priority (Hovardas et al., 2014). The results were consistent with other research (Hovardas et al., 2014; Gao et al., 2019) in which students' revisions were influenced, either fully or partly, by peer feedback accuracy. Allen and Mills (2014) reported that, although the number of inaccurate feedback was minimal in number, and that only less than half of the erroneous comments were used in revision, the inaccurate feedback negatively affected writing quality. In addition, the large predictive power of feedback accuracy shows that students are highly sensitive to the alignment and the accuracy of the suggested

TABLE 7 Correlations among peer feedback quality, features, focus, and implementation.

| | Variable | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Feedback quality | | | | | | | | | | |
| 2 | Identification | −0.02 | | | | | | | | | |
| 3 | Suggestion | 0.30** | −0.21** | | | | | | | | |
| 4 | Solution | −0.14** | −0.45** | −0.49** | | | | | | | |
| 5 | Explanation | 0.07** | −0.04** | −0.01 | −0.01 | | | | | | |
| 6 | Mitigating praise | 0.03** | −0.02 | 0.07** | −0.06** | 0.01 | | | | | |
| 7 | Meaning-level (reference: surface-level) | 0.22** | 0.29** | 0.32** | −0.56** | 0.04** | 0.07** | | | | |
| 8 | Gender | −0.02 | 0.02 | −0.01 | 0.01 | −0.03** | −0.02 | −0.02 | | | |
| 9 | Comment length | 0.08** | 0.08** | 0.13** | 0.07** | 0.22** | 0.09** | 0.15** | −0.03* | | |
| 10 | Draft 1 quality | −0.07** | −0.02 | 0.02 | 0.01 | −0.01 | −0.02 | −0.02* | 0.20** | −0.04** | |
| 11 | Implementation | 0.35** | −0.01 | −0.08** | 0.11** | 0.01 | −0.01 | −0.11** | 0.02 | −0.01 | −0.02* |

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

TABLE 8 Logistic regression analysis of the effect of peer feedback quality, features, and focus on implementation.

| Variable | Feedback quality (Model 5) | | | Feedback quality + features (Model 6) | | | Feedback quality + focus (Model 7) | | | Feedback quality + features + focus (Model 8) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *B* | *SE* | *OR* | *B* | *SE* | *OR* | *B* | *SE* | *OR* | *B* | *SE* | *OR* |
| Feedback quality | 0.42 | 0.02 | 1.52*** | 0.54 | 0.02 | 1.71*** | 0.49 | 0.02 | 1.64*** | 0.58 | 0.02 | 1.78*** |
| **Feedback features** | | | | | | | | | | | | |
| Identification | – | – | – | 0.18 | 0.08 | 1.20* | – | – | – | 0.22 | 0.09 | 1.25* |
| Suggestion | – | – | – | −0.80 | 0.10 | 0.45*** | – | – | – | −0.82 | 0.10 | 0.44*** |
| Solution | – | – | – | 0.57 | 0.10 | 1.77*** | – | – | – | 0.17 | 0.11 | 1.18 |
| Explanation | – | – | – | −0.07 | 0.18 | 0.93 | – | – | – | −0.09 | 0.18 | 0.92 |
| Mitigating praise | – | – | – | −0.04 | 0.30 | 0.96 | – | – | – | 0.05 | 0.30 | 1.05 |
| **Feedback focus** | | | | | | | | | | | | |
| Meaning-level (reference: surface-level) | – | – | – | – | – | – | −0.95 | 0.06 | 0.39*** | −0.76 | 0.08 | 0.47*** |
| **Control variables** | | | | | | | | | | | | |
| Gender | 0.15 | 0.11 | 1.16 | 0.13 | 0.11 | 1.14 | 0.14 | 0.11 | 1.15 | 0.13 | 0.11 | 1.14 |
| Comment length | −0.01 | 0.00 | 0.99*** | −0.01 | 0.00 | 0.99*** | −0.01 | 0.00 | 0.99** | −0.01 | 0.00 | 0.99* |
| Draft 1 quality | 0.00 | 0.01 | 1.00 | 0.00 | 0.01 | 1.00 | 0.00 | 0.01 | 1.00 | 0.00 | 0.01 | 1.00 |
| **Model fit statistics** | | | | | | | | | | | | |
| AIC | 6935.63 | | | 6649.04 | | | 6712.02 | | | 6551.55 | | |
| BIC | 6975.42 | | | 6721.99 | | | 6758.44 | | | 6631.12 | | |

$N = 5{,}606$. "–" means that the variables were not included in the model.
*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

solutions, which should reassure teachers who are hesitant to use peer feedback (Wu and Schunn, 2020a).

Moreover, the inclusion of feedback features and focus did not change the existing relationship between feedback accuracy and feedback implementation, indicating that the effect of feedback accuracy was constant and robust. Revision potential significantly predicted implementation only when feedback features/focus were considered (Model 2–4) and its effect became largest when features and focus were both included (Model 4). Obviously, the inclusion of feedback features and focus changed the observed relationship between revision potential and feedback implementation. The significant relationships among the revision potential, feedback features, and feedback focus indicated that feedback features and focus were crucial statistic confounds that should be considered when exploring the influencing factors of feedback implementation (see Table 5). In Model 1, the omission of feedback features and focus inevitably increased the variance of the error term. After feedback features and focus were included in Models 2–4, the variance of the

error term became smaller, and it was probably why revision potential became a significant predictor of feedback implementation (see Table 6). Given the positive effect of accuracy and revision potential on feedback implementation, EFL students should be instructed on detecting the flaws central to text improvement and addressing the flaws in the correct and substantial ways.

Secondly, the largest *OR* values of feedback quality indicated that its effect on implementation was significant and constant across all models (Model 5–8). A crucial message for practice is that, in addition to validity and reliability, which have been the focus of many earlier studies (Falchikov and Goldfinch, 2000; Cho et al., 2006), the quality of feedback can affect its effectiveness (Gielen et al., 2010). The central role of feedback quality in students' likelihood of feedback implementation is similar to that in Wu and Schunn's (2020a) study and it also confirms the significance of feedback quality in peer review (Hattie and Timperley, 2007; Walker, 2015). The effect of feedback quality became largest when feedback features and focus were considered together (Model 8). The consideration of feedback features and focus has provided a better explanation of feedback implementation because the reduced AIC-adjusted deviance in the full model (Model 8) in comparison with the baseline model (Model 5) suggested that the full model had stronger explanatory strength to feedback implementation. Compared with previous studies focusing on one or two comment-level factors (e.g., Lu and Law, 2012; Patchan et al., 2016), this study ventures further to explore the effect of multiple factors and their specific contributions to feedback implementation.

Although feedback quality significantly predicted implementation, it was important to note that students tended to implement more middle-range quality feedback. High-quality feedback is undoubtedly more facilitating to revision improvement, but implementing high-quality feedback is more challenging and students might have limited knowledge about how to handle the information delivered through such feedback (Wichmann et al., 2018). Since students also tend to screen out low-quality feedback by employing some decision-making strategies (Gielen et al., 2010; Hovardas et al., 2014), they tend to implement only those middle-range quality feedback which is presumably within their zone of competence. This indicates that more guidance is needed to encourage students to take the tougher task of incorporating high-quality feedback in future instructions.

Although feedback features, focus and other control variables are not the central foci of the current study, we have discussed these variables because they are theoretically and empirically important (Nelson and Schunn, 2009; Allen and Mills, 2014; Patchan et al., 2016). The positive role of identification on implementation in this study was similar to previous studies (Lu and Law, 2012; Wu and Schunn, 2020a). In terms of cognitive load, identifying a text problem is relatively easier than giving a suggestion, a solution, or an explanation, which partially explains why the amount of feedback with identification was the largest in the data. Suggestion was a significant negative predictor of implementation in this study as feedback with suggestion was usually general and sometimes vague for students to comprehend and take action. A follow-up analysis revealed that general advice was not helpful for students to address the text problems. For example, feedback like "*You should change some examples.*" usually ended up being ignored in text revision. Different from the findings in previous studies (Gielen et al., 2010; Wu and Schunn, 2020a), explanation had no effect on feedback implementation in the present study. It was possible that the small amount of explanatory feedback could hardly generate statistical power on feedback implementation. This might also explain why mitigating praise was not a significant predictor. Solution significantly predicted implementation in Model 6, but when feedback focus was jointly considered (Model 8), it turned insignificant, suggesting that feedback focus could mediate the relationship between feedback features and implementation.

Since meaning-level feedback significantly led to less implementation than low-level feedback did, it was obvious that students trended toward taking less challenging tasks (Gao et al., 2019). Students implemented more low-level feedback as meaning-level issues were found to be more difficult for students to address (Ene and Upton, 2014; Patchan et al., 2016). The negative correlation between comment length and implementation again proved that students tended to avoid repairs mentioned in long comments which might involve more suggestions or explanation to solve harder text problems. The findings about feedback features and focus indicate that teachers' guidance should be directed toward emphasizing the significance of helpful features (e.g., identification of problems) and instructing students to implement more meaning-level feedback.

## Conclusion

This study reveals that peer feedback quality can be more comprehensively and scalably explicated from two dimensions: accuracy and revision potential of feedback. The complexity of the predictive strength of feedback quality on implementation well demonstrates the different and interactive power of peer feedback quality, features, focus, and other variables in peer review. Yet, among all these elements, feedback quality plays a central role in determining peer feedback implementation.

Pedagogically, this study implies that improving peer feedback quality should strategically orient toward both accuracy and revision potential of feedback as accurate feedback with strong revision potential is most likely to lead to revision improvement when implemented. At the same time, while peer feedback training should prioritize feedback quality, special care should be given to encouraging students to take the pain of dealing with complex issues in revision by implementing high-quality feedback, as well as feedback with significant features such as identification of the problem and content/high-level focus feedback. Therefore, to improve the effectiveness of peer review, more importance should be attached to promoting peer feedback literacy in both providing and implementing high-quality feedback.

Some limitations of the present study should be considered. Firstly, with the support of *Peerceptiv,* peer review of this study was conducted anonymously online with participants from one course. As such, the generalization of the results of this study to other contexts involving different participants from other disciplines with different writing tasks should be exercised with caution. Secondly, this study focused on the effects of feedback quality on implementation, leaving the effects of providing high or low-quality feedback on students' own draft revision unexplored. Future research can adopt this two-dimensional feedback quality measurement to further test the effect of feedback quality on the feedback providers' learning performance to obtain a more comprehensive understanding of the significance of peer feedback quality in determining the effectiveness of peer review. Lastly, although carefully designed, this study is correlational in nature. In promoting feedback quality, intervention studies are needed in the future, and results from the present study can help locate the intervention foci.

## Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

## Author contributions

WH: conceptualization, methodology, software, data coding and curation, writing—original draft preparation, and writing—reviewing and editing. YG: conceptualization, methodology, writing—reviewing and editing, supervision, and funding acquisition. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the Major Program of National Fund of Philosophy and Social Science of China (CN) [grant number 18BYY114].

## 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



## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmed.2023.1201977/full#supplementary-material

## References

Allen, D., and Mills, A. (2014). The impact of second language proficiency in dyadic peer feedback. *Lang. Teach. Res.* 20, 498–513. doi: 10.1177/1362168814561902

Anson, I. G., and Anson, C. M. (2017). Assessing peer and instructor response to writing: a corpus analysis from an expert survey. *Assess. Writ.* 33, 12–24. doi: 10.1016/j.asw.2017.03.001

Baker, K. M. (2016). Peer review as a strategy for improving students' writing process. *Act. Learn. High. Educ.* 17, 179–192. doi: 10.1177/1469787416654794

Carson, J. G., and Nelson, G. L. (1994). Writing groups: cross-cultural issues. *J. Second. Lang. Writ.* 3, 17–30. doi: 10.1016/1060-3743(94)90003-5

Carson, J. G., and Nelson, G. L. (1996). Chinese students' perceptions of ESL peer response group interaction. *J. Second. Lang. Writ.* 5, 1–19. doi: 10.1016/s1060-3743(96)90012-0

Cho, K., Schunn, C. D., and Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *J. Educ. Psychol.* 98, 891–901. doi: 10.1037/0022-0663.98.4.891

Cui, Y., Schunn, C. D., Gai, X., Jiang, Y., and Wang, Z. (2021). Effects of trained peer vs. teacher feedback on EFL students' writing performance, self-efficacy, and internalization of motivation. *Front. Psychol.* 12:788474. doi: 10.3389/fpsyg.2021.788474

Denton, A. W. (2018). Improving the quality of constructive peer feedback. *Coll. Teach.* 66, 22–23. doi: 10.1080/87567555.2017.1349075

Dressler, R., Chu, M. W., Crossman, K., and Hilman, B. (2019). Quantity and quality of uptake: examining surface and meaning-level feedback provided by peers and an instructor in a graduate research course. *Assess. Writ.* 39, 14–24. doi: 10.1016/j.asw.2018.11.001

Elizondo-Garcia, J., Schunn, C. D., and Gallardo, K. (2019). Quality of peer feedback in relation to instructional design: a comparative study in energy and sustainability MOOCs. *Int. J. Instr.* 12, 1025–1040. doi: 10.29333/iji.2019.12166a

Ene, E., and Upton, T. A. (2014). Learner uptake of teacher electronic feedback in ESL composition. *System* 46, 80–95. doi: 10.1016/j.system.2014.07.011

Falchikov, N., and Goldfinch, J. (2000). Student peer assessment in higher education: a meta-analysis comparing peer and teacher marks. *Rev. Educ. Res.* 70, 287–322. doi: 10.3102/00346543070003287

Filius, R. M., de Kleijn, R. A. M., Uijl, S. G., Prins, F. J., van Rijen, H. V. M., and Grobbee, D. E. (2018). Strengthening dialogic peer feedback aiming for deep learning in SPOCs. *Comput. Educ.* 125, 86–100. doi: 10.1016/j.compedu.2018.06.004

Gao, Y., Schunn, C. D., and Yu, Q. (2019). The alignment of written peer feedback with draft problems and its impact on revision in peer assessment. *Assess. Eval. High. Educ.* 44, 294–308. doi: 10.1080/02602938.2018.1499075

Gielen, M., and de Wever, B. (2015). Structuring the peer assessment process: a multilevel approach for the impact on product improvement and peer feedback quality. *J. Comput. Assist. Learn.* 31, 435–449. doi: 10.1111/jcal.12096

Gielen, S., Peeters, E., Dochy, F., Onghena, P., and Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learn. Instr.* 20, 304–315. doi: 10.1016/j.learninstruc.2009.08.007

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hovardas, T., Tsivitanidou, O. E., and Zacharia, Z. C. (2014). Peer versus expert feedback: an investigation of the quality of peer feedback among secondary school students. *Comput. Educ.* 71, 133–152. doi: 10.1016/j.compedu.2013.09.019

Hu, C., and Zhang, Y. (2014). A study of college English writing feedback system based on M-learning. *Mod. Educ. Technol.* 24, 71–78.

Huang, S. (2018). A GEARed peer feedback model and implications for learning facilitation. *Assess. Eval. High. Educ.* 43, 1194–1210. doi: 10.1080/02602938.2018.1439881

Jacobs, H. L., Zinkgraf, S. A., Wormouth, D. R., Hartfiel, V. F., and Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowely, MA: Newbury House.

Kaufman, J. H., and Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instr. Sci.* 39, 387–406. doi: 10.1007/s11251-010-9133-6

Latifi, S., Noroozi, O., and Talaee, E. (2021). Peer feedback or peer feedforward? Enhancing students' argumentative peer learning processes and outcomes. *Br. J. Educ. Technol.* 52, 768–784. doi: 10.1111/bjet.13054

Leijen, D. A. J. (2017). A novel approach to examine the impact of web-based peer review on the revisions of L2 writers. *Comput. Compos.* 43, 35–54. doi: 10.1016/j.compcom.2016.11.005

Li, A. W. (2023). Using Peerceptiv to support AI-based online writing assessment across the disciplines. *Assess. Writ.* 57:100746. doi: 10.1016/j.asw.2023.100746

Li, L., Liu, X., and Steckelberg, A. L. (2010). Assessor or assessee: how student learning improves by giving and receiving peer feedback. *Br. J. Educ. Technol.* 41, 525–536. doi: 10.1111/j.1467-8535.2009.00968.x

Li, W., and Zhang, F. (2021). Tracing the path toward self-regulated revision: an interplay of instructor feedback, peer feedback, and revision goals. *Front. Psychol.* 11:612088. doi: 10.3389/fpsyg.2020.612088

Lu, J. Y., and Law, N. (2012). Online peer assessment: effects of cognitive and affective feedback. *Instr. Sci.* 40, 257–275. doi: 10.1007/s11251-011-9177-2

Lundstrom, K., and Baker, W. (2009). To give is better than to receive: the benefits of peer review to the reviewer's own writing. *J. Second. Lang. Writ.* 18, 30–43. doi: 10.1016/j.jslw.2008.06.002

Misiejuk, K., and Wasson, B. (2021). Backward evaluation in peer assessment: a scoping review. *Comput. Educ.* 175:104319. doi: 10.1016/j.compedu.2021.104319

Nelson, M. M., and Schunn, C. D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instr. Sci.* 37, 375–401. doi: 10.1007/s11251-008-9053-x

Nilson, L. B. (2003). Improving student peer feedback. *Coll. Teach.* 51, 34–38. doi: 10.1080/87567550309596408

Noroozi, O. (2018). Considering students' epistemic beliefs to facilitate their argumentative discourse and attitudinal change with a digital dialogue game. *Innov. Educ. Teach. Int.* 55, 357–365. doi: 10.1080/14703297.2016.1208112

Noroozi, O., Biemans, H., and Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *Internet High. Educ.* 31, 20–31. doi: 10.1016/j.iheduc.2016.05.002

Noroozi, O., Hatami, J., Bayat, A., van Ginkel, S., Biemans, H. J. A., and Mulder, M. (2020). Students' online argumentative peer feedback, essay writing, and content learning: does gender matter? *Interact. Learn. Environ.* 28, 698–712. doi: 10.1080/10494820.2018.1543200

Patchan, M. M., and Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: how students respond to peers' texts of varying quality. *Instr. Sci.* 43, 591–614. doi: 10.1007/s11251-015-9353-x

Patchan, M. M., Schunn, C. D., and Clark, R. J. (2018). Accountability in peer assessment: examining the effects of reviewing grades on peer ratings and peer feedback. *Stud. High. Educ.* 43, 2263–2278. doi: 10.1080/03075079.2017.1320374

Patchan, M. M., Schunn, C. D., and Correnti, R. J. (2016). The nature of feedback: how peer feedback features affect students' implementation rate and quality of revisions. *J. Educ. Psychol.* 108, 1098–1120. doi: 10.1037/edu0000103

Payant, C., and Zuniga, M. (2022). Learners' flow experience during peer revision in a virtual writing course during the global pandemic. *System* 105:102715. doi: 10.1016/j.system.2021.102715

Prins, F., Sluijsmans, D., and Kirschner, P. A. (2006). Feedback for general practitioners in training: quality, styles, and preferences. *Adv. Health Sci. Educ.* 11, 289–303. doi: 10.1007/s10459-005-3250-z

Prinsen, F. R., Volman, M. L. L., Terwel, J., and van den Eeden, P. (2009). Effects on participation of an experimental CSCL-programme to support elaboration: do all students benefit? *Comput. Educ.* 52, 113–125. doi: 10.1016/j.compedu.2008.07.001

Ramanathan, V., and Atkinson, D. (1999). Individualism, academic writing, and ESL writers. *J. Second. Lang. Writ.* 8, 45–75. doi: 10.1016/s1060-3743(99)80112-x

Rosalia, C., and Llosa, L. (2009). "Assessing the quality of online peer feedback in L2 writing". in *Hand-book of research on E-learning methodologies for language acquisition*. eds. Cássia Veiga Marriott R. de and P. L. Torrest (Hershey, PA: Information Science Reference), 322–338.

Rotsaert, T., Panadero, E., Schellens, T., and Raes, A. (2018). "Now you know what you're doing right and wrong!" peer feedback quality in synchronous peer assessment in secondary education. *Eur. J. Psychol. Educ.* 33, 255–275. doi: 10.1007/s10212-017-0329-x

Shi, Y. (2021). Exploring learner engagement with multiple sources of feedback on L2 writing across genres. *Front. Psychol.* 12:758867. doi: 10.3389/fpsyg.2021.758867

Sluijsmans, D. M. A., Brand-Gruwel, S., and van Merriënboer, J. J. G. (2002). Peer assessment training in teacher education: effects on performance and perceptions. *Assess. Eval. High. Educ.* 27, 443–454. doi: 10.1080/0260293022000009311

Strijbos, J. W., Narciss, S., and Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: are they critical for feedback perceptions and efficiency? *Learn. Instr.* 20, 291–303. doi: 10.1016/j.learninstruc.2009.08.008

Tseng, S. C., and Tsai, C. C. (2007). Online peer assessment and the role of the peer feedback: a study of high school computer course. *Comput. Educ.* 49, 1161–1174. doi: 10.1016/j.compedu.2006.01.007

Tsui, A. B. M., and Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *J. Second. Lang. Writ.* 9, 147–170. doi: 10.1016/S1060-3743(00)00022-9

van der Pol, J., van den Berg, B. A. M., Admiraal, W. F., and Simons, P. R. J. (2008). The nature, reception, and use of online peer feedback in higher education. *Comput. Educ.* 51, 1804–1817. doi: 10.1016/j.compedu.2008.06.001

van Steendam, E., Rijlaarsdam, G., Sercu, L., and van den Berg, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learn. Instr.* 20, 316–327. doi: 10.1016/j.learninstruc.2009.08.009

Walker, M. (2015). The quality of written peer feedback on undergraduates' draft answers to an assignment, and the use made of the feedback. *Assess. Eval. High. Educ.* 40, 232–247. doi: 10.1080/02602938.2014.898737

Walvoord, M. E., Hoefnagels, M. H., Gaffin, D. D., Chumchal, M. M., and Long, D. A. (2007). An analysis of calibrated peer review (CPR) in a science lecture classroom. *J. Coll. Sci. Teach.* 37, 66–73. doi: 10.2307/42992780

Wichmann, A., Funk, A., and Rummel, N. (2018). Leveraging the potential of peer feedback in an academic writing activity through sense-making support. *Eur. J. Psychol. Educ.* 33, 165–184. doi: 10.1007/s10212-017-0348-7

Winstone, N. E., Nash, R. A., Parker, M., and Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: a systematic review and a taxonomy of recipience processes. *Educ. Psychol.* 52, 17–37. doi: 10.1080/00461520.2016.1207538

Wu, Z. (2019). Lower English proficiency means poorer feedback performance? A mixed-methods study. *Assess. Writ.* 41, 14–24. doi: 10.1016/j.asw.2019.05.001

Wu, W., Huang, J., Han, C., and Zhang, J. (2022). Evaluating peer feedback as a reliable and valid complementary aid to teacher feedback in EFL writing classrooms: a feedback giver perspective. *Stud. Educ. Eval.* 73:101140. doi: 10.1016/j.stueduc.2022.101140

Wu, Y., and Schunn, C. D. (2020a). When peers agree, do students listen? The central role of feedback quality and feedback frequency in determining uptake of feedback. *Contemp. Educ. Psychol.* 62:101897. doi: 10.1016/j.cedpsych.2020.101897

Wu, Y., and Schunn, C. D. (2020b). From feedback to revisions: effects of feedback features and perceptions. *Contemp. Educ. Psychol.* 60:101826. doi: 10.1016/j.cedpsych.2019.101826

Wu, Y., and Schunn, C. D. (2021). From plans to actions: a process model for why feedback features influence feedback implementation. *Instr. Sci.* 49, 365–394. doi: 10.1007/s11251-021-09546-5

Wu, Y., and Schunn, C. D. (2023). Assessor writing performance on peer feedback: exploring the relation between assessor writing performance, problem identification accuracy, and helpfulness of peer feedback. *J. Educ. Psychol.* 115, 118–142. doi: 10.1037/edu0000768

Zong, Z., Schunn, C. D., and Wang, Y. Q. (2020). Learning to improve the quality peer feedback through experience with peer feedback. *Assess. Eval. High. Educ.* 46, 973–992. doi: 10.1080/02602938.2020.1833179

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 25 July 2023
DOI 10.3389/fpsyg.2023.1125483



OPEN ACCESS

EDITED BY
Rui Li,
Hunan University, China

REVIEWED BY
Hao Xu,
Beijing Foreign Studies University, China
Yanhua He,
Shanghai International Studies University, China
Mohammad H. Al-khresheh,
Northern Border University, Saudi Arabia

*CORRESPONDENCE
Jing Huang
✉ jphuang@eduhk.hk

RECEIVED 16 December 2022
ACCEPTED 10 July 2023
PUBLISHED 25 July 2023

CITATION
Zhao JZ and Huang J (2023) A comparative study of frequency effect on acquisition of grammar and meaning of words between Chinese and foreign learners of English language.
*Front. Psychol.* 14:1125483.
doi: 10.3389/fpsyg.2023.1125483



# A comparative study of frequency effect on acquisition of grammar and meaning of words between Chinese and foreign learners of English language

Jizheng Zhao[1] and Jing Huang [2]*

[1]School of Foreign Languages and Business, Shenzhen Polytechnic University, Shenzhen, Guangdong, China, [2]Department of English Language Education, Education University of Hong Kong, Hong Kong, Hong Kong SAR, China

Frequency effect on vocabulary acquisition has been widely investigated in second language acquisition (SLA) research, whereas comparative studies of vocabulary acquisition of learners from different language types, such as hieroglyphic writing and alphabetic writing, are still rarely found. This type of studies could be of great significance in exploring some unique characteristics of how second language learners of native languages of different writing perceive and acquire second language. Using artificial words of alphabetic writing and low-frequency English words as experimental materials, this study aims to compare the effect of frequency on the acquisition of grammar and meaning of alphabetic words between Chinese learners of the hieroglyphic native language and foreign learners of alphabetic native languages. Specifically, the study intends to find out whether frequency effect plays the key role in language acquisition; to what extent frequency effect affects language acquisition; and whether there are any differences between learners of different language types for vocabulary acquisition in terms of frequency effect. The results show that Chinese and foreign learners of English language have no significant differences as a whole in terms of type of languages affecting the acquisition of grammar and meaning of artificial words and English words, indicating the difference in the type of mother tongue might not be the factor causing differences on grammar and meaning acquisition of vocabulary. Learner types, language types, frequency and part of speech of a word have interaction effect toward the acquisition of grammar and meaning of a word. However, exposure frequency of vocabulary plays the determining role in the acquisition of grammar and meaning of words.

KEYWORDS

**frequency effect, grammar, meaning, alphabetic, hieroglyphic, second language acquisition (SLA)**

## Introduction

Many researchers (Larsen-Freeman, 1997; MacWhinney, 1999; Ellis, 2002a,b, 2008; Schwartz and Causarano, 2007) claimed that frequency has a major role in second language acquisition. Ambridge et al. (2015) indicated that frequency effects were ubiquitous in virtually every domain of human cognition and behavior, from the perception of facial attractiveness (Grammer and Randy, 1994) and the processing of musical structure (Temperley, 2007) to language change (Bybee, 2010) and adult sentence processing

(Ellis, 2002a). Ambridge et al. (2015) claimed that frequency effects are pervasive in children's first language acquisition and argued therefore that any successful account of language acquisition, from whatever theoretical standpoint, must be frequency sensitive to the extent that it can explain these effects. Ellis and Ogden (2015) further emphasized that the same conclusion follows from 60 years of psycholinguistic research into the fluent language processing that culminates from acquisition: Language processing is exquisitely sensitive to usage frequency at all levels of representation, e.g., phonology and phonotactics, reading, spelling, lexis, morph syntax, formulaic language, language comprehension, grammaticality, sentence production, and syntax.

Ellis (2002a) pointed out that the frequency effect, after 40 years of exile, returned to researchers' focus again. According to Leech (2011), three theoretical positions that have been gaining momentum since the 1990s, all implicitly or explicitly give frequency a role in the workings of language: usage-based linguistics, cognitive linguistics, and construction grammar. Ellis (2002a) believed that language acquisition is cumulative example-based learning of thousands of constructions, as well as a frequency-based process of abstracting internal rules. Regularities of language emerge when learners are exposed to categories and prototypes. The frequency effect plays an important role in explaining sociolinguistic variants and language changes.

Andringa and Rebuschat (2015) held similar ideas for the role of frequency in language acquisition, indicating that statistical learning is an incremental accumulation of language knowledge on the basis of input distribution characteristics, with learners being very sensitive to the input distribution characteristics. A significant characteristic of statistical learning is that it will emerge automatically and unconsciously when people are exposed to language input.

In fact, since the late 20th century, many theories in the linguistics sphere, including implicit learning theory (Ellis, 1994, 2002a), dynamic system theory (DST) (Larsen-Freeman, 1997, 2007; Verspoor et al., 2008), construction grammar theory (Langacker, 1987), connectionism (Rumelhart and Melland, 1986, 1987), emergentism (Elman et al., 1996; Ellis and Schmidt, 1998; Ellis and Larsen-Freeman, 2006), and usage-based language theory (Bybee, 2006; Zhao, 2017), which are from the same theoretical paradigm and inextricably related, tend to support the view that language learning mechanism is not different from the other cognitive mechanism that language is acquired by using.

Besides, embodied philosophy represented by Lakoff and Johnson (1999) also emphasized that human language is derived from language use and formed through the interaction of brain, body and the environment. It proposed a theory of body-mind unity and internal-external unity (unity of man, nature and society) which provides a good foundation for taking some other factors into consideration.

Despite the fact that research on frequency effect on SLA has been extensively carried out over decades, comparative studies between EFL learners of different language types remain scant. What role does exposure frequency play in target vocabulary acquisition? Could it be possible that the difference of writing forms of native languages has different frequency effect on vocabulary acquisition of the target language? Does exposure frequency interact with the other factors, such as learner type, language type and the grammar of a word, in affecting target vocabulary acquisition? Adopting an experimental design (Creswell, 2012; Creswell and Guetterman, 2019), the present study set out to explore all these questions.

## Literature review

Harrington and Denis (2002) classify the frequency effect as "an attribute of individual experience" and "an attribute of the linguistic environment." The former is called "task frequency," and the latter "distribution frequency." This study mainly focuses on task frequency, at which a learner is exposed to a linguistic item.

### Frequency rates and vocabulary acquisition

For the purpose of determining what might be the "most" appropriate exposure frequencies for vocabulary acquisition, researchers have carried out extensive studies on the frequency effect on this from different perspectives. Studies on the correlation between exposure frequency and the acquisition and retention of vocabulary are one of the perspectives that some researchers focus on. Saragi et al. (1978) showed that the correlation between vocabulary acquisition and vocabulary frequency was 0.34, which presents the frequency effect on learning but indicates that learners need to be exposed to a word 10 or more times (a common number for mastering vocabulary knowledge) before achieving a significant effect on vocabulary knowledge acquisition. Horst et al. (1998, p. 215) showed that the correlation coefficient between exposure frequency and acquisition is 0.49, indicating at least 8 or more times of exposure required for vocabulary acquisition; that is, with fewer than 8 times of exposure, the acquisition effect would be difficult to predict. The study found that notional word acquisition had a higher acquisition score, and the images had a significant effect on the acquisition. The study by Waring and Takaki (2003) found that learners could have a 50% probability of recall and comprehension of a word 3 months later if they were exposed to a word in the target language at least 8 times. If the learners were exposed to the word 18 times, it was likely for them to remember the meaning of the word after 3 months. Therefore, they recommended that learners be exposed to a new word more than 20 times for acquisition. Rott (1999) studied the effects of vocabulary exposure frequency on vocabulary acquisition and retention in the mid-level language learners' reading process. The results showed that the students exposed to the words (2, 4, or 6 times) were significantly better at mastering vocabulary knowledge than the students who were not exposed to the words; both 4 and 2 exposures resulted in no significant changes in the acquisition of vocabulary input and output knowledge, while 6 exposures had a significant effect on the acquisition of the two kinds of knowledge compared with 4 exposures.

Studies on exposure frequency and acquisition of various knowledge of vocabulary are another perspective that researchers focus on. Webb's (2007) study focused on the frequency effect and seven aspects of vocabulary knowledge. Learners are exposed to the words 1, 3, 7, or 10 times. The results showed that each exposure to a word increased knowledge in at least one variable. If learners

are exposed to the word 10 times, they acquire objective word knowledge. The mastery of word knowledge, however, may require more than 10 exposures. Chen and Truscott (2010) studied whether L1 lexicalization affects vocabulary acquisition. The lexicalized words referred to words having equivalent translations and fixed linguistic items in Chinese, while the non-lexicalized words had no equivalent translations in Chinese. The results showed that the frequency of exposure to lexicalized words had a greater effect than non-lexicalized words in acquisition, demonstrating that the number of exposures had a positive and significant impact on learning. However, for the acquisition of non-lexicalized words, both the immediate post-test and the delayed post-test showed reduced acquisition. Even after 7 exposures, it is still impossible to acquire non-lexicalized words.

In addition, Sun (2014) examined the effect of contextual richness on vocabulary acquisition and studied the relationship between frequency and vocabulary knowledge acquisition (see also Webb, 2007). Zhang and Qi (2009), using natural authentic reading materials, studied incidental vocabulary acquisition. Song and Sardegna (2014), focusing on the grammatical perspective, studied the frequency effect on the acquisition of English prepositions and showed that frequency exposure to propositions of various contexts and learners' participation in output activities could help to acquire the target features. Aka (2019) investigated the frequency effect on the acquisition of the grammar structure of to-infinitive used as noun, indicating frequent exposure to target grammar items repeatedly helps learners notice a grammatical rule and will contribute positively to grammar acquisition. Zhang (2020) finds that the frequency effect on the processing of formulaic sequences by Chinese native speakers is significant. Zhang and Zhang (2022) studied the developmental features of the receptive-productive continuum of L2 academic vocabulary, and the results showed that there was a significant positive correlation between subjects' overall proficiency of academic vocabulary and the frequency level of the corresponding vocabulary. Although the character of frequency in this study is distributive, the researchers emphasized that a higher distributive frequency vocabulary meant more probabilistic and contacting opportunities in input and that the vocabulary had more chances to be activated.

## Embodied cognition and SLA

To make the present study more justified, this research has also taken into consideration the interplay of various internal and external factors. Atkinson (2010, p. 599) stated:

> *. . .conceptions of cognition have changed radically over the past century. . .. . ..toward extended and embodied views of cognition. Extended cognition conceptualizes mind/brain as inextricably tied to the external environment, while embodied cognition views cognitive activity as grounded in bodily states and action. These two approaches are related because bodies link minds to the world–we experience, understand, and act on the world through our bodies.*

Boden (2006) indicated that instead of being the self-contained logical system posited by cognitivism, cognition depends heavily on the external environment. Atkinson (2002) developed the notion of sociocognitive perspective on SLA and advocated that language and language acquisition as simultaneously occurring and interactively constructed both "in the head" and "in the world" (p. 525).

According to embodied psychology and language cognition (Wang, 2012), language acquisition should focus on the integration of language cognition and the physical and external environment, as well as the role that the body and the environment play in the cognition process. The theory pays more attention to the physical body, the local environment (situation) and the interaction of the nerve system with the corresponding external environment. Human brain, body and environment are constantly changing and interacting. The true cognitive system is a unified system consisting of all three. Therefore, a common point of embodied philosophy and cognitive linguistics is that categories, concepts, reasoning, and thought of human beings are formed through people's physical experience, and language is gradually formed through people's cognitive processing relying on the interaction of the sensory organs with the real world. Language acquisition is the result of multiple interactions between subject (human being) and object (environment). These ideas are in line with Atkinson's (2010, p. 612) sociocognitive view of SLA:

> *. . .sociocognitive approaches to SLA are based on this tripartite premise: (i) Mind, body, and world are in continuous processes of interaction alignment; (ii) These processes are partly public; and (iii) In being public, they are learnable. Thus, if cognition is the site of learning, it is extended, embodied cognition that makes learning possible, at least in part.*

Besides, embodiment theories are regarded as being capable of complementing usage-based approaches and should be incorporated into existing L2 theories (Patterson, 2021). Using usage-based and embodiment approaches, Patterson (2021) investigated second language listening functor (function words) comprehension probability. Transcription of functors were used as the dependent variable and frequency, word length, and Minkowski3 sensorimotor ratings as independent variables. The results showed that greater frequency, longer word length, and higher Minkowski3 ratings were found to facilitate comprehension and significantly increase the probability that a functor was transcribed. Frequency rates derived from spontaneous L1 oration and conversations were found to be significant.

## Usage-based approach and frequency effect

In fact, usage-based approach is widely used in frequency effect studies of SLA (Bybee, 2006; Ellis et al., 2008; Ellis and Larsen-Freeman, 2009). Usage-based linguistics argues that language acquisition takes place through implicit learning (using cognitively generic learning strategies) of patterns of form and meaning encountered in language input. Ellis and Ogden (2015, p. 283) held the view:

> *Learning, memory, and perception are all affected by frequency of usage: the more times we experience something, the stronger our memory for it, and the more fluently it is accessed. The more recently we have experienced something, the stronger our memory for it. The more times we experience conjunctions of features, the more they become associated in our minds and the more these subsequently affect perception and categorization; so a stimulus becomes associated to a context and we become more likely to perceive it in that context.*

Usage-based linguistics thus recognizes the impact of language usage on language cognition representation. It emphasizes that as users are exposed to language tokens, they classify their forms in different abstract forms. This classification process forms a network that includes speech, semantics, and pragmatics. This type of network is subject to language frequency. The usage-based language theory actually regards language knowledge as a set of automatic, generalized sentence patterns.

In terms of the influence of frequency on SLA, the usage-based theory holds that: (1) L2 language learners find it difficult to learn language because of a lack of a mother tongue acquisition environment; (2) L2 learners have the comparison mechanism for language decoding and output as well as the mechanism for linguistic and non-linguistic classification. These mechanisms can be used for acquiring new languages. The only requirement for L2 learners is to have sufficient exposure to second language (L2). At the same time, chunking and automaticity processes require a wealth of links between language and non-language to reach fluency.

## Studies using samples of different language types

Studies of this type could be of great importance for exploring unique characteristics such as how second language learners of different native languages perceive and acquire second language; how different related factors affect SLA; what role exposure frequency plays in the acquisition process; and whether frequency may have any universal effect on SLA for learners of different native languages. Therefore, research in this line, may involve learners of different native languages, learners' native cultures and life environments, their perception of different languages, the interaction of linguistic and non-linguistic factors while learners are learning a new language.

The study by Chen et al. (2020) might be one of the few comparative studies on the frequency effect on SLA of learners of different native language types. Using artificial words of alphabetic writing as experimental materials, they investigated the acquisition of alphabetic word forms between Chinese learners of the hieroglyphic native language and foreign learners of the alphabetic native language. The results showed that the difference in the character pattern of the mother tongue could result in disparity of the acquisition rate of the character pattern, and the word acquisition rate of the same character patterns was higher. The results also showed that input frequency could go beyond the difference of mother tongue and shared some common features, indicating that the frequency of being exposed to the language can overcome and transcend the barriers of language differences during language acquisition.

Perez-Paredes and Bueno-Alastuey's (2019) study consisted of subjects of native speakers (NSs) and Non-native speakers (NNSs) of Chinese, German and Spanish. The research explored the most frequent certainty adverbs in the extended LOCNEC and their frequency and use in three datasets of the LINDSEI (Chinese, German, and Spanish LINDSEI components). The study yielded a complex picture and no simple rule could be drawn from the data on the use of stance adverbs by learners of different native languages. An important finding relevant to the present study is that NSs and Chinese frequencies of use for adverbs were not significantly different. The researchers believed that this might be attributed to that the two groups approached the task in ways different from the German and Spanish speakers. In this study, an examination of the pragmatic contexts of using the certainty adverbs revealed that both NSs and NNSs restricted their semantic choice to classic epistemic meanings with few instances of more complex pragmatic meanings. Complex might be the results of this study, we can still find that the learners, in spite of the differences of their native languages, share more similarities than differences in using the target language.

Ament et al. (2020) explored the distribution of pragmatic marker (PM) use by English as a Foreign Language (EFL) speakers and English native speakers (NSs). Participants were second-year (N1/423), and third-year (N1/418) business undergraduates, and a NS control group (N1/410). Via English-medium instruction (EMI), the researchers increased learners' contact with English to explore the use of textual PMs in their oral communication. The results indicated that the EMI groups used PMs for causal, contrast and sequential functions at similar frequencies as NSs, and that the NSs used PMs significantly more often for continuation and elaboration functions and significantly less opening and closing functions compared to the EMI groups. The study suggested that EMI may play an important role in facilitating the acquisition of some functions of PMs, whereas other PMs, such as elaboration markers, may take longer to acquire.

Zhang and Fang (2020) investigated frequency effect on collocation processing of native speakers of English (NSs) and Chinese EFL learners (NNSs). Same-translated collocations and different-translated collocations were chosen as the experimental materials. Online acceptability judgment task of English collocations was used to measure subjects' performance. The study showed that both NSs and NNSs processed more accurately same-translated collocations but not faster than judging different-translated collocations; NNSs' language proficiency modulated the effects of constituent word frequency and collocational frequency on the processing output; and lexical frequency played a modulating role in the processing of all types of collocations for both NSs and NNSs. The results indicated that the ultimate goal of second language learners was to infinitely approach the overall processing of collocations in the native language. Zhang and Fang (2020) assumed that frequency is the determining factor for collocation acquisition, and frequency of exposure is ultimately experience, an example, and the use of language. In terms of frequency effects, there is no bipolar debate between NSs and

NNSs; it is a gradual transition and evolution from dependence on rules to overall synthesis as the frequency of contact increases and language proficiency grows.

Based on the thorough review of related literature in this section, we may draw the following conclusions: (1) Exposure frequency plays an important role in language acquisition. (2) The human being's body and brain may interact with the environment and unify to affect language acquisition. (3) Learners have language processing mechanism which can equip them with the ability to distinguish linguistic and non-linguistic factors. While learners expose themselves to the input-rich environment in which linguistic factors and non-linguistic factors interact constantly, they are supposed to be able to achieve language fluency. Therefore, (4) learners' native languages and cultures, learners' mental and physical factors, learning contexts and environments, the interaction between learner physical condition and environment, the interaction of linguistic and non-linguistic factors, are all possible factors affecting SLA. (5) Comparative studies on frequency effect of different language types toward second language vocabulary acquisition are still at its initial stage, and therefore, more studies in this line are needed.

## Research questions

This study wants to continue with the line of comparative research on vocabulary acquisition and concentrates on the acquisition of grammar and meaning of words between learners of different language types. The purpose of this study is to find out whether frequency effect, among all the factors relevant to SLA, plays the key role in language acquisition and to what extent it affects language acquisition; and whether frequency exposure of language might have some universal effect on SLA across learners of different cultures.

This study assumes that language could be a symbolic icon representing culture. People's value and perception of the world could be embedded in languages and they may affect the process of SLA implicitly. Therefore, in this experiment we choose two types of learners of English whose native languages are very different in forms: Chinese learners of hieroglyphic writing and non-native speakers of alphabetic writing. To avoid culture bias for Chinese learners of hieroglyphic writing, and foreign learners of alphabetic, a language which could both symbolize alphabetic writing and without much cultural embodiment might be an appropriate choice as one of the target languages, for which an artificial Keki language (McCandliss et al., 1997) is chosen. Besides, the study also aims to investigate what the situation might be while learners acquire a real language. Hence, the study uses the low-frequency words in high-frequency category of Corpus of Contemporary American English (COCA), for the purpose of both symbolizing the form of real alphabetic language to the greatest extent, and with the least possibility of representing its meaning.

The study addresses the following research questions (RQs):

> *RQ1. Do Chinese and foreign learners of English differ in the frequency effect on the acquisition of grammar and meaning of artificial words and English words?*
>
> *RQ2. What is the general role of frequency in Chinese and foreign learners' acquisition of grammar and meaning of artificial and English words?*
>
> *RQ3. Does the interaction of factors such as learner type, language type, frequency and part of speech influence lexical acquisition?*
>
> *RQ4. Does the acquisition of grammar and meaning of words vary in accordance with the difference in language type?*

## Materials and methods

### Subjects sampled for the experiment

To maintain the validity of the experimental data, we chose 30 subjects for both Chinese group and foreign group. And 30 subjects are regarded sufficient for a psychological experiment (Chen, 2005, p. 43). The following criteria were established to keep the homogeneity of the subjects, including (1) all subjects had not previously lived or studied in countries of native English language; (2) they were all tertiary level students, and their English proficiency should be at the same level; (3) they all had normal visual acuity (corrected or uncorrected). The second-year Chinese students of English major whose native tongue is hieroglyph and foreign students whose native language is alphabetic writing were the two target subject groups. Both the Chinese learners and foreign learners were in a same university in southern China.

To keep consistent the proficiency level of the participants, we invited the English teachers of these two target groups to evaluate these students' English level, based on these students' formative achievements (e.g., quiz scores) and the results of semester final examination to eventually choose 60 Chinese and 60 foreign volunteer students of similar English proficiency level as the candidate participants. Further, we adopted College English Test Band 6 (CET6) July 2020 as the tool to test the subjects' English proficiency. Given that some subjects are foreign students, the sections of Writing and Translation were eliminated and only sections of Listening (total score = 248.5) and Reading (total score = 248.5) were kept. The testing procedures of these two sections were strictly implemented as required in the real test. The results showed that there was no significant difference in terms of English proficiency between the two groups ($T$ = 0.075, $P$ = 0.092). We sampled 30 subjects from each group whose test results were among the middle range ($M$ = 388 for the Chinese group; $M$ = 392 for the foreign group) of the 60 candidate participants in each group. The foreign group consisted of students from Russia, Republic of Korea, Kazakhstan, Slovakia, and Italy.

### Ethical considerations

The experiment was conducted with the participants' informed consent. Before the experiment, we explained the purpose and content of the experiment to the participants. We emphasized that

different national and ethnic cultures and different writing forms of their native languages would be used for academic purpose only, without prejudice against any specific cultures and writing forms. We respected the freedom of participants and allowed them to withdraw from the experiment as they wish. All possible measures were taken in the experiment to ensure that participants did not experience any adverse reactions due to their participation in the experiment. The participants were informed that their performance during the experiment and the results of the experiment would be kept strictly confidential. They would be awarded ¥40 each for participation.

## Experimental materials

### Artificial words and low frequency English words

Two language types are chosen for the experiment: artificial Keki language and English language. The artificial Keki language includes 68 artificial words (see **Supplementary Appendix 1**). The composition of the words follows the rules of C(C)VC(C)V (C stands for consonant, V vowel). There are many similarities in composition between Keki and English words; the only difference is that the words in Keki end in vowels. The uniqueness of Keki words means that they have some features of English spelling but are different from English. These features can properly reflect the characteristics of alphabetic writing with no similarities to hieroglyphics and therefore have favorable representation and test validity for measuring differences in words between alphabetic writing and hieroglyphics. The materials consist of two parts, one of which is used in the learning phase and the other in the test phase. Part 1 materials in the learning phase include artificial words (see **Supplementary Appendix 2**) and low-frequency English words (see **Supplementary Appendix 3**). Artificial words are from the artificial Keki language. We chose the target artificial words according to the parts of speech of target words needed for the experiment. Therefore, some more artificial words were created based on the word formation rules of the Keki language to satisfy the requirement of experiment. The experimental materials include 4 groups of notional words; each has 6 words, including 2 nouns, 2 verbs, and 2 adjectives, for a total of 24 words. The exposure frequencies of the four groups of words were 1 time, 3 times, 7 times, and 10 times. The corresponding learning materials are low-frequency English words. The grouping form, part of speech in the groups, and frequency of words presented are identical to the artificial words.

The 24 English words in the four groups of the experiment are low-frequency words selected from the sampling list of COCA, extracted from every 7 words in the list of the top 60,000 high-frequency words in the COCA corpus, and including a total of 8,574 words. On this basis, we select 24 notional words from back to front in the list. Therefore, the 24 English words used in the experiment are low-frequency words in the 60,000 high-frequency words in the COCA corpus. Such selection can both satisfy learners' perception of high-frequency words and ensure that students probably have not been exposed to such words to the greatest extent possible.

The reason we use artificial words and English vocabulary at the same time in the experiment for meaning and grammar lies in the fact that the artificial Keki language has the characteristic of alphabetic writing, but it is different from English, which has a certain neutrality. For Chinese learners of hieroglyphs, Keki language has similarities with English language in its form. For foreign students of alphabetic writing, it is similar to their native language, as well as English. Furthermore, the critical difference between the artificial words and real words lies in that the former is assumed not containing cultural information, while the latter contains cultural information. The use of such artificial words can commendably test the real condition of learners of English from hieroglyph and alphabetic backgrounds on the grammar and meaning of alphabetic writing. For the adoption of English to experiment on Chinese and foreign learners, its purpose is to test, in real language, whether there is a difference between them in the acquisition of grammar and the meaning of vocabulary and whether it differs from the grammar and meaning of vocabulary of the artificial language. If differences exist and are distinct, it is suggested that the word form difference of characters might also be an important reason for the acquisition difference of grammar and meaning of vocabulary of Chinese and foreign learners with different word forms in their mother tongues, and the language environment and cultural differences behind vocabulary acquisition are worthy of study.

Part 2 materials of the experiment are testing materials, including test questions of artificial words (see **Supplementary Appendix 4**) and English low-frequency words (see **Supplementary Appendix 5**). Sentence completion with multiple choices is adopted. There is one space in each sentence, which requires the subjects to choose one right answer from the four options and fill in the blank. The four options include distractors developed based upon misconceptions of word meaning and language type and with special attention given to grammar. For example, test on the artificial word *gonta*:

gonta

*She seems very gonta with the result, for that is all what she can do.*

*A. sad B. happy C. greatly D. gone.*

Among the four options, two adjectives, including "sad" and "happy," one adverb, "greatly," and one past participle "gone," are included. When presenting the target words in E-Prime, apart from the meaning of the target words, the grammatical element–part of speech of the word–is also presented, such as "gonta, adj. happy." Coupled with the connection of the target word with the word before and after it in the sentence, learners make a judgment: an adjective should be chosen for this space, and its meaning should be happy. Therefore, the choice of right answer not only requires the subjects to infer the meaning but also the part of speech of the target word.

### Measurement of grammar and meaning of vocabulary

The grammar of vocabulary involved in this experiment is presented through the part of speech of words. Considering the distribution of part of speech of words in language and in avoidance of the impact of close-class words such as articles and prepositions on the test results, three types of open-class words–noun, verb and adjective–are selected for this experiment. The accuracy of words with different parts of speech selected by the subjects in

the experiment can be regarded as the acquisition rate of the grammar knowledge of words by them. The accuracy of acquisition of the meanings of words is determined by the correctness of the meanings of words selected by the subjects. In other words, if learners select the correct answer during the test period in the experiment, it is supposed that they have mastered the grammar and meaning of such words. Therefore, in this experiment, the accuracy of acquisition of grammar and meaning of different language types is determined by selection of the right answers from multiple choices.

## Experimental design

Factorial design (Creswell, 2012, p. 311) was adopted and a multi-factor mixed design was created. Four factors used as independent variables are learner type, language type, part of speech of the target word and exposure frequency of target words. Factor as dependent variable is subjects' achievements in the test task. The purpose of this design is to test the main effect of frequency and acquisition and at the same time the interaction effect of all the factors for acquisition. Specifically multi-factor $2 \times 2 \times 3 \times 4$ mixed design is used in the experiment. Independent variable 1 (variables between subjects) is learner type (two groups: foreign students from non-native English-speaking countries and Chinese learners of English); independent variable 2 (within subjects) is language type (artificial language and English language); independent variable 3 (within subjects) is the part of speech of the target word (adjective, noun and verb); and independent variable 4 is the exposure frequency of target items, including four frequencies in total (1 time, 3 times, 7 times, and 10 times). In the experimental design, repetition within and between the frequencies of words was avoided. In addition, words of the four frequencies were presented randomly, which aimed to prevent students from feeling tired or guessing the answers during the learning and testing period. The dependent variable is tested by using multiple choice questions. According to the number of target words, we designed six sentences in which the target words are underlined and four options are provided for each question. Other than the test on understanding of meanings, distractors are also designed to test the subject's grammar knowledge. A score of 5 points is assigned for each question, 5 points for a right choice and zero for a wrong choice. Therefore, the full score for the target word of each part of speech is 10 points for each exposure frequency. The full score of the four frequencies is 120 points. The score is counted through a computer.

## Experimental procedure

Learning-recognition paradigm is widely used in psychological experiments (e.g., Liao and Zhang, 2012; Zhang and Xing, 2012). The procedure is divided into two steps: "learning" and "recognition"; "recognition" is immediately implemented after the subjects complete the task at the "learning" stage.

The learning and recognition steps are computer-based and programmed and recorded by E-prime. In the experiment, after one level of exposure frequency (1, 3, 7, or 10) of words, the recognition test is carried out immediately. During the learning period, the instruction is first presented to the two subject groups of Chinese and foreign learners, which is shown in both Chinese and English.

When presenting the target word, three types of information are shown in each page on the screen–target word, part of speech and meaning of the target word. Before the presentation of each artificial word, a string of "*" is shown on the screen, lasting for 500 ms, priming the subjects' attention, followed by artificial words presented, lasting for 8 s for each word without an interval. The subjects rest for 10 s after learning a sequence.

After the break, the subjects come to the recognition test phase for the experiment. The presentation time for each word is 15 s without an interval. The recognition test is conducted immediately after learning.

# Results

According to the questions under study, we sorted the experimental data and performed statistical analysis with SPSS17.0. Four copies of invalid test materials in each group in the experiment were rejected; 26 copies of valid test materials in each group were used for analysis. To learn the overall situation of the acquisition of target words in the sentence context of Chinese and foreign learners, we performed descriptive statistics. The acquisition of artificial words and English words is shown in **Tables 1**, **2**, respectively.

From the perspective of the acquisition effect of artificial words, in exposure frequency 1, foreign students have the best acquisition effect on adjectives ($M = 8.08$), and Chinese learners have the worst acquisition effect on nouns and verbs ($M = 6.5$). In terms of the acquisition result of three different parts of speech in exposure frequency 1, the acquisition rate of foreign learners is higher than that of Chinese learners. In exposure frequency 3, the acquisition effect on nouns by foreign students is the best ($M = 9.23$), and on verbs is the worst ($M = 6.92$), and the average of verb acquisition of Chinese students and that of foreign students is the same. From the view of the overall effect of exposure frequency 3, foreign learners show higher acquisition of adjectives and nouns and are commensurate with Chinese students in the acquisition of verbs. In the acquisition of artificial words of exposure frequency 7, the acquisition effect on adjectives by foreign students is the best ($M = 9.62$), and on adjectives and verbs by Chinese students is the worst ($M = 8.46$). From the general situation of part-of-speech acquisition of artificial words of exposure frequency 7, the average acquisition rate of foreign students is higher than that of Chinese students. From the acquisition of artificial words of exposure frequency 10, foreign and Chinese students have the best acquisition effect on nouns and the former on verbs (the mean of the three is the same, $M = 9.04$), and Chinese learners have the worst acquisition effect on verbs ($M = 8.27$). In the 10-time exposure frequency condition, the difference in the acquisition rate of different vocabularies by Chinese and foreign learners is smaller than that when the exposure frequency is 3 and 7.

From the perspective of the overall effect of the acquisition rate of artificial words, frequency is still a major factor that leads to acquisition differences. With the increase in exposure frequency, the acquisition rate of words of all the parts of speech increases. However, under the same frequencies, the acquisition rate of

TABLE 1 Acquisition mean of artificial words with different parts of speech at different frequencies of Chinese and foreign learners.

| | Learner type | Case number | Mean | Std. deviation | Mean of std. error |
|---|---|---|---|---|---|
| A1adj | Foreign learner | 26 | 8.08 | 3.187 | 0.625 |
| | Chinese learner | 26 | 6.92 | 3.486 | 0.684 |
| A1n | Foreign learner | 26 | 7.50 | 3.240 | 0.635 |
| | Chinese learner | 26 | 6.54 | 3.679 | 0.722 |
| A1v | Foreign learner | 26 | 7.50 | 3.240 | 0.635 |
| | Chinese learner | 26 | 6.54 | 4.188 | 0.821 |
| A3adj | Foreign learner | 26 | 8.27 | 3.144 | 0.617 |
| | Chinese learner | 26 | 7.12 | 3.788 | 0.743 |
| A3n | Foreign learner | 26 | 9.23 | 1.840 | 0.361 |
| | Chinese learner | 26 | 7.31 | 3.803 | 0.746 |
| A3v | Foreign learner | 26 | 6.92 | 4.019 | 0.788 |
| | Chinese learner | 26 | 6.92 | 4.019 | 0.788 |
| A7adj | Foreign learner | 26 | 9.62 | 1.961 | 0.385 |
| | Chinese learner | 26 | 8.46 | 3.397 | 0.666 |
| A7n | Foreign learner | 26 | 9.42 | 1.629 | 0.319 |
| | Chinese learner | 26 | 9.04 | 2.010 | 0.394 |
| A7v | Foreign learner | 26 | 8.65 | 3.019 | 0.592 |
| | Chinese learner | 26 | 8.46 | 3.397 | 0.666 |
| A10adj | Foreign learner | 26 | 8.46 | 2.746 | 0.538 |
| | Chinese learner | 26 | 8.65 | 3.019 | 0.592 |
| A10n | Foreign learner | 26 | 9.04 | 2.835 | 0.556 |
| | Chinese learner | 26 | 9.04 | 2.835 | 0.556 |
| A10v | Foreign learner | 26 | 9.04 | 2.457 | 0.482 |
| | Chinese learner | 26 | 8.27 | 2.426 | 0.476 |

Number (1,3,7,10) = frequency (1,3,7,10); A, artificial; adj., adjective; n, noun; v, verb.

foreign students is generally higher than that of Chinese students (except at frequency 10, for the acquisition of adjectives, the mean was 8.46 of the acquisition rate of foreign students and 8.65 of Chinese students). In terms of the acquisition rate of the part of speech, the acquisition rate of adjectives and nouns is high, and that of verbs is low, but such a difference decreases with the increase in exposure frequencies.

Meanwhile, the data indicate that the common characteristic of Chinese and foreign learners in the acquisition of artificial words was that the acquisition level of all the words was higher than 50%. The acquisition of adjectives at exposure frequency 7 by foreign learners was the highest ($M$ = 9.62), and that of nouns and verbs at exposure frequency 1 by Chinese learners was the lowest ($M$ = 6.45). Even the lowest acquisition rate was 10.45% higher than chance. This result proves that, regardless of the native language family of Chinese and foreign learners, in the sentence context, a similar effect exists in the exposure frequency of their acquisition of artificial words.

Table 2 below shows the acquisition of English words of the subjects. At exposure frequency 1, the acquisition of adjectives by Chinese students was the best ($M$ = 7.88), and for verbs, it was the worst ($M$ = 4.04). From the acquisition result of three parts of speech at exposure frequency 1, the acquisition rate of nouns and verbs by foreign students was higher than that of Chinese students, while for the acquisition rate of adjectives, Chinese students were higher than foreign students. At exposure frequency 3, the acquisition of nouns by foreign students was the best ($M$ = 9.81) and poorest for verbs ($M$ = 8.46), identical to the Chinese students. In terms of the general effect at exposure frequency 3, foreign learners and Chinese learners showed improvements in the acquisition of nouns and adjectives ($M$ = 9.42 for the latter), while the acquisition level of verbs was equal for both. For the acquisition of English words at an exposure frequency of 7, the acquisition of verbs by Chinese students was the best ($M$ = 8.85) and lowest for adjectives ($M$ = 7.31 for both subject groups). Identification of part of speech and meaning of English vocabulary at exposure frequency 7 proved highest for verbs, then nouns, followed by adjectives. Here, the subject groups differed only slightly. The identification of verbs by Chinese students ($M$ = 8.85) was slightly higher than that of foreign students ($M$ = 8.08) and that of adjectives and nouns by both were the same ($M$ = 7.31 for adjectives, $M$ = 7.69 for nouns). In the acquisition of English words at an exposure frequency of 10, the performance with adjectives by Chinese students was the best ($M$ = 9.42), with the lowest score demonstrated by Chinese students on verbs ($M$ = 8.27) and by foreign learners on adj. ($M$ = 8.46). Similar to the results for artificial words, the

TABLE 2 Descriptive statistics of Chinese and foreign learners' acquisition of English words of different parts of speech under different frequencies.

| | Learner type | Student number | Mean | Std. deviation | Mean of std. error |
|---|---|---|---|---|---|
| E1adj | Foreign learner | 26 | 6.35 | 3.622 | 0.710 |
| | Chinese learner | 26 | 7.88 | 3.514 | 0.689 |
| E1n | Foreign learner | 26 | 4.62 | 2.418 | 0.474 |
| | Chinese learner | 26 | 4.23 | 1.840 | 0.361 |
| E1v | Foreign learner | 26 | 4.42 | 2.580 | 0.506 |
| | Chinese learner | 26 | 4.04 | 2.010 | 0.394 |
| E3adj | Foreign learner | 26 | 8.85 | 2.572 | 0.504 |
| | Chinese learner | 26 | 9.42 | 1.629 | 0.319 |
| E3n | Foreign learner | 26 | 9.81 | 0.981 | 0.192 |
| | Chinese learner | 26 | 9.04 | 2.457 | 0.482 |
| E3v | Foreign learner | 26 | 8.46 | 2.746 | 0.538 |
| | Chinese learner | 26 | 8.46 | 2.353 | 0.462 |
| EF7adj | Foreign learner | 26 | 7.31 | 3.234 | 0.634 |
| | Chinese learner | 26 | 7.31 | 3.803 | 0.746 |
| E7n | Foreign learner | 26 | 7.69 | 3.530 | 0.692 |
| | Chinese learner | 26 | 7.69 | 2.909 | 0.570 |
| E7v | Foreign learner | 26 | 8.08 | 3.486 | 0.684 |
| | Chinese learner | 26 | 8.85 | 2.572 | 0.504 |
| E10adj | Foreign learner | 26 | 9.23 | 1.840 | 0.361 |
| | Chinese learner | 26 | 9.42 | 1.629 | 0.319 |
| E10n | Foreign learner | 26 | 9.04 | 2.010 | 0.394 |
| | Chinese learner | 26 | 9.23 | 1.840 | 0.361 |
| E10v | Foreign learner | 26 | 9.23 | 1.840 | 0.361 |
| | Chinese learner | 26 | 9.04 | 2.010 | 0.394 |

Number (1,3,7,10) = frequency (1,3,7,10); E, English; adj., adjective; n, noun; v, verb.

performance differences at an exposure frequency of 10 on the parts of the speech task by Chinese and foreign learners are smaller than those at exposure frequencies of 7 and 3.

Regarding the overall results with English vocabulary, frequency is still a major factor that influences task performance. With the increase in exposure frequency, the performance with words from all parts of speech increases. However, under the same frequencies, the acquisition rate of foreign students in low-frequency exposure (such as frequency 1) is generally higher than that of Chinese students. With increasing frequency, the difference between them decreases. The growth is not linear, but at frequency 7, the acquisition rate fell back, and the overall acquisition rate was lower than that at frequency 3. At an exposure frequency of 10, the acquisition rate is largely improved again. In terms of the identification of the part of speech (except for the verbs at frequency 7, $M = 8.85$ for Chinese students and $M = 8.08$ for foreign students), the identification of adjectives and nouns was high and that of verbs was low, but such a difference decreased with increasing exposure frequencies.

Similar to the acquisition of meaning and grammar of artificial words, the universality of the frequency effect that transcends the native language family at statistical significance is also generated during the acquisition of meaning and grammar of English words. Frequency has a universal effect beyond the level of chance for both foreign learners and Chinese learners. Although the data show that the acquisition rate of nouns and verbs of Chinese and foreign learners on single exposure is low (40.4–46.2%), this does not mean that the word form of language type leads to low acquisition of grammar and meaning of words only by Chinese learners, since the acquisition rate of both Chinese learners and foreign learners is similarly low. In contrast, it might properly indicate that low exposure frequency has no significant effect on the acquisition of meaning and grammar of any word forms regardless of whether they are artificial or English.

To learn the relation between experimental factors and accurately understand the influence of frequency on the identification of different parts of speech of artificial words by Chinese and foreign learners, the researcher conducted a repeated measures variance analysis, as shown in Table 3. Four independent variables are involved in this experiment, including the learner, language type, part of speech and frequency. The dependent variable is the score for word recognition.

The data (see Table 3) show that the main effect on frequency is significant ($F = 53.491$; $p = 0.000 < 0.05$), and the main effect of part of speech is significant ($F = 6.953$; $p = 0.001 < 0.05$). While a significant interaction exists among language * frequency

TABLE 3 Variance analysis of repeated measurement of grammar and meaning of artificial words.

| Source | Type III sum of squares | df | Mean square | *F* | Sig. |
|---|---|---|---|---|---|
| Language | 28.926 | 1 | 28.926 | 2.023 | 0.161 |
| Language × learner type | 54.167 | 1 | 54.167 | 3.789 | 0.057 |
| Frequency | 1362.901 | 3 | 454.300 | 53.491 | 0.000 |
| Frequency × learner type | 11.058 | 3 | 3.686 | 0.434 | 0.729 |
| Part of speech | 60.136 | 2 | 30.068 | 6.953 | 0.001 |
| Part of speech × learner type | 9.495 | 2 | 4.748 | 1.098 | 0.338 |
| Language × frequency | 521.554 | 3 | 173.851 | 20.537 | 0.000 |
| Language × frequency × learner type | 10.737 | 3 | 3.579 | 0.423 | 0.737 |
| Language × part of speech | 30.088 | 2 | 15.044 | 3.364 | 0.039 |
| Frequency × part of speech | 205.248 | 6 | 34.208 | 6.265 | 0.000 |
| Frequency × part of speech × learner type | 46.274 | 6 | 7.712 | 1.412 | 0.209 |
| Language × frequency × part of speech | 112.220 | 6 | 18.703 | 4.000 | 0.001 |
| Language × frequency × part of speech × learner type | 20.393 | 6 | 3.399 | 0.727 | 0.628 |
| Learner type | 25.962 | 1 | 25.962 | 0.475 | 0.494 |

Significance value < 0.05.

TABLE 4 Comparison between acquisition of artificial and English words.

| Dep. variable: score | | | | |
|---|---|---|---|---|
| Languages | Frequency | Mean | Std. deviation | Case |
| Artificial | 1.00 | 7.1617 | 0.40212 | 6 |
| | 3.00 | 7.6000 | 0.82149 | 6 |
| | 7.00 | 8.9217 | 0.44634 | 6 |
| | 10.00 | 8.7450 | 0.31998 | 6 |
| | Total | 8.1071 | 0.90935 | 24 |
| English | 1.00 | 5.2583 | 1.53150 | 6 |
| | 3.00 | 8.9967 | 0.54010 | 6 |
| | 7.00 | 7.8317 | 0.56623 | 6 |
| | 10.00 | 9.2350 | 0.10114 | 6 |
| | Total | 7.8304 | 1.80014 | 24 |

($F$ = 20.537; $p$ = 0.000 < 0.05), language * part of speech ($F$ = 3.364; $p$ = 0.039 < 0.05), frequency * part of speech ($F$ = 6.265; $p$ = 0.000 < 0.05), and language * frequency * part of speech ($F$ = 4.000; $p$ = 0.001 < 0.05), namely, under their mutual action, significant differences in the overall acquisition rate of meaning and grammar of words are observed.

However, the data show that the main effects of language type (artificial words and English) ($F$ = 2.023; $P$ = 0.161) and learner type (Chinese and foreign learners) ($F$ = 0.475; $P$ = 0.494) are insignificant, suggesting that the difference in the acquisition of grammar and meaning of artificial words and English at low frequencies by Chinese and foreign learners is not apparent, indicating that more similarity and consistency are reflected between them.

To determine the detailed differences in acquisition between the artificial words and English words, the present study performed a descriptive analysis (see **Table 4**) and paired-$T$ test (Table not presented) on the frequency effect on the acquisition of artificial words and English words.

The result demonstrates that the average score of the acquisition of artificial words ($M$ = 8.1071) is higher than that of English words ($M$ = 7.8304). The highest score occurs in frequency 10 for English words, while the lowest score occurs in frequency 1 for English words. The paired-$T$ test result ($T$ = 0.370; $P$ = 0.736 > 0.05; two-tailed) indicates that there is no significant difference between artificial word acquisition and English word acquisition among Chinese and foreign learners.

## Discussion

From the experimental results, we find that the acquisition of grammar and meaning of artificial words by Chinese and foreign learners is complex and influenced by multiple factors. According to usage-based theory (Bybee, 2006; Tyler, 2010; Wang, 2011), any real language is used in a context and affected by the factors in the context. Language system and language competence base fundamentally on the use of language, and

language system is exemplar-based and is gradually formed by learners' frequent exposure to real communication situations. Besides, embodied cognitive linguistics emphasizes the unified influence of human brain, human body and the environment toward language acquisition. Learners' culture could be one of the factors, which implicitly affects SLA.

In this section, we discuss the experimental results in relation to the four questions that the current study set out to address.

> *RQ1. Do Chinese and foreign learners of English differ in the acquisition of grammar and meaning of artificial words and English words?*

The descriptive statistics for Chinese and foreign learners' acquisition of meaning and grammar of artificial words, variance analysis result (main effect of Chinese and foreign learners is insignificant), and multiple comparisons for frequency effect show that there is no significant difference in the acquisition of grammar and meaning of artificial words between both types of learners. This indicates that Chinese and foreign learners only differ slightly in the acquisition of the part of speech and meaning of artificial words of alphabetic writing, which is distinct from the result of Chen et al.'s (2020) study of word form acquisition of artificial words of Chinese and foreign learners. However, as a whole, foreign learners' acquisition of vocabulary at different exposure frequencies is better than that of Chinese learners.

Regarding the reasons why there are few significant differences between them in the acquisition of grammar and meaning of artificial words, we believe that although artificial words are closer to English words with regard to word form, compared with the meaning and grammar of target words, the difference between the two languages on word form is more obvious, and meaning may be interlinked or similar in the native language of Chinese and foreign learners. Furthermore, in this experiment, the target words are tested in the context of sentences, which means that the subjects (either Chinese learners or foreign learners) have more clues for obtaining knowledge of the target word than in the context of a single word, as Yang and Zhang (2021) indicated that frequency is sometimes embodied in one's world knowledge and is the result of one's past experience. Concerning the reason why the acquisition of the part of speech and meaning of artificial words by foreign learners is generally better than that of Chinese learners, we think it might be because the word form structure of artificial words is more similar to that of foreign learners' L1, which can make foreign learners pay less attention to the processing of word forms, while more attention resources can be used for grammatical and semantic recognition of words. Chinese learners of hieroglyphs do not have such cognitive prerequisites. The data reflect that with the increase in acquisition frequency, the difference between them continually decreases. This means that when learners of different types of native language are exposed to words frequently, their competence for grammar and competence for meaning converge.

Based on the experimental result, we may infer that for the acquisition of words in a new language, in the initial stage, the more similar the word forms are, the better the acquisition effect is. The similarity of word form is a major cause that leads to rapid mastering of language for learners of alphabetic writing. However, with the extension of learning time and the increase in exposure frequency, the gap between them will narrow. Hieroglyphic learners might speed up the acquisition of new languages after they adapt to them and integrate the new languages into their own language system, including the meaning and grammar of new languages.

Another important finding in artificial word acquisition is that, for Chinese learners whose character pattern of their L1 is obviously different from alphabetic writing, the impact of the exposure frequency on the acquisition of grammar and meaning of vocabulary has reached the level of significance, indicating that the effect of frequency has transcended language types and has similar functions to the grammar and meaning acquisition of vocabulary of different language types.

The acquisition data of grammar and meaning of English words of Chinese and foreign learners are more complex. In terms of the exposure frequency and acquisition rate, the characteristics of fluctuation and change are shown. The English vocabulary acquisition rate of the three parts of speech at exposure frequency 1 is generally low; it increases greatly at frequency 3, decreases at exposure frequency 7, and increases obviously at frequency 10 again. The non-linear learning curve reflects what is seen in reality. After acquiring substantial new knowledge, learners begin to internalize and reconstruct knowledge in the brain, compare and assimilate with existing knowledge, and then acquire new knowledge after the reconstruction.

In terms of the acquisition of words of different parts of speech, at the same exposure frequency, the acquisition of adjectives is the best, followed by nouns and then verbs. This is similar to the phenomenon observed in daily teaching, in which performance with adjectives and nouns is better than with verbs. This result is very similar to Horst et al.'s (1998) finding that notional word acquisition had a higher acquisition score, and the images had a significant effect on the acquisition. This might also be related to the easier identification of the meaning of adjectives and nouns. As this experiment only involves the grammatical characteristics of the part of speech of words, other than the number, case, tense and voice of words, the influence of the saliency of word meaning on word acquisition is easier to show.

From the perspective of learners, the data of acquisition of English words show that the acquisition similarity of the grammar and meaning is larger than the difference of Chinese and foreign learners. This might explain that even though the word form of learners' L1 is different, when two groups of learners learn a language at the same time for a period of time, they will no longer be influenced by the word form, grammar and meaning of their L1 and show more similarities in the acquisition of the target language. This point of view is also confirmed in **Table 2** by the result of Chinese and foreign learners' acquisition of different parts of speech at different frequencies. According to embodied theories (Boden, 2006; Wang, 2008; Atkinson, 2010), we assume that during the learning process the target culture embedded in the target language increasingly enhanced its effect toward the learners' (Chinese or foreign) acquisition and constantly interact with learners' cognition, and as a result, the acquisition achievement of Chinese learners and foreign learners tends to assimilate.

The data in **Table 2** are in line with the results presented in **Table 1** that Chinese and foreign learners have achieved good results in terms of the exposure frequency of artificial words and revealed the same characteristics of Chinese and foreign learners

that the acquisition rate for the words with the same part of speech by both learners is low at exposure frequency 1. More importantly, these characteristics are shown in the real language–the acquisition of English vocabulary. This further proves that the exposure frequency plays an important role for Chinese learners who use non-alphabetic writing, and the exposure frequency surpasses the language family of its L1 and has similar functions to lexical acquisition.

From the analysis of RQ1, it is found that at the initial period of vocabulary acquisition, effect of language types embedded with native cultural factor for language acquisition is obvious. But as learners expose them more to vocabulary, this effect tends to fade, and the effect of frequency increases. According to embodied philosophy and usage-based theory, we may assume that native context is always the first factor exerting important influence on SLA. Nevertheless, this situation may change with the change of relevant factors. In this study, frequency effect shows its great increasing impact on SLA. This finding is supported by Ament et al.'s (2020) study which suggested that EMI, with an emphasis on both context effect and frequency effect, plays an important role in facilitating the acquisition of some functions of PMs. In the present study, it is also found that the cultural factor, specifically the writing of native language in this study, changes while learners learning a new type of writing. With increasing exposures, the frequency effect becomes the most prominent factor in SLA. On one hand, as Chinese learners learn a language of alphabetic writing, they may be influenced directly by the obvious difference of the writing at the beginning of the learning; on the other hand, when learners of different native language learn a same language, the culture influence tends to converge and the influence of new language will gradually overpass the influence of the original languages. During this period, effect of exposure frequency becomes the key factor.

> *RQ2. What is the general role of frequency in Chinese and foreign learners' acquisition of grammar and meaning of artificial and English words?*

In general, frequency has a prominent effect on the acquisition of grammar and meaning of artificial words and English words, indicating that the change of exposure frequency will result in acquisition change, the higher the frequency of exposure, the better the acquisition of grammar and meaning. This is supported by Qi and Wang's (2020) study, which, based on the viewpoint of a usage-based approach, explored how input frequency and semantic feature affect language acquisition device and showed how with the increasing contact with specific language structure, learners gradually extract language use rules from these language constructs and establish the mapping relationship between structure and meaning in the brain. Therefore, frequency is the fundamental mechanism of language acquisition.

However, in this regard, the acquisition of artificial words and English words is differential. With the increase of frequency, the acquisition rate of artificial words increases by and large, but the increase is first fast and then slow, and the difference at frequency 7 and frequency 10 is not obvious. This shows that with grammar and meaning of artificial words, the function of frequency is obvious in the first stage, but with the increase in frequency, the growth of acquisition slows down. We can deduce that for grammar and meaning acquisition of artificial words, the plateau phenomenon of exposure frequency might exist, demonstrating that after a certain exposure frequency is reached, new acquisition becomes difficult.

For the acquisition of grammar and meaning of English words, the function of frequency fluctuates, and such fluctuations might still exist in the relationship between subsequent frequency increase and lexical acquisition rate, which differs from the frequency effect on that of artificial words. This might suggest that the frequency effect on grammar and meaning acquisition of vocabulary in real languages is more inclined to vary due to the influences of implicit culture and meaning, and for artificial languages, with no possibilities of being embedded with cultural elements, the effect of frequency tends to increase linearly. The analysis of RQ 2 proves that language could not be separated from its culture and environment. All the factors could explicitly or implicitly influence the acquisition of language, thus make the results irregular.

> *RQ3. Does the interaction of factors such as learner type, language type, frequency and part of speech influence the acquisition of grammar and meaning of vocabulary?*

From the results of the multivariate analysis of Chinese and foreign learners' acquisition of grammar and meaning of words, the frequency and part of speech are seen to be important factors that cause significant differences in vocabulary acquisition. This suggests that frequency is an important factor in promoting the acquisition of artificial words and English in general. As Larsen-Freeman (1976) indicated, frequency may be the only important factor that leads to acquisition change. It seems that in any language, the identification of parts of speech varies largely, but the common thing is that performance with adjectives and nouns is generally better than with verbs, and at the same time, it relates to the salience of words of different parts of speech and memorability of such words.

Regarding the combined effect of various factors on vocabulary acquisition, significance exists in four types of interactions. From the effects of the interaction of different factors, we find that language types, frequency and part of speech of a word are the three factors that can usually combine to cause significant differences in the acquisition of grammar and meaning of words, which indicates that to obtain good results in the acquisition of grammar and meaning, in addition to some single factors, such as frequency and part of speech, the combination of these factors can also play a very important role. This finding further supports Zhang and Fang's (2020) study on frequency effect on collocation, which showed that language proficiency, constituent word frequency, lexical frequency are all factors influencing the acquisition of collocation. Second language learners need to infinitely approach the overall processing of collocations in native language to acquire the collocations. Frequency effect is a gradual transition and evolution from dependence on rules to overall synthesis as the input frequency increases and NNSs' language proficiency improves.

The choice of the four factors of learner type, language type, frequency and part of speech as the variables for interaction analysis reflect our assumption that they might well represent the key elements in language acquisition: learner, language (embedded in culture), language use, and language system. The interaction of these factors is supposed to well present how language is processed

in the real context. This in fact affirms the Inseparability Principle by Atkinson (2010) that mind, body, and world work together in SLA.

> *RQ4 Does the acquisition of grammar and meaning of words vary in accordance to the difference of language type?*

As **Table 4** shows, in general, the acquisition of grammar and meaning of artificial words is better than that of English words. However, the trend is not regular, for which one level of frequency effect, such as frequency 1, might be better for artificial words, another level of frequency effect, such as frequency 3, is better for English words. The insignificance of the paired-*T* test of the two types of languages also indicates that the type difference of language does not cause the acquisition difference. From the analysis above, we might conclude that the type of language, whether artificial or real, is not the factor that can greatly affect the acquisition of grammar and meaning of a language. In other words, despite the different types of languages, learners tend to process their grammar and meaning in more or less the same way, which shows that human beings share more commonalities than differences in the use of languages in their language mechanism. These findings are very similar to the results of Perez-Paredes and Bueno-Alastuey's (2019) study, revealing that although learners of different native languages vary in frequency in using certainty adverbs, NSs, and NNSs share more similarities than differences in language use. These findings also provide more evidences that exposure frequency and usage-based approaches have strong explanatory power for SLA (Bybee, 2006; Ellis and Larsen-Freeman, 2009; Ambridge et al., 2015; Patterson, 2021).

What we need to attend is that in this study, the language types refer to the two types of languages (artificial and real) under the same category of alphabetic writing, while in the study by Chen et al. (2020), the language types refer to a more general category of hieroglyphic Chinese and alphabetic Keki artificial language. Despite the differences in languages at micro-level or macro-level, the results of these two experiments show that the frequency effect can help to transcend language barriers to make language learners reach a similarly high level of acquisition of a word as native speakers.

## Implications

This study has the following pedagogical implications: (1) Exposure frequency is important for their SLA, regardless of learners' native language. Therefore, teachers should take into account exposure frequency of vocabulary when designing teaching materials or creating teaching tasks; (2) Context embedded with the target language culture should be created to raise learners' cultural awareness of target language and facilitate learners' acquisition vocabulary; (3) Attention should be paid to the factors of language learners, their native languages, input frequency and grammar of words, as these factors may interact with each other to affect learners' acquisition of second language vocabulary; (4) Different parts of speech of a word may be processed differently and take different amounts of time for acquisition. Therefore, teachers are suggested to attend to these differences and design appropriate tasks for the acquisition of different words.

## Limitations and further research recommendations

Limitations of this study should also be noted: (1) The sample size is relatively small which might affect the generalizability of the study. (2) We are mainly concerned with how the overall differences between the hieroglyphic writing and alphabetic writing and their influences toward the acquisition of meaning and grammar of words, assuming that alphabetic writing, whatever its native language is, will have the same influence on SLA. This assumption may ignore the individual traits of different alphabetic languages, which may vary in its effect during the process of SLA; (3) Embodied theory was quoted as an important theoretical basis in this study and culture was supposed to play an important role in affecting SLA. However, the exploration on how culture exactly works in affecting SLA and how it interacts with exposure frequency for SLA was not examined in this study; (4) This study is restricted to the investigation of vocabulary acquisition in sentence context. Future experiment could be designed for study at discourse level; (5) The measurement of acquisition of meaning and grammar relied on the same method of multiple choice. This might obscure their acquisition differences.

For future research, we suggest that comparative study of vocabulary acquisition be carried out between EFL learners of a specific country of native alphabetic writing language and native EFL Chinese learners. We also encourage researchers to carry out studies of this comparative type on the acquisition context of discourse level. New experiment paradigms of eye-tracking and ERP are recommended for exploring the nuances between the learners of different language types.

# Conclusion

The present study, as one of the few of its kind, sheds light on the frequency effect on grammar and meaning acquisition by learners of different language types. First, despite different language types of learners, regarding acquisition in its initial stage, the more similar the word forms are, the better the acquisition effect would be. After being frequently exposed to certain words, learners of different language types tend to converge in competence for grammar and meaning of the words. Secondly, learners of different language types share more similarities than differences for grammar and meaning acquisition. As second language learning progresses, target culture tends to enhance its effect toward the learners' acquisition increasingly. Thirdly, learner types, language types, part of speech of a word and exposure frequency interact and have combining interaction effect toward the acquisition. Finally, the results of this experimental study suggest that exposure frequency could possibly be the determining factor in the acquisition of grammar and meaning of words. The effect of frequency might transcend language types and has similar functions to the grammar and meaning acquisition of vocabulary of different language types.

## 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 study was approved by the Research Ethics Committee of South China Normal University, China where the first author completed his Ph.D. thesis (which the current manuscript is based on). The patients/participants provided their written informed consent to participate in this study.

## Author contributions

JZ led the research project and contributed to the research design, data collection, data analysis, manuscript drafting, and revising. JH analyzed the data, reviewed, and revised the manuscript. Both authors have approved the submission.

## Funding

This work was supported by: (1) High Level Discipline Cultivation Project of Shenzhen Polytechnic (Project No. 6022311008S) and (2) Foreign Language Program of 2022 Guangdong Philosophy and Social Sciences (Project No. GD22WZX02-10).

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1125483/full#supplementary-material

## References

Aka, N. (2019). Reading performance of Japanese high school learners following a one-year extensive reading program. *Read. Foreign Lang.* 31, 1–18.

Ambridge, B., Kidd, E., Rowland, C. F., and Theakston, A. (2015). The ubiquity of frequency effects in first language acquisition. *J. Child Lang.* 42, 239–273. doi: 10.1017/S030500091400049X

Ament, J., Baron-Pares, J., and Perez-Vidal, C. (2020). A study on the functional uses of textual pragmatic markers by native speakers and English-medium instruction learners. *J. Pragm.* 156, 41–53. doi: 10.1016/j.pragma.2019.09.009

Andringa, S., and Rebuschat, P. (2015). New directions in the study of implicit and explicit learning: An introduction. *Stud. Sec. Lang. Acquisit.* 37, 185–196. doi: 10.1017/S027226311500008X

Atkinson, D. (2002). Toward a sociocognitive approach to second language acquisition. *Mod. Lang. J.* 86, 525–545. doi: 10.1111/1540-4781.00159

Atkinson, D. (2010). Embodied cognition and second language acquisition. *Appl. Ling.* 31, 599–622. doi: 10.1093/applin/amq009

Boden, M. (2006). *Mind as Machine: A History of Cognitive Science*. Oxford: Clarendon Press.

Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, J. (2010). *Language, Usage and Cognition*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511750526

Chen, C., and Truscott, J. (2010). The effect of repetition and L1 lexicalization on incidental vocabulary acquisition. *Appl. Ling.* 31, 693–713. doi: 10.1093/applin/amq031

Chen, C. F., Zhou, R., and Zhao, Z. J. (2020). A comparative study of frequency effect on word form acquisition of alphabetic writing between Chinese learners and Foreign learners of English language. *For. Lang. Res.* 37, 63–68. doi: 10.13978/j.cnki.wyyj.2020.04.010

Chen, J. L. (2005). *Modern Foreign language teaching-theory and method*. Shanghai: Shanghai Foreign Language Education Press.

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*, 4th Edn. Boston, MA: Pearson Education.

Creswell, J. W., and Guetterman, T. C. (2019). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*, 6th Edn. Ann Arbor, MI: ProQuest Ebook Central.

Ellis, N. (1994). *Implicit and Explicit Learning of Languages*. Cambridge, MA: Academic Press.

Ellis, N., and Larsen-Freeman, D. (2006). Language emergence: Implications for applied linguistics- Introduction to the special issue. *Appl. Ling.* 27, 558–589. doi: 10.1093/applin/aml028

Ellis, N., and Schmidt, R. (1998). Rules or associations in the acquisition of morphology? The frequency by regularity interaction in human and PDP learning of morphosyntax. *Lang. Cogn. Process.* 13, 307–336. doi: 10.1080/016909698386546

Ellis, N., Simpson-Vlach, R., and Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Q.* 42, 375–396. doi: 10.1002/j.1545-7249.2008.tb00137.x

Ellis, N. C. (2002a). Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition. *Stud. Sec. Lang. Acquisit.* 24, 143–188. doi: 10.1017/S0272263102002024

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Stud. Sec. Lang. Acquisit.* 24, 297–339. doi: 10.1017/S0272263102002140

Ellis, N. C. (2008). "Usage-based and form-focused language acquisition: The associative learning of constructions, learned-attention, and the limited L2 endstate," in *Handbook of Cognitive*, eds P. Robinson and N. C. Ellis (London: Routledge), 372–405.

Ellis, N. C., and Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Lang. Learn.* 59, 90–125. doi: 10.1111/j.1467-9922.2009.00537.x

Ellis, N. C., and Ogden, D. C. (2015). Language cognition: Comments on Ambridge, Kidd, Rowland, and Theakston "the ubiquity of frequency effects in first language acquisition". *J. Child Lang.* 42, 282–286. doi: 10.1017/S0305000914000646

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., and Parisi, D. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Grammer, K., and Randy, T. (1994). Human (Homo sapiens) facial attractiveness and sexual selection: The role of symmetry and averageness. *J. Comp. Psychol.* 108, 233–242. doi: 10.1037/0735-7036.108.3.233

Harrington, M., and Denis, S. (2002). Input driven language learning. *Stud. Sec. Lang. Acquisit.* 24, 261–268. doi: 10.1017/S0272263102002103

Horst, M., Cobb, T., and Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Read. For. Lang.* 11, 207–223.

Lakoff, G., and Johnson, M. (1999). *Philosophy in the Flesh:The Embodied Mind and Its Challenges to Western Thought*. New York, NY: Basic Books.

Langacker, R. W. (1987). *Foundations of cognitive grammar.Vol.1 Theoretical prerequisites*. Redwood City, CA: Stanford University Press.

Larsen-Freeman, D. (1997). Chaos/Complexity/Science and second language acquisition. *Appl. Ling.* 18, 141–165. doi: 10.1093/applin/18.2.141

Larsen-Freeman, D. E. (1976). An explanation for the morpheme acquisition order of second language learners. *Lang. Learn*. 26, 125–134. doi: 10.1111/j.1467-1770.1976.tb00264.x

Larsen-Freeman, D. (2007). Reflecting on the cognitive-social debate in second language acquisition. *Mod. Lang. J.* 91, 773–787. doi: 10.1111/j.1540-4781.2007.00668.x

Leech, G. (2011). Why frequency can no longer be ignored in ELT? *For. Lang. Teach. Res.* 43, 3–20.

Liao, Y., and Zhang, Q. (2012). The influence of affective congruency between target and context during encoding on memory retrieval. *J. Psychol. Sci.* 5, 1083–1090. doi: 10.16719/j.cnki.1671-6981.2012.05.018

MacWhinney, B. (1999). *The Emergence of Language*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

McCandliss, B. D., Possner, M. I., and Givón, T. (1997). Brain plasticity in learning visual words. *Cogn. Psychol.* 33, 88–110. doi: 10.1006/cogp.1997.0661

Patterson, A. (2021). Predicting second language listening functor comprehension probability with usage-based and embodiment approaches. *Int. J. Bilingual.* 25, 772–788. doi: 10.1177/13670069211000851

Perez-Paredes, P., and Bueno-Alastuey, M. C. (2019). A corpus-driven analysis of certainty stance adverbs: Obviously, really and actually in spoken native and learner English. *J. Pragm.* 140, 22–32. doi: 10.1016/j.pragma.2018.11.016

Qi, J., and Wang, H. (2020). The influence of input frequency and semantics on Chinese English Learners' acquisition of dative structure based on mixed effect model. *For. Lang. Educ.* 2, 76–80+113. doi: 10.16362/j.cnki.cn61-1023/h.2020.02.014

Rott, S. (1999). The effect of exposure frequency on intermediate language learner's incidental vocabulary acquisition through reading. *Stud. Sec. Lang. Acquisit.* 21, 589–619. doi: 10.1017/S0272263199004039

Rumelhart, D., and Melland, J. M. (1986). "On learning the past tense of English verb," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol.2: Psychological and Biological Models*, eds J. M. Clelland and D. Rumelhart (Cambridge, MA: MIT Press).

Rumelhart, D., and Melland, J. M. (1987). "Learning the past tense of English verbs: Implicit rules or parallel processing?," in *Mechanism of Language Acquisition*, ed. B. M. Whinney (Mahwah, NJ: Lawrence Erlbaum Associates).

Saragi, T., Nation, P., and Meister, F. (1978). Vocabulary learning and reading. *System* 6, 72–78. doi: 10.1016/0346-251X(78)90027-1

Schwartz, M., and Causarano, P. L. (2007). The role of frequency in SLA: An analysis of gerunds and infinitives in ESL written discourse. *Arizona Work. Pap. SLA Teach.* 14, 43–57.

Song, J., and Sardegna, V. (2014). EFL learners' incidental acquisition of English prepositions through enhanced extensive reading instruction. *RELC J.* 45, 67–84. doi: 10.1177/0033688214522623

Sun, H. M. (2014). The effects of exposure frequency and contextual richness in reading on Chinese EFL learners' vocabulary acquisition. *Chin. J. Appl. Ling.* 37, 86–129. doi: 10.1515/cjal-2014-0006

Temperley, D. (2007). *Music and Probability*. Cambridge, MA: MIT Press. doi: 10.7551/mitpress/4807.001.0001

Tyler, A. (2010). Usage-based approaches to language and their applicationss to second language learning. *Annu. Rev. Appl. Ling.* 30, 270–291. doi: 10.1017/S0267190510000140

Verspoor, M., de Bot, K., and Lowie, W. (2008). "Input and second language development from a dynamic systems perspective," in *Input Matters*, eds T. Piske and M. Young-Scholter (Bristol: Multilingual Matters), 62–81. doi: 10.21832/9781847691118-006

Wang, C. M. (2011). Usage-based view on language acquisition. *For. Lang. China* 8:1. doi: 10.13564/j.cnki.issn.1672-9382.2011.05.003

Wang, Y. (2008). The fourth paper on the new growth of linguistics: exploration on philosophy of language, post philosophy of language and embodied philosophy. *For. Lang. Res.* 4, 2–10. doi: 10.16263/j.cnki.23-1071/h.2008.04.005

Wang, Y. (2012). Linguistics frontier in the view of postmodern philosophy: Embodied humanism and cognitive linguistic. *J. For. Lang.* 11, 17–26.

Waring, I., and Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader. *Read. For. Lang.* 15, 130–163.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Appl. Ling.* 28, 46–65. doi: 10.1093/applin/aml048

Yang, Y. Q., and Zhang, H. (2021). Study on the role of verb types in the delay of principle B effect in second language acquisition. *For. Lang. Res.* 4, 56–63. doi: 10.16263/j.cnki.23-1071/h.2021.04.009

Zhang, P., and Fang, N. (2020). Effects of lexical frequency, semantic congruency, and language proficiency on English collocational processing by Chinese EFL learners. *For. Lang. Teach. Res.* 4, 532–640. doi: 10.19923/j.cnki.fltr.2020.04.005

Zhang, S., and Zhang, H. (2022). Developmental features of the receptive-productive continuum of L2 academic vocabulary. *Mod. For. Lang.* 5, 1–13.

Zhang, X., and Qi, L. X. (2009). Incidental vocabulary acquisition in natural reading. *For. Lang. Teach. Res.* 4, 303–321.

Zhang, Y. (2020). Effects of frequency, semantic transparency on the processing of formulaic sequences by Chinese native speakers and intermediate and advanced learners. *J. PLA Univ. For. Lang.* 43, 35–43.

Zhang, Z. L., and Xing, Q. (2012). Mood congruent effect in different negative emotion. *Stud. Psychol. Behav.* 5, 373–377.

Zhao, J. Z. (2017). Studies on frequency effect in language learning-perspectives of psychology and linguistics. *For. Lang. Teach. Res. Basic Educ.* 5, 15–21.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 17 August 2023
DOI 10.3389/fpsyg.2023.1126106



OPEN ACCESS

EDITED BY
Hamdollah Ravand,
Vali-E-Asr University of Rafsanjan, Iran

REVIEWED BY
Rajab Esfandiari,
Imam Khomeini International University, Iran
Dayu Jiang,
Wuhan University, China
Peida Zhan,
Zhejiang Normal University, China

*CORRESPONDENCE
Yaru Meng
✉ maryann@mail.xjtu.edu.cn

RECEIVED 17 December 2022
ACCEPTED 28 July 2023
PUBLISHED 17 August 2023

CITATION
Meng Y, Wang Y and Zhao N (2023) Cognitive diagnostic assessment of EFL learners' listening barriers through incorrect responses.
*Front. Psychol.* 14:1126106.
doi: 10.3389/fpsyg.2023.1126106



# Cognitive diagnostic assessment of EFL learners' listening barriers through incorrect responses

Yaru Meng[1]*, Ya Wang[1] and Ningning Zhao[2]

[1]School of Foreign Studies, Xi'an Jiaotong University, Xi'an, Shaanxi, China, [2]Xi'an Chanba No.2 Middle School, Xi'an, Shaanxi, China

English as Foreign Language (EFL) learners' cognitive processes have been a research focus in listening assessment. Most studies use correct responses as data, but undervalue the rich information of the incorrect answers or options (in the case of multiple choice questions, MCQ). However, the MCQ distractors are often intentionally designed to reveal learners' problems or barriers. In order to diagnose the EFL learners' listening barriers through incorrect responses, Cognitive Diagnostic Models (CDMs) for bugs were adopted, hence the name Bug-CDMs. First, five EFL listening barrier attributes were identified and two Bug Q-matrices were developed to comparatively analyze the learner's responses with different Bug-CDMs. The results revealed that Bug-GDINA was the optimal model, and the most prevalent barriers were semantic understanding and vocabulary recognition. These barriers confirmed both compensatory and non-compensatory relationships in causing listening comprehension failures. The study proved the feasibility of Bug-GDINA in diagnosing listening barriers from the incorrect responses. Limitations and suggestions for further research were also proposed.

KEYWORDS

**cognitive diagnostic assessment, EFL listening barriers, incorrect responses, cognitive diagnostic model, Bug-GDINA**

## 1. Introduction

Though listening is a major language skill (Vandergrift and Goh, 2012; Rost, 2016), it is still the thinly profiled, least understood and particularly under-researched skill (Vandergrift and Goh, 2012; Wolf et al., 2019; He et al., 2022), and the learners' underlying cognitive process is even less addressed (Buck, 2001; Harding et al., 2015). The same is true of Chinese college EFL learners (Xu and Nie, 2016), as a result, understanding their cognitive processes has been a focus in EFL listening research.

Cognitive Diagnostic Assessment (CDA), as a new generation of measurement, has demonstrated its advantages in combining cognitive process and psychometrics (Leighton and Gierl, 2007). It can not only reveal learners' cognitive processes and mastery or non-mastery of subskills intended by test items (Sawaki et al., 2009), but CDA also helps provide tailored feedback for subsequent remediation and guidance (Yi, 2017; He et al., 2022). This can be achieved by adopting appropriate cognitive diagnostic models (CDMs), which are latent class models used for classifying students based on their skill profiles. Current CDA studies in L2 listening focus on identifying whether a particular learner has mastered some specific language subskills (known as attributes) such as "listening for details" or "listening for main ideas" based on the correct responses or options (in the case of multiple choice questions, MCQ hereafter)

(e.g., Min and Xiong, 2019; He et al., 2022), but they undervalue the rich information of students' incorrect options or distractors which are often intentionally designed and likely to be chosen to reveal specific problems in knowledge or skills (Ozaki et al., 2019), and can also provide useful diagnostic feedback (Stout et al., 2022). Therefore, both incorrect and correct listening responses (or MCQ distractors) deserve equal attention and treatment.

Goh (2000) used a general term "listening problem" to refer to the real-time processing difficulties directly related to cognitive procedures at various stages of comprehension. From cognitive diagnostic perspective in psychometrics, "Bug" is used to refer to misconceptions, lack of skills (Kuo et al., 2016), or the thinking processes interfering with learning (Bradshaw and Templin, 2014). The current study adopts MCQ incorrect responses as data and bug-CDMs as analytical framework to identify the root causes of failures. The distractors often reflect both common problems and misconceptions (Haladyna and Rodriguez, 2013; Jones, 2020), therefore, "listening barriers" (Dunkel, 1991) may be better for the internal and external characteristics to hamper L2 comprehension, though sometimes "problems" will also be used interchangeably.

Up to now, very few attempts were made employing CDA to analyze incorrect responses in L2. One possible challenge might be the lack of appropriate CDMs to accurately identify students' barriers at sufficiently fine-grained levels (Lee, 2015). Fortunately, Bug-CDMs like Bug-DINO (the bug deterministic inputs, noisy "or" gate mode) (Kuo et al., 2016) were developed, which inspires us to make attempts to analyze the incorrect listening responses. The current study is an early attempt to cognitively diagnose the root causes using Bug-CDMs. The findings would contribute to remedial learning and teaching, test development, and particularly to the design of high-quality MCQ distractors. In the sense, the study is significant methodologically, pedagogically and psychometrically.

# 2. Literature review

## 2.1. EFL listening barriers and the cognitive processes

Listening comprehension is arguably the most complex and multi-dimensional cognitive process (Buck, 2001; Vandergrift and Goh, 2012; Field, 2013), involving both linguistic and non-linguistic competence (Bachman and Palmer, 1996; Buck, 2001). Vandergrift (2007) pointed out that tracking L2 listening barriers is as important as required skills. They can help infer learners' problematic cognitive processes as well as their interactions. The most systematic study of ESL listening processing problems is by Goh (2000) on Chinese learners in Singapore based on Anderson (1995) three-phase model of perception, parsing and utilization. Perception is the physiological process of receiving auditory signals for processing; parsing is the mapping of the perceived input onto the information from long-term memory, corresponding to bottom-up processing; utilization is drawing on listeners' world knowledge to fill in gaps in their mental representation of the message, analogous to top-down comprehension. Goh's 10 problems are derived from the learners' self-reports, weekly diaries and interviews. The 5 most common problems she identifies are: (1) quickly forgetting what is heard (parsing); (2) not recognizing words or unfamiliar alternative pronunciations of words they know (perception); (3) understanding words but not their intended message (utilization); (4) neglecting the next part when thinking about meaning (perception); and (5) inability to form a mental representation (parsing). Since four of them are in the perception and parsing stage, it may indicate that the fluent listening comprehension is often obstructed by limited lexical and syntactic knowledge (Juan and Abidin, 2013; Vafaee and Suzuki, 2020; Yeldham, 2022). This is further confirmed by Cao et al. (2016) two new types of problems as "Confused about unexpected word appearance" and "Unsure about the meaning of words." Goh (2000) study made an enormous contribution by establishing the correlation between the three cognitive stages and the learners' reported difficulties, moving listening barrier research forward. However, there are different versions of this correlation, which has since become an issue (Zhang et al., 2010; Feng and Xu, 2022), hence needing more fine-grained models for this purpose. Since then, different listening comprehension models have been adopted to facilitate a better understanding of the listening process, including listening barriers. These models include Vandergrift and Goh (2012) 3 process model, Field (2013) "two level" model and Rost (2016) four category model. Rukthong and Brunfaut (2019) integrate the above models into 5 sub-skills of acoustic-phonetic decoding, word recognition, parsing, semantic processing, and pragmatic processing, with the former three stages as lower-level processes and the latter two as higher-level processes. This integrated model incorporates top-down and bottom-up, and is also more able to demonstrate the complex and interactive nature of both high-level and low-level cognitive processes involved in L2 listening. Therefore, it may shed light on learners' barrier identification.

## 2.2. Methods in L2 listening barriers research

To investigate learners' listening comprehension problems, various procedures and empirical methods have been used, including listening diaries and interviews (early attempts like Goh, 2000), think-aloud protocols (Hwang, 2004), and questionnaires (Zhang and Zhang, 2011; Noroozi et al., 2014). For example, Hwang (2004) studied listening problems with high school students by asking them to verbally describe what they heard, and then reflected on and wrote down any barriers they encountered in listening. These retrospection techniques provide opportunities for listeners to recall the listening experience and offer insights into their listening processes at a later moment in time, though the reliability of their results is limited by the reliability of the measurements or, in the case of the recall protocols, inter-rater reliability. Listening barrier analysis based on test scores tells us something about the outcome based on correct answer, i.e., the level of listening success, and may verify comprehension, but it reveals very little about how students arrived at comprehension or, more importantly, how comprehension failed. In view of this, many scholars have adopted the analysis of wrong responses to diagnose learners' lack of skills and help overcome their difficulties. This resembles error analysis, while in assessing L2 listening problems, error analysis is most often focused on dictation tasks (Cho, 2021). Kao and Kuo (2021) used 12 MCQ items from TOEIC Bridge test and asked learners' to self-report on their incorrect listening items for diagnosing the real sources of their problems in EFL contexts. They also used Goh (2000) top five

problems for mediation moves, but their insufficient psychometric-based validity evidence may have caused inaccurate diagnoses. On the other hand, this use of wrong options in MCQ items for future mediation inspired further attempts.

The diagnosing power of MCQs, and especially the messages in their options, was pointed out by Briggs et al. (2006). He demonstrated that multiple-choice items based on construct can provide diagnostic information. Andrade and Heritage (2017) made one more step forward by saying that "When each of the possible answer choices in an item is linked to student understanding, an item-level analysis of student responses can reveal what individual students know" (P.46). It's true that the wrong choices are not simply wrong. Each is wrong in a way that reflects a common gap or misconception. There is some evidence that diagnostic, multiple-choice items are actually better than open-ended items at eliciting students' true understanding, perhaps because the items probe students' thinking by offering plausible incorrect answers (Steedle and Shavelson, 2009). The incorrect answers that demonstrate incomplete understanding, errors in reasoning, or misconceptions are useful to teachers, who can use them to identify next steps in instruction (Andrade and Heritage, 2017). But precisely how the incorrect options in listening tests reflect the learners' cognitive processes and barriers is an area seldom ventured into. Cognitive Diagnostic Assessment, as one recent psychometric development, prides itself on combining qualitative and quantitative evidence for fine-grained validation and feedback, thus giving it potential for tapping into the diagnostic information from incorrect options.

## 2.3. Cognitive diagnostic assessment for listening barriers

Compared with the traditional tests that simply rank-order examinees on a one-dimensional continuum, Cognitive Diagnostic Assessment (CDA) can comprehensively investigate the multidimensional cognitive processes, thus inferring the non-observable knowledge state of an individual based on the observable response data (Leighton and Gierl, 2007; Rupp et al., 2010). It can explore learners' differences in internal cognitive processes or knowledge structures so as to offer pedagogically useful information for subsequent learning and teaching remedies (Jang et al., 2015; Chen and Chen, 2021). Under the framework of CDA, three major iterative procedures should be undertaken in order to obtain diagnostic results about learners' specific abilities, namely, cognitive attribute identification, Q-matrix construction, and data analysis. Cognitive attributes refer to the cognitive skills, strategies, and knowledge that learners might need to correctly complete a given task. They are often associated with test items and their relationships are represented in a two-dimensional incidence called a Q-matrix, which is expressed with a "0" or a "1," indicating an item not requiring or requiring an attribute (Rupp et al., 2010). In conjunction with learners' item response data, the Q-matrix is then inputted into certain psychometric models called cognitive diagnostic models (CDMs) for data analysis during which model-data fit statistics are evaluated. If a mismatch is identified, one should revise the primary attributes and the Q-matrix until an appropriate fit is achieved. With all those procedures completed, the diagnostic results showing students' mastery profiles can be obtained both at group and individual levels. Due to CDA' s great potential for discovering learners' underlying performance, it has been widely used in the field of language assessment.

Most CDA-based listening studies explore what attributes can best represent L2 listening ability (Sawaki et al., 2009; Meng, 2013) and the underlying inter-attribute relationships (Yi, 2017; Min and Xiong, 2019; Dong et al., 2021). Therefore, the target attributes of most studies are listening sub-skills. No wonder Harding et al. (2015) worried that the regular CDA could not help find the root causes of students' barriers. To address this, careful attention should be given to incorrect options, to exploit diagnostic information and pinpoint students' cognitive problems. Kuo et al. (2016, 2018) employed Bug-CDMs to diagnose students' cognitive problems based on their incorrect responses in mathematical multiple-choice items, proving the feasibility of using Bug-CDMs for problematic cognitive diagnosis, and thus inspiring us to consider their application in the field of language testing. In other words, Bug-CDMs may have the potential to assist in diagnosing listening comprehension problems.

Similar to the basic assumptions of the normal GDINA models, all bug-related models are realized through the modification of the latent variables into "bugs," and can then be used to analyze the bug data. But the non-saturated models like Bug-DINA or Bug-DINO include only specific parameters, hence with limited generality. For example, with Bug-DINO model, a learner cannot get a correct answer if he possesses one or more barriers the item measures (Kuo et al., 2016, 2018); with the Bug-DINA model, if and only if a learner has all the barriers the item measures, will he be more likely to get the answer wrong. In contrast, Bug-GDINA, parallel to the GDINA model which is commonly used in CDM studies and accommodates all the possible major and interactive effects between the attributes (Rupp et al., 2010), has the same advantages and can also realize greater generality. But Bug-GDINA is also often penalized for violating the principle of parsimony for model application with its inclusiveness.

## 2.4. Research questions

Bug-CDMs' development and application are in their infancy even in psychometrics, and their potential in solving EFL learning problems has not yet been explored. So the current study attempts to tap the feasibility of Bug-CDMs in diagnosing listening processing problems through the incorrect options. To accomplish this purpose, the following two research questions are addressed:

RQ1: To what extent can cognitive diagnostic assessment identify EFL listening barriers?

RQ2: To what extent can diagnostic results help reveal the relationships between listening barrier attributes?

# 3. Methods

## 3.1. Participants

A total of 1,121 EFL college students (approximately 17–19 years old) with intermediate language proficiency (National Matriculation English Test, roughly equivalent to CEFR B1 level) participated in the listening test. They were all freshmen with an average of over 6 years of prior EFL learning experience, and were recruited from six universities at different levels (2 top-tier, 3 s-tier and 1 third-tier) in

western China. Six and twelve test takers, respectively, volunteered to participate in the first and second verbal protocol sessions, with the former targeted at modifying the identified attributes, and the latter aiming to validate the Q-matrix. In addition, eight content experts were invited to code the incorrect options as barrier attributes. These experts not only had considerable experience in language test development and teaching, but were also familiar with cognitive diagnostic approaches.

## 3.2. Instruments

The diagnostic listening tests were selected from the item pool of PELDiaG system (Personalized English Learning Diagnosis and Guidance system) designed for diagnostic purposes (Meng, 2013; Dong et al., 2021; Meng and Fu, 2023) and the CET 4 test. The Cronbach's alpha of the test is 0.724, and the KMO is 0.848 ($p < 0.001$), indicating good reliability. The item selection process involved two steps: first, IRT (Item response theory) analysis was conducted to obtain the item parameters. Items with good discrimination ($a > 0.3$) and difficulty ($3.0 \geq b \geq -3.0$) were chosen. Afterwards, the quality of incorrect options was analyzed through the response frequency, which means that the distractors with low-frequency (<5%) were screened out (Haladyna and Downing, 1993). In the end, a total of 19 multiple-choice items were selected. There were three sections with 11 short dialogue items, 1 long conversation, and 2 passages. The topics covered familiar ones like shopping, education, public transportation and technology. The tests were administered in paper-and-pencil format at regular class times, and the responses were scored dichotomously.

Five Bug-CDM models were compared for the optimal model-data fit, including Bug-DINO, Bug-DINA, Bug-GDINA, Bug-RRUM, and Bug-ACDM. The comparison was done using the "GDINA" package (version 2.9.3) (Ma and de la Torre, 2020) embedded in R studio, which provides a platform for a series of cognitively diagnostic analyses for dichotomous and polytomous responses.

## 3.3. Procedures

The procedures in this study bore a strong resemblance to those in diagnosing listening skills which included five major stages: identifying the attributes, constructing the Q-matrix, validating the Q-matrix, selecting optimal CDM and generating feedback.

### 3.3.1. Identifying the barrier attributes

First, the researchers proposed a preliminary list of common EFL listening barriers based on listening cognitive processing models and literature review. Then, 12 students were recruited to participate in verbal reports for their encountered barriers in finishing the target listening items, according to which the authors' barrier attribute names and definitions were derived. For example, the fifth attribute B5 "barrier in making pragmatic inferences" was modified because some students reported the overuse of prior knowledge. The following excerpt from Participant 1, who answered the item (Item 16) in a passage incorrectly, might help to illustrate.

Item 16: What can be inferred from the passage?

A. **Helping others brings positive emotions. (Correct Answer)**
B. Volunteering benefits the receivers more.
C. Everyone needs help and friends.
D. Helping others means power.

*Transcript: … Volunteer is to help, and when we help others, we satisfy our needs to have a degree of control over our world. When people see smiles and satisfaction in person being helped, they can feel happier in their hearts…*

> Participant 1: *"In fact, I haven't grasped all information the speaker mentioned here, but I believe if the volunteer service is provided, many people (receivers) can get benefits from it. So I decide to choose option B."*

In this case, the barrier impeding participant 1 was not the lack of sufficient prior knowledge, as this barrier is often interpreted. Rather, he over-used his prior knowledge about the volunteer service, resulting in an over-generalization and his neglect of key information. In other words, since he had not fully understood the original information, he was prone to overusing background knowledge. This attribute was accordingly redefined as "barrier due to the lack of or over-generalization of prior knowledge." Students' verbal data also revealed that sometimes the incorrect answer may not result from a single barrier, but the interactive effect of two or more barriers. For example, when participant 1 was trying to solve item 3, he not only overgeneralized prior knowledge (B5), but also missed the explicitly expressed information (B4). In short, with verbal data analysis, the barrier attributes were modified and validated. See Table 1 for the final version.

### 3.3.2. Constructing the preliminary Q-matrix

The development of the Q-matrix was informed by two sources of information: the final version of barrier attributes and expert judgment. First, eight experts were invited to participate a Q-matrix coding training workshop on how to analyze the MCQ distractors for barriers. Then they individually decided whether a certain barrier (or barriers) was involved in each incorrect option for an item, based on the examination of the above barrier definitions. According to the attribute coding results, if more than half of experts reached an agreement on a certain barrier, then it was identified as a listening barrier for the item. Otherwise, the barrier was rejected. Hence the preliminary Q-matrix (Q1) was developed (see items for modification on Table 2).

### 3.3.3. Validating the Q-matrix

In order to modify the item-barrier mapping, two steps were taken: verbal report 2 and the data-driven method. In terms of the former, an example is provided here to illustrate the process.

Item 11: What can be inferred about Phillip?

A. He'll go to the party with the woman.
B. He will not meet the man at the party.
C. **He has changed his plans. (Correct Answer)**
D. He has to work late.

*Transcript:*

*(Woman): I talked to Phillip today, and he said he'll be come to the party.*

*(Man): So he CAN come after all.*

Originally, the expert-coded barrier was "understanding semantic meanings" (B4), which meant the listener probably chose the wrong

TABLE 1 Modified barriers.

| Cognitive models | Related research | Verbal reports excerpts | Barrier attributes | Definition of barrier attributes |
|---|---|---|---|---|
| Acoustic- phonetic decoding | Goh (2000), Graham (2006), Namaziandost et al. (2019), and Nushi and Orouji (2020) | "I clearly heard 'camp.'" (But the mentioned word is "cab") | B1: Identifying speech | Unable to identify the auditory, phonetic, and phonological features. |
| Word recognition | Goh (2000), Juan and Abidin (2013), Cao et al. (2016), Namaziandost et al. (2019), Vafaee and Suzuki (2020), and Alharbi and Al-Ahdal (2022) | "… I did not catch most of the words in the text…." | B2: Recognizing vocabulary | Unable to identify words or phrases in a speech stream or activate the relevant word knowledge. |
| Parsing | Goh (2000), Cao et al. (2016), and Vafaee and Suzuki (2020) | "…I got all the words in the sentence, but I still couldn't understand the meaning of the sentence." | B3: Understanding syntactic or semantic structures | Unable to understand the syntactic or semantic structures of the language to generate the local representations of text (clause level). |
| Semantic processing | Goh (2000), Graham (2006), and Cao et al. (2016) | "… I just follow the audio to understand the specific contents, but ignore to connect with the information speakers have mentioned before, so it's hard to get the general idea of the paragraph." | B4: Understanding semantic meanings | Unable to identify or synthesize explicit information at multiple locations to generate the coherent representations of text (discourse level). |
| Pragmatic processing | Goh (2000), Graham (2006), Juan and Abidin (2013), Cao et al. (2016), and Alharbi and Al-Ahdal (2022) | "(I believe) If the volunteer service is provided, many people will get benefits from it." This student overextended his prior knowledge about volunteering instead of referring to text's original contents. | B5: Making pragmatic inferences | Unable to infer implicit contents due to the lack of or over-generalization of prior knowledge, and misunderstanding texts' linguistic information and communicative contexts. |

TABLE 2 The final Q-matrix.

| Item | Ob1 | Ob2 | Ob3 | Ob4 | Ob5 | Items | Ob1 | Ob2 | Ob3 | Ob4 | Ob5 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Item1 | 1 | 0 | 0 | 0 | 0 | Item11 | **1** | 0 | 0 | 1 | 0 |
| Item2 | 1 | 1 | 0 | 0 | 0 | Item12 | 0 | 0 | 0 | 1 | 1 |
| Item3 | 0 | 1 | 0 | 1 | 0 | Item13 | 0 | 1 | 0 | **1** | 1 |
| Item4 | 1 | 1 | 0 | 1 | 0 | Item14 | 0 | 0 | 0 | 0 | 1 |
| Item5 | 0 | 0 | 1 | 0 | 0 | Item15 | 0 | 0 | **1** | 1 | 1 |
| Item6 | 0 | 0 | 1 | 1 | 1 | Item16 | 0 | 0 | 0 | 1 | 1 |
| Item7 | 0 | 0 | 1 | 1 | 0 | Item17 | 0 | 0 | 1 | 1 | 1 |
| Item8 | 0 | 0 | 1 | 1 | 0 | Item18 | 0 | 0 | 0 | 1 | 1 |
| Item9 | 0 | 0 | 0 | 1 | 0 | Item19 | 0 | 0 | 0 | 1 | 1 |
| Item10 | 0 | 0 | 0 | 0 | 1 | | | | | | |

1 indicates the corresponding barrier is measured by a particular item. 0 indicates otherwise. Bold font numbers indicate revisions from the preliminary Q-matrix.

answer when he could not understand or synthesize the speakers' intention. Nevertheless, further evidence was collected through interviewing participants. For example, participant 2 stated:

> *"… the woman said Phillip will come to the party, and then the man probably agree with her because he said Philip can come. But I'm not very sure about this … it's a little bit like he can't come. In fact, I didn't catch clearly what the man said is* ***can or can't****."*

According to this participant, one of the barriers he encountered was that he could not distinguish the phonetic features between "can" and "cannot," and therefore, we added the attribute B1 in our preliminary Q-matrix 1.

To empirically revise and validate the Q-matrix, the data-driven method was employed using the GDINA package. Based on the initial analysis, suggestions for revision of the Q-matrix were put forward. For example, it was suggested that B4 "barrier in understanding semantic

TABLE 3 Model fit comparison of Bug-CDMs using the final Q-matrix.

| Bug-CDMs | Npars | Relative fit | | | Absolute fit | |
|---|---|---|---|---|---|---|
| | | −2LL | AIC | BIC | Max zr | Max zl |
| Bug-DINO | 69 | 26126.38 | 26264.37 | 26610.83 | 4.1628 | 4.2254 |
| Bug-DINA | 69 | 26126.5 | 26264.49 | 26610.95 | 4.1191 | 4.2199 |
| Bug-GDINA | 133 | **25822.3** | **26088.3** | 26756.1 | **3.3777** | **3.4044** |
| Bug-RRUM | 91 | 25941.44 | 26123.44 | **26580.36** | 4.0201 | 3.9800 |
| Bug-ACDM | 91 | 25945.94 | 26127.95 | 26584.86 | 3.5097 | 3.5591 |

Critical z-score = 3.649, 4.044 for $\alpha = 0.05$, 0.01, respectively (with the Bonferroni Correction) (Chen et al., 2013). Bold font indicates optimal fit values.

meanings" be deleted for item 11. Similarly, for items 13 and 15, B3 "barrier in understanding syntactic or semantic structures" was recommended. But considering that psychometric analysis should not be the only sources for Q-matrix revision (Aryadoust and Luo, 2022), the study also included expert judgment. The experts were asked to analyze the items and suggested barriers, and then to reach a consensus after discussion. Final revisions to some of the items were then completed (see Table 2 for Q-matrix 2). Again, taking item 2 as an illustration, when it was not successfully answered, it was more likely that students were unable to identify the phonetic features, and at the same time had difficulties in activating word knowledge.

### 3.3.4. Selecting the optimal CDM

As mentioned above, five Bug-CDMs were employed and compared in order to choose the optimal model. Both relative and absolute fit statistics were obtained (see Table 3). In terms of the former at the test level, the maximum z-scores (denoted as zr) of the residual between the observed and predicted Fisher-transformed correlation of item pairs was produced; in terms of the latter, the residual between the observed and predicted log-odds ratio (LOR) of item pairs (denoted as zl) were produced. It can be seen that the Bug-GDINA has the better model-data fit (Max zr =3.3777, $p = 0.125 > 0.05$; Max zl = 3.4044, $p = 0.113 > 0.05$), followed by the Bug-ACDM (Max zr =3.5097, $p = 0.08 > 0.05$; Max zl = 3.5591, $p = 0.06 > 0.05$). However, BIC for Bug-GDINA was not the lowest, since it generally imposes a penalty on highly parameterized models (Murphy, 2012). But overall, based on the values of absolute fit and relative fit, as well as the saturated model's capacity in identifying complex relationships of listening comprehension, Bug-GDINA proved to be the optimal one for further analysis.

## 4. Results

RQ1: To what extent can cognitive diagnostic assessment identify EFL listening barriers?

At group-level, we can see in Figure 1 the overall barrier profile for students' listening comprehension through "attribute prevalence," which shows learners' mastery probability of each barrier attribute, ranging from 24.79 to 68.19%. As Figure 1 makes clear, "understanding semantic meanings" (B4) is the most prominent barrier impeding learners' listening comprehension, while "identifying speech" (B1) is the least challenging factor. The second most serious problem (63.26%) is a difficulty in "recognizing vocabulary" (B2). The probabilities of encountering barriers in "understanding syntactic or semantic structures" (B3) and "making pragmatic inferences" (B5) were fairly close together (39.46 and 37.38%, respectively), suggesting their similar hindering effects on learners' listening comprehension.

At individual-level, all students were classified into different latent classes. With 5 attributes in the current study, a total of 32 theoretically possible latent classes were identified. The top eight most dominant patterns are presented in Table 4. The five numbers in each latent class correspond to 5 barrier attributes, again with "1" denoting the presence of a certain barrier and "0" otherwise. As is shown, two flat profiles "11111" and "00000" enjoy the top percentages, indicating the possession of all the attributes is the most popular profile, and the second one is with none of these attributes. The remaining six jagged profiles demonstrate learners' strengths and weaknesses (He et al., 2021). When only one barrier is possessed, the latent class profiles were more likely to be 00010 (15.24%) and 01000 (9.04%), as well as a relatively high percentage of 01010 (12.22%), all indicating the prevalence of B2 and B4 in the current sample. Latent class profiles "01110" and "11001" mean that those learners have three barriers, with the class probabilities of 5.13 and 1.67%, respectively. In the same vein, the profile "01111" suggests that learners (11.94%) possess four barriers with the exception of B1, meaning that it is probably the least prevalent barrier.

However, it is worth noting that the 0/1 classification may overgeneralize learners' knowledge states to some degree (Du and Ma, 2021). Considering this, "person parameter estimation" was used to represent to what extent individual learners possessed a certain barrier attribute, especially with the same total scores. This further reveals the personalized differences, as Learner No. 12 and No. 22 illustrate in Figure 2. They exhibit different personal attribute patterns, though

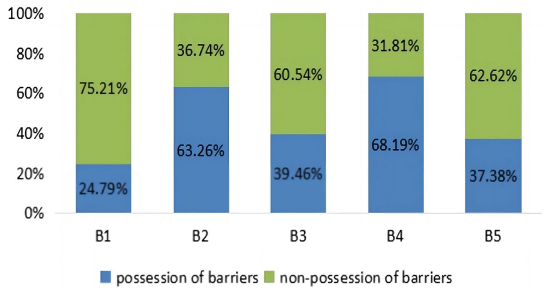



FIGURE 1
Attribute prevalence.

TABLE 4 Eight dominant latent classes and posterior probabilities.

| Latent class | Posterior probability | Latent class | Posterior probability |
|---|---|---|---|
| 11111 | 18.43% | 01111 | 11.94% |
| 00000 | 17.10% | 01000 | 9.04% |
| 00010 | 15.24% | 01110 | 5.13% |
| 01010 | 12.22% | 11001 | 1.67% |

both achieved the same total score of 12. The major barriers for Learner No. 12 were B2 and B4; while for learner No. 22, B2, B3, and B4 were the main issues obstructing listening success.

RQ2: To what extent can diagnostic results help reveal the relationships between listening barrier attributes?

In terms of the inter-relations between the barrier attributes, the tetrachoric correlations between the attributes and the item profiles can be obtained from the Bug-GDINA estimation, as shown in Table 5. The attribute tetrachoric correlation is based on the base-rate probabilities of attribute mastery (Toprak and Cakir, 2021). Generally, values above 0.70 are regarded as strong, 0.50–0.70 as moderate, and below 0.50 as weak (Aryadoust, 2018). In this study, moderate to strong correlations between barrier attributes were identified. As indicated, 5 out of 10 (50%) were strongly correlated (see the bold values), suggesting their strong tetrachoric correlations. For example, B3 exhibited moderate to strong relationships with all the other attributes, probably indicating that B3 alone could not cause listening failure, whereas when combined with other attributes, it may hinder students' listening cognitive processing (as learner No. 22's case in Figure 2). However, the correlation coefficient between B2 and B4 was the lowest (0.44), suggesting a certain degree of independence, which is also demonstrated in the case of learner No. 12.

To illustrate how the interrelationships between attributes can lead to an incorrect response to one particular item, four item profiles extracted from the Bug-GDINA estimation are presented in Table 6. Column 3–6 demonstrate the probability of an incorrect response to each item under different attribute patterns. As is shown, item 3's failure mainly results from B2 and B4, and the learner's possession of either of the two attributes or both would indiscriminately result in a high probability of a wrong response. This indicates the non-compensatory relation between B2 and B4. This is also the same for item 8. In contrast, B2 and B5 in item 13 demonstrate a compensatory relationship because the probability of an incorrect response is much lower when either barrier is present, but only when both barriers occur simultaneously do learners commit an error. The same is true for B1 and B4 in item 11.

Item 8 below is taken as an example to further explain the relationship between B3 and B4. If learners could not understand the syntactic structure of the subjunctive mood (B3) in the woman's utterance, it would be difficult for them to choose the correct option A. Otherwise, they would choose an incorrect answer, such as option D, based on irrelevant common knowledge (because the bad traffic may also lead to traffic accidents). Only when learners understand and synthesize utterances by both the man and woman, can they generate coherent representations (B4) and have a high probability of choosing the correct answer, indicating their non-compensatory interaction.

TABLE 5 Tetrachoric correlations between the attributes.

| | B1 | B2 | B3 | B4 | B5 |
|---|---|---|---|---|---|
| B1 | 1 | | | | |
| B2 | 0.53 | 1 | | | |
| B3 | **0.76** | **0.83** | 1 | | |
| B4 | **0.75** | 0.44 | 0.62 | 1 | |
| B5 | 0.58 | **0.79** | **0.92** | 0.62 | 1 |

The bold values in table show strong tetrachoric correlations.

TABLE 6 The four selected item profiles.

| Item | Attribute Pattern | P(00) | P(10) | P(01) | P(11) |
|---|---|---|---|---|---|
| 3 | B2-B4 | 0.12 | 0.83 | 0.89 | 0.94 |
| 8 | B3-B4 | 0.29 | 0.65 | 0.74 | 0.88 |
| 13 | B3-B5 | 0.22 | 0.52 | 0.45 | 0.78 |
| 11 | B1-B4 | 0.18 | 0.34 | 0.42 | 0.76 |

P (00) represents the non-possession of neither barrier, P (10) represents the possession of the first barrier attribute, P (01) represents the possession of the second barrier attribute, and P (11) represents the possession of both barriers.

Item 8: What happened to the speakers?

**(A) They missed the train because of the bad traffic. (Correct Answer).**

(B) They arrived at the railway station just in time.

(C) They barely caught the bus to the railway station.

(D) They had a traffic accident on the way to the station.

*Transcription:*

> (Woman): If the traffic wasn't so bad, we could have arrived at the railway station in time to catch the train.
>
> (Man): What a shame! We'll have to wait for the next train.

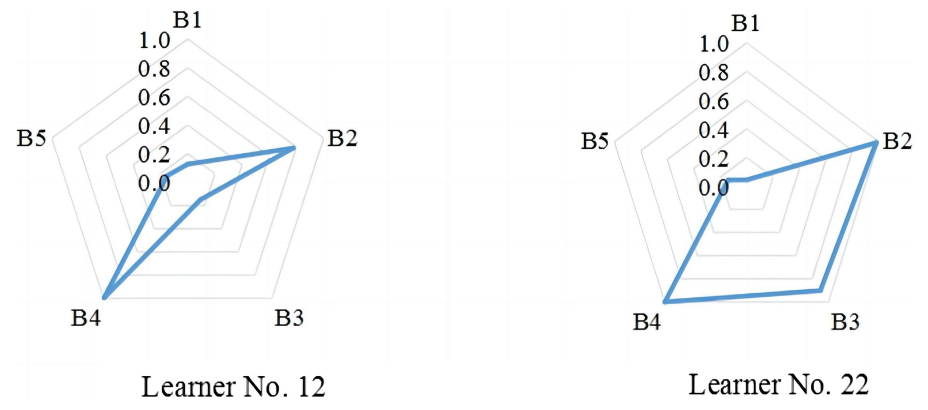



FIGURE 2
Different person parameter estimation of learners with the same total score.

# 5. Discussion

Unlike previous research on skill-based CDA in the language domain, this study set out to explore the feasibility of employing Bug-CDMs to investigate EFL learners' cognitive processing barriers in listening. The five major barrier attributes include "identifying speech" (B1), "recognizing vocabulary" (B2), "understanding syntactic or semantic structures" (B3), "understanding semantic meanings" (B4), and "making pragmatic inferences" (B5). Comparison analysis of five available Bug-CDMs revealed that Bug-GDINA was best fitted to the data. That means it carries the same qualities of a saturated model as GDINA, showing itself to be the most suitable model for uncovering the attributes of the underlying barriers affecting listening comprehension and their interactions with each other. Moreover, from the different mastery profiles of two cases (No. 12 and No. 22) with the same score, this study also stresses the advantage of Bug-CDMs in providing fine-grained diagnostic information, facilitating both instructors and learners with group and individualized feedback for the future improvement of their most problematic skills (Jang et al., 2015).

## 5.1. EFL learners' major listening barriers

By employing Bug-CDMs, the study addresses Chinese EFL learners' listening barriers. It contributes to a better understanding of different cognitive demands. The results reveal that the barriers in "understanding semantic meanings (B4)" and "recognizing vocabulary (B2)" are found to be the most prevalent causes for comprehension failure, followed by barriers in "understanding syntactic or semantic structures (B3)," "making pragmatic inferences (B5)," and "identifying speech (B1)." In general, they are quite distinct from the traditional listening problems defined in some studies such as "fast speed of delivery and difficulties in concentrating" (Flowerdew and Miller, 2005), "the shortage of background knowledge of English vocabulary" (Juan and Abidin, 2013), and "misunderstandings of speakers' accents" (Namaziandost et al., 2019). Possible reasons for the differences may be found in the different theoretical rationales and classification criteria. The current study is, first of all, based on the cognitive processing model, which is top-down oriented. At the same time, bottom-up evidence like learners' verbal reports serves as complementary evidence. In addition, the thus identified barrier attributes are mapped onto the learners' incorrect options to produce a bi-dimensional Q-matrix with emphasis on the stable cognition level of sustained misunderstanding or misuse. These multiple sources of evidence for Q-matrix construction ensured the reliability and validity of diagnostic inferences (Jang, 2009). By contrast, most listening problem studies are based on questionnaires, verbal reports, diaries or interviews (e.g., Goh, 2000; Graham, 2006; Nushi and Orouji, 2020; Alharbi and Al-Ahdal, 2022), which cannot establish such close relations between wrong answers and listening mechanisms.

Though different in terms, the current results do not radically deviate from the previous findings. As mentioned above, the most frequent barrier is understanding semantic meanings at discourse level (B4), suggesting that students find it difficult to generate a coherent representation of the processed utterance. This, to a large extent, corresponds to the problem "unable to form the mental representation of words heard" by Goh (2000) and Zhai and Liu (2010), since both are about meaning construction barriers. Results from verbal reports in this study further showed that although some learners were able to handle the semantic structures (B3) to generate the local representations at sentence level, it was still very challenging for them to connect the text already processed with the incoming new text (Cao et al., 2016). This finding confirms what Becker (2016) stated: analyzing and connecting different pieces of information often involves extra cognitive load and thus increases the difficulty of listening comprehension. This overloading is also due to the multitasking nature of listening, as Field (2009) as well as Zhai and Aryadoust (2022) pointed out, when listeners are presented with concurrent audio input, item stem and option reading, and answering, they are engaged in multitasking. This probably spreads their attention across multiple tasks, interfering with the creation of a coherent situation representation (Aryadoust et al., 2022). Besides, this might also be related to test-taking strategies when comprehension is obstructed, which was reflected in some verbal reports noting "keyword matching test-taking strategies." This meant that they often looked for keywords or phrases in the test items, and then matched them with what had been heard in order to locate the answer (Namaziandost et al., 2019). This strategy indicated learners' use of local-level processing (Field, 2009). In other words, they paid more attention to lexical matches instead of generating a global meaning representation of the texts (Zhai and Aryadoust, 2022).

The second most common barrier is a difficulty in understanding a listening text with many unfamiliar words (B2). Here, unfamiliar words not only involve those that are not acquired in written or oral form, but also those that sound unfamiliar. The current finding is in line with many studies which also proved that vocabulary-related problems prohibited the proper understanding of the listening content (e.g., Goh, 2000; Graham, 2006; Cao et al., 2016; Namaziandost et al., 2019; Alharbi and Al-Ahdal, 2022). In the latest listening barrier study conducted by Alharbi and Al-Ahdal (2022), students confessed to having difficulties in identifying the oral form of familiar words, making it harder to activate the relevant phonological and semantic information. Similarly, Goh (2000), Graham (2006), and Namaziandost et al. (2019) also found that most problems reported by learners were associated with vocabulary knowledge. For instance, students may miss the key information of the listening material when they are preoccupied with recalling spoken words or with new terminology, and this may interfere with the ongoing cognitive process. When examining the verbal report data, the authors found this interference was especially true in long passage listening tasks. Based on this, it is apparent that vocabulary and word recognition play a significant role in listening comprehension (Aryadoust, 2017).

The current results reveal that learners rarely possess barriers in identifying speech, which seems to conflict with Nushi and Orouji (2020) who found that the most significant listening difficulties are pronunciation-based ones such as "connected speech" and "not familiar with phonological features like assimilation or deletion of sounds." One possible reason may be the different participants and research methods in Nushi and Orouji's study. Their investigation was based on teachers' views on EFL learners' listening difficulties through questionnaire and semi-structured interviews.

The findings of this study suggest that EFL learners need to improve their spoken word recognition and their ability to synthesize a global meaning representation of the texts. Once they are aware of these potential problems, they can take tailored remedial actions to cope with them (Graham, 2006). Furthermore, by targeting the problematic areas that affects their comprehension most, instructors could make use of limited teaching time more profitably (Goh, 2000). Ideally, researchers

and teachers should work closely to address these barriers and so help learners enhance their listening comprehension ability.

## 5.2. Relationships between listening barriers

The tetrachoric correlations and item profiles make it possible to uncover the relationships between different barrier attributes. The results of this study confirm that there are both compensatory and non-compensatory relationships between the listening barrier attributes. The exploration of their interactions is also possible because the saturated Bug-GDINA model can capture both compensatory and non-compensatory mechanisms (Chen et al., 2013) and fits the data best. In other words, it contains multiple latent traits in such a way that failure on an item (or a task) requires multiple barriers or their interactions. For example, barriers B2 and B4 could either stand alone or combine together to hamper the success of listening cognitive processing (as in the case of item 3 profile in Table 6). One possible reason may lie in the intermediate listening proficiency of our participants. The previous studies reported that due to the limited linguistic knowledge, less-skilled listeners rely primarily on bottom-up (i.e., phonetic and lexical levels) processing (Rost, 2016; Min and Xiong, 2019), so they are more likely to encounter challenges with lower-level processing (Vandergrift and Baker, 2015). More proficient learners are often able to flexibly shift between top-down and bottom-up processing to activate linguistic and contextual information simultaneously (Nix, 2016; Furuya, 2021). The intermediate learners in the current study would less likely experience the lower-level processing difficulties such as phonetic perception, although the vocabulary recognition barrier is still commonly present. On the other hand, compared with highly proficient counterparts, learners in this study still encounter difficulty in higher-level processing such as understanding semantic meanings (B4). This can also be reflected in the personal attribute pattern of the learner No. 12, who ranks at the intermediate level (63% of the total) and demonstrates listening barriers mainly in vocabulary recognition (B2) and semantic meanings (B4). This result informs teachers that the two barriers should be tackled as a matter of priority, and even concurrently, in their remedial instruction for EFL intermediate learners.

In addition, this study revealed that the barrier in understanding syntactic or semantic structures (B3) alone does not hinder listening success unless combined with other barriers such as B5. This can be explained by the example of item 13 in Table 6, which demonstrates compensation of the two. This finding seems to be consistent with the opinion of Nix (2016), stating that though bottom-up processing plays a fundamental role in listening, the mastery of top-down processing is still indispensable, which indicates the interaction between lower and higher processing in listening (Min and Xiong, 2019). Specifically, comprehending the syntactic and semantic structures often involves a bottom-up process (Field, 2013; Rukthong and Brunfaut, 2019), and if these structures are not perceived, recognized or comprehended, learners prefer to utilize background knowledge and contextual cues such as prosody to compensate for the loss in lower-level processing (He and Xiong, 2021; He et al., 2022). In other words, learners can strategically adopt top-down processing to facilitate inference-making tasks (Min and Xiong, 2019; Chen and Chen, 2021). Another possible reason may be correlated with the difficulty level of listening barrier attributes. According to the diagnostic results of attribute prevalence, the average mastery probability of learners on B3 and B5 is 39.46 and 37.38% respectively, indicating both barriers are relatively less difficult or less likely to be encountered. As Ravand (2016) states, the interrelationship between skill attributes may vary with their difficulty level, and the interaction between skills with lower difficulty tends to be more compensatory. Based on the results of this study, it would also be possible for barrier attributes with less difficulty to demonstrate a compensatory nature.

# 6. Conclusion

This study addresses Chinese college EFL learners' listening barriers by cognitively diagnosing their incorrect options through the application of Bug-CDMs. It advances the curent CTT- or IRT- based uni-dimensional barrier research to the psychometric multi-dimentional CDA-based research, which can also reflect the attribute barriers' interactions within an item. Considering Bug-CDMs' limited application in EFL listening assessment, this study can be seen as a significant step toward their feasibility and usefulness in L2 research.

However, the study is not without limitations. First, we did not investigate barriers across learners with different proficiency levels. For better understanding and interpretation of barriers, future research is recommended to address this area. Second, this study focuses only on the response level of an item, i.e., whether the answer is correct or incorrect, not the information at the individual option level. Considering this, one more step is desired to diagnose students' weaknesses from the option level, which may help improve the diagnostic accuracy, thereby contributing to the more targeted and in-depth feedback.

Despite the aforementioned limitations, the current study has great implications in that it confirmed that Bug-CDMs can offer valuable insights into how listening barriers are distributed among EFL learners, and how these barriers interact in complex ways. The findings would inspire EFL teachers to provide effective remedial instructions. At the same time they inform the high-quality MCQ test design and development.

# Data availability statement

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

# Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

# Author contributions

YM: study design, data collection, supervision, proposal writing, multiple draft writing, and revision. YW: final data analysis, original proposal writing, multiple draft writing, and revision. NZ: data collection and preliminary data analysis. All authors contributed to the article and approved the submitted version.

## Funding

This work was supported by the National Social Science Fund of China (Grant No. 16BYY096), the NEEA and BC English Assessment Research Grants (Grant No. EARG2020007), and Shaanxi Province Social Science Foundation (Grant No. 2022K005).

## Conflict of interest

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

## Publisher's note

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

## Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1126106/full#supplementary-material

## References

Alharbi, M. A., and Al-Ahdal, A. A. M. H. (2022). Communication barriers in the EFL classroom: is poor listening the culprit that obstructs learning? *Interact. Learn. Environ.*, 1–15. doi: 10.1080/10494820.2022.2098776

Anderson, J. R. (1995). *Cognitive psychology and its implications*. New York, NY: Worth.

Andrade, H. L., and Heritage, M. (2017). *Using formative assessment to enhance learning, achievement, and academic self-regulation* Routledge.

Aryadoust, V. (2017). *Teaching and assessing reading comprehension to improve literacy: an integrated model and its application. Keynote speech presented at English literacy pedagogy in 21st century: striving for multiple literacies in classroom practices, Indonesia*.

Aryadoust, V. (2018). A cognitive diagnostic assessment study of the listening test of the Singapore–Cambridge general certificate of education O-level: application of DINA, DINO, G-DINA, HO-DINA, and RRUM. *Int. J. Listen.* 35, 1–24. doi: 10.1080/10904018.2018.1500915

Aryadoust, V., Foo, S., and Ng, L. Y. (2022). What can gaze behaviors, neuroimaging data, and test scores tell us about test method effects and cognitive load in listening assessments? *Lang. Test.* 39, 56–89. doi: 10.1177/02655322211026876

Aryadoust, V., and Luo, L. (2022). The typology of second language listening constructs: a systematic review. *Lang. Test.* 40, 375–409. doi: 10.1177/02655322211126604

Bachman, L. F., and Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global information. *J. Engl. Acad. Purp.* 24, 1–13. doi: 10.1016/j.jeap.2016.07.004

Bradshaw, L., and Templin, J. (2014). Combining item response theory and diagnostic classification models: a psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika* 79, 403–425. doi: 10.1007/s11336-013-9350-4

Briggs, D. C., Alonzo, A. C., Schwab, C., and Wilson, M. (2006). Diagnostic assessment with ordered multiple choice items. *Educ. Assess.* 11, 33–63. doi: 10.1207/s15326977ea1101_2

Buck, G. (2001). *Assessing listening*. Beijing: Foreign Language Teaching and Research Press.

Cao, X., Yamashita, N., and Ishida, T. (2016). How non-native speakers perceive listening comprehension problems: implications for adaptive support technologies. In Collaboration Technologies and Social Computing: 8th International Conference, Collab Tech 2016, Kanazawa, Springer Singapore, 89–104.

Chen, H., and Chen, J. (2021). Investigating the relationships between listening skills and genre competence through cognitive diagnosis approach. *SAGE Open* 11, 1–14. doi: 10.1177/21582440211061342

Chen, J., De La Torre, J., and Zhang, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *J. Educ. Meas.* 50, 123–140. doi: 10.1111/j.1745-3984.2012.00185.x

Cho, H. (2021). An analysis of listening errors by Korean EFL learners from self-paced passage dictation. *Phonet. Speech Sci.* 13, 17–24. doi: 10.13064/ksss.2021.13.1.017

Dong, Y., Ma, X., Wang, C., and Gao, X. (2021). An optimal choice of cognitive diagnostic model for second language listening comprehension test. *Front. Psychol.* 12:608320. doi: 10.3389/fpsyg.2021.608320

Du, W., and Ma, X. (2021). Probing what's behind the test score: application of multi-CDM to diagnose EFL learners' reading performance. *Read. Writ.* 34, 1441–1466. doi: 10.1007/s11145-021-10124-x

Dunkel, P. (1991). Listening in the native and second/foreign language: toward an integration of research and practice. *TESOL Q.* 25, 431–457. doi: 10.2307/3586979

Feng, G., and Xu, J. (2022). Investigating English listening comprehension difficulties based on the language comprehension theory (in Chinese). *Foreign Lang. World.* 4, 49–56.

Field, J. (2009). *Listening in the language classroom*. Cambridge: Cambridge University Press.

Field, J. (2013). "Cognitive validity" in *Examining listening: research and practice in assessing second language listening*. eds. A. Geranpayeh and L. Taylor (Cambridge: Cambridge University Press), 77–151.

Flowerdew, J., and Miller, L. (2005). *Second language listening: theory and practice*. Cambridge, UK: Cambridge University Press.

Furuya, A. (2021). How do listening comprehension processes differ by second language proficiency? Top-down and bottom-up perspectives. *Int. J. Listen.* 35, 123–133. doi: 10.1080/10904018.2019.1694411

Goh, C. C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System* 28, 55–75. doi: 10.1016/s0346-251x(99)00060-3

Graham, S. (2006). Listening comprehension: the learners' perspective. *System* 34, 165–182. doi: 10.1016/j.system.2005.11.001

Haladyna, T. M., and Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educ. Psychol. Meas.* 53, 999–1010. doi: 10.1177/0013164493053004013

Haladyna, T. M., and Rodriguez, M. C. (2013). *Developing and validating test items*. London, UK: Routledge.

Harding, L., Alderson, J., and Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: elaborating on diagnostic principles. *Lang. Test.* 32, 317–336. doi: 10.1177/0265532214564505

He, L., Jiang, Z., and Min, S. (2021). Diagnosing writing ability using China's standards of English language ability: application of cognitive diagnosis models. *Assess. Writ.* 50:100565. doi: 10.1016/j.asw.2021.100565

He, L., and Xiong, L. (2021). A new approach to the development of cognitive processing models in second language listening comprehension: a method combining cognitive diagnosis with standard setting (in Chinese). *Foreign Lang. World*. 4, 35–43.

He, L., Xiong, L., and Min, S. (2022). Diagnosing listening and reading skills in the Chinese EFL context: performance stability and variability across modalities and performance levels. *System* 106, 102787–102719. doi: 10.1016/j.system.2022.102787

Hwang, M. H. (2004). Exploring listening comprehension difficulties observed among Korean high school students. *J. Curric. Eval.* 7, 275–301. doi: 10.29221/jce.2004.7.1.275

Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Lang. Assess. Q.* 6, 210–238. doi: 10.1080/15434300903071817

Jang, E. E., Dunlop, M., Park, G., and Van Der Boom, E. H. (2015). How do young students with different profiles of reading skill mastery, perceived ability, and goal orientation respond to holistic diagnostic feedback? *Lang. Test.* 32, 359–383. doi: 10.1177/0265532215570924

Jones, G. (2020). "Designing multiple-choice test items" in *The Routledge handbook of second language acquisition and language testing*. eds. P. Winke and T. Brunfaut (New York, USA: Routledge), 90–101.

Juan, W. X., and Abidin, M. J. Z. (2013). English listening comprehension problems of students from China learning English in Malaysia. *Lang. India* 13, 14009–14022.

Kao, Y. T., and Kuo, H. C. (2021). Diagnosing l2 English learners' listening difficulties and learning needs through computerized dynamic assessment. *Interact. Learn. Environ.* 31, 2219–2243. doi: 10.1080/10494820.2021.1876738

Kuo, B. C., Chen, C. H., and de la Torre, J. (2018). A cognitive diagnosis model for identifying coexisting skills and misconceptions. *Appl. Psychol. Meas.* 42, 179–191. doi: 10.1177/0146621617722791

Kuo, B. C., Chen, C. H., Yang, C. W., and Mok, M. M. C. (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educ. Psychol.* 36, 1115–1133. doi: 10.1080/01443410.2016.1166176

Lee, Y. W. (2015). Diagnosing diagnostic language assessment. *Lang. Test.* 32, 299–316. doi: 10.1177/0265532214565387

Leighton, J., and Gierl, M. (Eds.) (2007). *Cognitive diagnostic assessment for education: theory and applications*. Cambridge, UK: Cambridge University Press.

Ma, W., and de la Torre, J. (2020). An empirical Q-matrix validation method for the sequential generalized DINA model. *Br. J. Math. Stat. Psychol.* 73, 142–163. doi: 10.1111/bmsp.12156

Meng, Y. (2013). *Developing a model of cognitive diagnostic assessment for college EFL listening*. Unpublished doctoral dissertation, Shanghai International Studies University, China.

Meng, Y., and Fu, H. (2023). Modeling mediation in the dynamic assessment of listening ability from the cognitive diagnostic perspective. *Mod. Lang. J.* 107, 137–160. doi: 10.1111/modl.12820

Min, S., and Xiong, L. (2019). Exploring the compensatory mechanism of EFL listening ability (in Chinese). *Modern Foreign Lang.* 42, 254–266.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. Massachusetts, USA: The MIT Press.

Namaziandost, E., Neisi, L., Mahdavirad, F., and Nasri, M. (2019). The relationship between listening comprehension problems and strategy usage among advance EFL learners. *Cogent. Psychol.* 6:1691338. doi: 10.1080/23311908.2019.1691338

Nix, J. (2016). Measuring latent listening strategies: development and validation of the EFL listening strategy inventory. *System* 57, 79–97. doi: 10.1016/j.system.2016.02.001

Noroozi, S., Sim, T. S., Nimehchisalem, V., and Zareian, G. (2014). Developing an instrument for Iranian EFL learners' listening comprehension problems and listening strategies. *Ad. Lang. Lit. Stud.* 5, 63–69. doi: 10.7575/aiac.alls.v.5n.3p.63

Nushi, M., and Orouji, F. (2020). Investigating EFL teachers' views on listening difficulties among their learners: the case of Iranian context. *SAGE Open* 10:215824402091739. doi: 10.1177/2158244020917393

Ozaki, K., Sugawara, S., and Arai, N. (2019). Cognitive diagnosis models for estimation of misconceptions analyzing multiple-choice data. *Behaviormetrika* 47, 19–41. doi: 10.1007/s41237-019-00100-9

Ravand, H. (2016). Application of a cognitive diagnostic model to a high-stakes reading comprehension test. *J. Psychoeduc. Assess.* 34, 782–799. doi: 10.1177/0734282915623053

Rost, M. (2016). *Teaching and researching listening. 3rd ed.*. London: Longman.

Rukthong, A., and Brunfaut, T. (2019). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Lang. Test.* 37, 31–53. doi: 10.1177/0265532219871470

Rupp, A. A., Templin, J. L., and Henson, R. A. (2010). *Diagnostic measurement: theory, methods, and applications*. New York, NY: Guildford Press.

Sawaki, Y., Kim, H. J., and Gentile, C. (2009). Q-matrix construction: defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Lang. Assess. Q.* 6, 190–209. doi: 10.1080/15434300902801917

Steedle, J., and Shavelson, R. (2009). Supporting valid interpretations of learning progression level diagnoses. *J. Res. Sci. Teach.* 46, 699–715. doi: 10.1002/tea.20308

Stout, W., Henson, R., and DiBello, L. (2022). Optimal classification methods for diagnosing latent skills and misconceptions for option-scored multiple-choice item quizzes. *Behaviormetrika* 50, 177–215. doi: 10.1007/s41237-022-00172-0

Toprak, T. E., and Cakir, A. (2021). Examining the L2 reading comprehension ability of adult ELLs: developing a diagnostic test within the cognitive diagnostic assessment framework. *Lang. Test.* 38, 106–131. doi: 10.1177/0265532220941470

Vafaee, P., and Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Stud. Second. Lang. Acquis.* 42, 383–410. doi: 10.1017/S0272263119000676

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Lang. Teach.* 40, 191–210. doi: 10.1017/S0261444807004338

Vandergrift, L., and Baker, S. (2015). Learner variables in second language listening comprehension: an exploratory path analysis. *Lang. Learn.* 65, 390–416. doi: 10.1111/lang.12105

Vandergrift, L., and Goh, C. M. (2012). *Teaching and learning second language listening: metacognition in action*. New York, NY: Routledge.

Wolf, M., Muijselaar, M., Boonstra, A., and Bree, E. (2019). The relationship between reading and listening comprehension: shared and modality-specific components. *Read. Writ.* 32, 1747–1767. doi: 10.1007/s11145-018-9924-8

Xu, J., and Nie, R. (2016). A comprehensive survey of English ability of freshmen in key universities in China —a case study of freshmen in 2014 (in Chinese). *Foreign Language World*. 1, 18–26.

Yeldham, M. (2022). Examining the interaction between two process-based L2 listening instruction methods and listener proficiency level: which form of instruction most benefits which learners? *TESOL Q.* 56, 688–712. doi: 10.1002/tesq.3089

Yi, Y. S. (2017). Probing the relative importance of different attributes in L2 reading and listening comprehension items: an application of cognitive diagnostic models. *Lang. Test.* 34, 337–355. doi: 10.1177/0265532216646141

Zhai, J., and Aryadoust, V. (2022). The metacognitive and neurocognitive signatures of test methods in academic listening. *Front. Psychol.* 13:930075. doi: 10.3389/fpsyg.2022.930075

Zhai, F., and Liu, Z. (2010). A cognitive analysis of three-phase model of listening comprehension obstacles and its countermeasures (in Chinese). *J. Changchun Univ. Sci. Technol.* 5, 114–115. doi: 10.15977/j.cnki.cn32-1555/g4.2013.04.005

Zhang, Z., Lu, L., and Tan, X. (2010). Cognitive analysis of relationships between language proficiency and listening comprehension problems (in Chinese). *Foreign Lang. Learn. Teach. Pract.* 1, 33–38.

Zhang, Z., and Zhang, L. J. (2011). Developing a listening comprehension problem scale for university students' metacognitive awareness. *J. Asia TEFL* 8, 161–189.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 26 October 2023
DOI 10.3389/fpsyg.2023.1122471



OPEN ACCESS

EDITED BY
Oliver Niebuhr,
University of Southern Denmark, Denmark

REVIEWED BY
Mafuyu Kitahara,
Sophia University, Japan
Takeshi Nozawa,
Ritsumeikan University, Japan
Jeff Holliday,
University of Kansas, United States

*CORRESPONDENCE
Kakeru Yazawa
✉ yazawa.kakeru.gb@u.tsukuba.ac.jp
James Whang
✉ jamesw@snu.ac.kr

RECEIVED 13 December 2022
ACCEPTED 03 October 2023
PUBLISHED 26 October 2023

CITATION
Yazawa K, Whang J and Escudero P (2023) Australian English listeners' perception of Japanese vowel length reveals underlying phonological knowledge.
*Front. Psychol.* 14:1122471.
doi: 10.3389/fpsyg.2023.1122471



# Australian English listeners' perception of Japanese vowel length reveals underlying phonological knowledge

Kakeru Yazawa[1]*, James Whang[2]* and Paola Escudero[3]

[1]Institutes of Humanities and Social Sciences, University of Tsukuba, Tsukuba, Japan, [2]Department of Linguistics, Seoul National University, Seoul, Republic of Korea, [3]The MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, NSW, Australia

Speech perception patterns are strongly influenced by one's native phonology. It is generally accepted that native English listeners rely primarily on spectral cues when perceiving vowels, making limited use of duration cues because English lacks phonemic vowel length. However, the literature on vowel perception by English listeners shows a marked bias toward American English, with the phonological diversity among different varieties of English largely overlooked. The current study investigates the perception of Japanese vowel length contrasts by native listeners of Australian English, which is reported to use length to distinguish vowels unlike most other varieties of English. Twenty monolingual Australian English listeners participated in a forced-choice experiment, where they categorized Japanese long and short vowels as most similar to their native vowel categories. The results showed a general tendency for Japanese long and short vowels (e.g., /ii, i/) to be categorized as Australian English long and short vowels (e.g., /iː, ɪ/ as in "heed," "hid"), respectively, which contrasts with American English listeners' categorization of all Japanese vowels as tense regardless of length (e.g., /ii, i/ as both "heed") as reported previously. Moreover, this duration-based categorization was found not only for Australian English categories that contrast in duration alone (e.g., /ɐː, ɐ/ as in "hard," "hud") but also for those that contrast in both duration and spectra (e.g., /oː, ɔ/ as in "hoard," "hod"), despite their spectral mismatch from the corresponding Japanese vowels (e.g., /aa, a/ and /oo, o/). The results, therefore, suggest that duration cues play a prominent role across all vowel categories—even nonnative—for Australian English listeners. The finding supports a feature-based framework of speech perception, where phonological features like length are shared across multiple categories, rather than the segment-based framework that is currently dominant, which regards acoustic cues like duration as being tied to a specific native segmental category. Implications for second and foreign language learning are discussed.

KEYWORDS

**Australian English, Japanese, cross-linguistic perception, vowel, phonological feature, length, acoustic cue, duration**

## 1. Introduction

Languages differ as to which acoustic cues are phonologically meaningful and in what way. Some languages such as Arabic, Czech, Japanese, and Swedish utilize vowel duration phonemically (International Phonetic Association, 1999), where long and short vowels of the same quality convey different lexical meanings (e.g., *ii* [iː] "good"—*i* [i] "stomach" in

Japanese).[1] A few languages such as Estonian even employ a more complex, three-way distinction (Asu and Teras, 2009): *kalu* [kɑlu] "fish" (partitive plural)—*kaalu* [kɑːlu] "scales" (genitive singular)—*kaalu* [kɑːːlu] "scales" (partitive singular). English, on the other hand, is said to lack phonological vowel length, since changes in vowel duration alone would not change the meaning of the word (e.g., *Do it!* [du ɪt]—*Doooo iiiit!* [duː ɪːt]). Such cross-linguistic differences in native phonology are known to shape speech perception patterns (Jacquemot et al., 2003; Escudero et al., 2009; Mazuka et al., 2011; Lipski et al., 2012; Yazawa et al., 2020). For instance, it is generally accepted that English listeners rely primarily on spectral cues and little on duration cues when perceiving native and nonnative vowels because vowel length is not phonemic in English (Hillenbrand et al., 2000; McAllister et al., 2002; Hirata, 2004; Dietrich et al., 2007; Kondaurova and Francis, 2008; Nishi et al., 2008; Mugitani et al., 2009; Karpinska et al., 2015). The literature on vowel perception by English listeners, however, shows a marked bias toward American English (AmE), with the phonological diversity among different varieties of English being largely overlooked. The current study therefore examines the perception of Japanese vowel length by native listeners of Australian English (AusE), which is reported to use length to distinguish vowels unlike most other varieties of English.

Previous research on AmE listeners has found a marginal role of duration as a perceptual cue for vowel identity. Hillenbrand et al. (2000) tested native AmE listeners on synthesized /hVd/ tokens with altered vowel durations, finding a small overall effect of duration on their vowel identification. While some vowel contrasts such as /ɛ/-/æ/ and /ʌ/-/ɑ(ɔ)/ were significantly affected by duration, those that differ systematically in duration such as /i/-/ɪ/, /u/-/ʊ/, /ɪ/-/e/-/ɛ/ were minimally affected. Similarly, Kondaurova and Francis (2008) used synthetic *beat-bit* tokens varying in nine perceptually equidistant spectral and durational steps, finding that AmE listeners relied predominantly on vowel spectra. The primacy of spectral cues has also been found in cross-linguistic and second language (L2) perception by AmE listeners. Of particular relevance to the current study, Nishi et al. (2008) found that AmE listeners categorized Japanese long and short vowels (embedded in /hVba/, spoken by four male Japanese speakers in citation and sentence forms) as most similar to AmE tense vowels regardless of length (Table 1). The duration of the Japanese vowels was thus being ignored, although a small effect of duration was found in the categorization of Japanese /ee/-/e/ and /aa/-/a/, possibly reflecting the status of AmE /ɛ/-/æ/ and /ʌ/-/ɑ(ɔ)/ discussed above. Hirata (2004) further tested whether first-language (L1) AmE listeners can learn to correctly identify Japanese vowel length contrasts through supervised perceptual training. The result showed a statistically significant improvement from pre-test (overall 39% correct) to post-test (about 54% correct for the sentence condition and 80% correct for the isolated condition), indicating that the length contrasts are difficult yet learnable for AmE listeners. Finally, the observed underutilization of vowel duration by AmE listeners is an influence of native phonology, as AmE-learning 18-month-old infants can detect changes in vowel duration in the same way as Japanese adults but do not interpret the changes as lexically contrastive (Dietrich et al., 2007; Mugitani et al., 2009).

Much less is understood about listeners of other varieties of English, which warrants attention since different varieties of a language can exhibit divergent perceptual patterns (Escudero and Boersma, 2004; Escudero and Williams, 2012; Escudero et al., 2012; Williams and Escudero, 2014). Again using synthetic *beat-bit* tokens differing in spectral and duration steps, Karpinska et al. (2015) found that English listeners from England, Scotland, Wales, Ireland, New Zealand, and Singapore behaved similarly to AmE listeners, showing primary reliance on vowel spectra. Thus, it appears that listeners of most varieties of English are perceptually alike, i.e., underutilizing duration for vowel identity. However, the study also found a distinct perceptual pattern in AusE listeners, who relied primarily on duration rather than spectra. Williams et al. (2018) extended this finding by showing that duration, along with vowel inherent spectral change (VISC), is a crucial cue for AusE listeners to distinguish *bid* from *bead* and *beard*. This makes AusE listeners an exception, at least regarding high front vowels. Chen et al. (2014) also found that AusE-learning 18-month-olds perceive the durational difference between AusE /ɐː/ and /ɐ/ as lexically contrastive, suggesting that the duration-based perception extends to non-high-front vowels. AusE listeners' sensitivity to vowel duration has been documented in their nonnative perception as well. Tsukada (2012) conducted an AXB discrimination test of vowel length contrasts in Arabic (/ii, aa, uu/—/i, a, u/) and Japanese (/ii, ee aa, oo, uu/—/i, e, a, o, u/) by native Arabic, Japanese, and AusE listeners, where the Arabic and Japanese groups were expected to outperform the AusE "control" group in nonnnative perception because "the extent to which vowel duration is used contrastively in Australian English is likely to be more limited than in Arabic or Japanese" (Tsukada, 2012 p. 511). Contrary to the expectation, the study found no significant advantage of Arabic and Japanese listeners over AusE listeners, who achieved an overall discrimination accuracy of 82% for Arabic vowels and 75% for Japanese vowels despite both languages being nonnative. This, in turn, indicates that AusE listeners are generally sensitive to vowel duration.

The results of Tsukada (2012) show that AusE listeners are able to use duration to discriminate nonnative Japanese vowels. The current study further tests how AusE listeners identify Japanese long and short vowels as their native categories in a forced-choice perception experiment. The distinction between discrimination and identification is important here, since the ability to detect changes in acoustic-phonetic vowel duration does not entail that length is part of phonological vowel identity, as the aforementioned studies on AmE-learning infants have demonstrated (Dietrich et al., 2007; Mugitani et al., 2009). Unlike AmE listeners whose categorization of Japanese vowels was largely unaffected by length (Table 1), AusE listeners may categorize Japanese long and short vowels into different AusE categories according to length (Table 2).[2] If so, this would indicate that length determines

1 Japanese has five distinct qualities /i, e, a, o, u/, which form five short (1-mora) and long (2-mora) pairs (Keating and Huffman, 1984). The long vowels can be considered phonologically as a sequence of two short vowels and therefore are transcribed with double letters (/ii, ee, aa, oo, uu/) in this paper.

2 In this study, AusE vowels with and without the "ː" symbol are treated as being long and short, respectively.

TABLE 1 Categorization of Japanese vowels by AmE listeners (in percentage, bold = modal responses).

| | | Japanese vowel stimuli | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | /ii/ | /i/ | /ee/ | /e/ | /aa/ | /a/ | /oo/ | /o/ | /uu/ | /u/ |
| Perceived AmE vowel | /i/ | **99** | **95** | | | | | | | | |
| | /ɪ/ | 1 | 4 | 2 | 16 | | | | | | |
| | /eɪ/ | | | **94** | **76** | | | | | | |
| | /ɛ/ | | 1 | 5 | 8 | | | | | | |
| | /æ/ | | | | | 2 | 3 | | | | |
| | /ɑ(ɔ)/ | | | | | **89** | **57** | | | | |
| | /ʌ/ | | | | | 9 | 39 | | 1 | 1 | 3 |
| | /oʊ/ | | | | | | | **99** | **95** | 2 | 1 |
| | /u/ | | | | | | | 1 | 2 | **92** | **91** |
| | /ʊ/ | | | | | | | | 1 | 5 | 5 |

Adapted from Nishi et al. (2008), citation condition.

TABLE 2 AusE vowel categories and example words (Cox and Palethorpe, 2007).

| Vowel | Word | Vowel | Word |
|---|---|---|---|
| /iː/ | *heed* | /ɪ/ | *hid* |
| /eː/ | *haired* | /e/ | *head* |
| /ɜː/ | *heard* | /æ/ | *had* |
| /ɐː/ | *hard* | /ɐ/ | *hud* |
| /oː/ | *hoard* | /ɔ/ | *hod* |
| /ʉː/ | *food* | /ʊ/ | *hood* |
| /ɪə/ | *feared* | | |

phonological vowel identity in AusE, making it an exception among the many varieties of English thought to lack contrastive length.

A theoretically important question pertinent to the above prediction is whether AusE listeners' use of duration in vowel categorization would vary depending on the type of Japanese vowel. In AusE, only a subset of vowel categories contrast in duration alone (/eː/-/e/ and /ɐː/-/ɐ/), while others contrast in both duration and spectra (Cox, 2006; Cox and Palethorpe, 2007; Ratko et al., 2022). It is thus possible that AusE listeners more readily use duration when perceiving Japanese vowels that spectrally match the former (e.g., /ee/-/e/ and /aa/-/a/) than those matching the latter (e.g., /ii/-/i/). Alternatively, given their general sensitivity to vowel duration in nonnative length discrimination (Tsukada, 2012), AusE listeners may equally utilize duration in categorizing all Japanese vowels. These two possibilities are closely related to segment- and feature-based frameworks of speech perception. Current models of cross-linguistic perception generally subscribe to the segment-based view. For example, the Perceptual Assimilation Model (PAM; Best, 1995; Best and Tyler, 2007) and the Speech Learning Model (SLM; Flege, 1995; Flege and Bohn, 2021) explain cross-linguistic perception patterns as a result of nonnative sounds being assimilated to or classified as equivalent to existing native segmental categories. A common implicit premise of these models is that the use of acoustic cues in the assimilation or classification process is specific to each native category. Thus, if duration is an important cue for certain categories but not for others in the L1, then the categorization of nonnative sounds assimilated to the former categories will be duration-dependent and that of those assimilated to the latter categories will not be. The alternative, feature-based view derives from the "feature" hypothesis, which asserts that "L2 features not used to signal phonological contrast in L1 will be difficult to perceive for the L2 learner" (McAllister et al., 2002, p. 230). A crucial assumption underlying this view is that a feature is available to the whole phonological system, independent of specific categories. Thus, if a length feature exists in L1 phonology, then the use of duration cues owing to the feature may extend beyond certain L1 categories with the feature.

Previous studies of cross-linguistic length perception with other languages provide mixed evidence for the above two frameworks. Chládková et al. (2013) examined pre-attentive sensitivity to duration in native and nonnative vowels across Dutch, Czech, and Spanish, using electroencephalography (EEG) to measure mismatch negativity (MMN). Dutch was of particular interest because its phonological status of vowel duration is rather unclear, with all vowel categories that contrast in duration also contrasting in spectra (e.g., /aː/ in *maan* "moon"—/ɑ/ in *man* "man"). It was found that Dutch listeners' sensitivity to duration was comparable to Czech listeners' but greater than Spanish listeners' when the vowel quality was [a] (i.e., native to all three languages), suggesting that Dutch uses vowel duration phonemically as in Czech. However, Dutch listeners' sensitivity to duration was significantly reduced compared to Czech listeners when the quality was changed to Estonian [ɤ] (i.e., nonnative to all three languages). This suggests that Dutch listeners do not disentangle duration cues from spectral cues, perhaps due to their obligatory co-occurrence, consequently confining the phonemic use of vowel duration to native vowel categories. While the result needs to be treated with caution because no significant difference in MMN was found within Dutch listeners between the native and nonnative conditions, the overall finding aligns with the segment-based view. Chládková et al. (2015b)

further examined Dutch listeners' perceptual sensitivity to duration in [a] and [ɑ] qualities and found a larger MMN amplitude for the former. This suggests that duration is phonemically relevant for the *maan* vowel that is represented as "long" but phonemically unspecified for the *man* vowel, providing further evidence that the use of duration is vowel-specific in Dutch.[3]

In contrast, the aforementioned study of McAllister et al. (2002) lends support to the feature-based view. The study compared the perception and production accuracy of L2 Swedish vowel length by L1 Estonian, AmE, and Spanish participants, who had all lived in Sweden for at least 10 years. It was found that the L1 AmE and Spanish groups performed significantly worse than the L1 Estonian group, which was taken as evidence for the transfer of a vowel length feature that is present in Estonian but is absent in AmE or Spanish. Of particular note from the results is that the Estonian group was indistinguishable from native Swedish controls in their implementation of duration during production. In Swedish, the length of vowels and consonants are in complementary distribution in stressed syllables—a short consonant follows a long vowel and a long consonant follows a short vowel—which differs from Estonian where vowels and consonants have independent length. The results, therefore, suggest that Estonian speakers were able to learn and implement the complementary duration for consonants and vowels in L2 Swedish, despite no such relationship existing in their L1. Adding to this finding, Pajak and Levy (2014) found that native listeners of a language with vowel length contrasts showed enhanced discrimination of nonnative consonant length contrasts (i.e., geminates). These results together imply the existence of a length feature that is shared across vowel and consonant categories, which seems accessible in nonnative and L2 perception. However, given that the above two studies focus on perception accuracy while those in support of the segment-based view (Chládková et al., 2013, 2015b) focus on perceptual sensitivity, these sets of evidence may not be strictly compatible with each other, thus leaving room for further investigation.

Following the discussion above, the predictions going into the current study are summarized as follows. First, if vowel length is indeed used phonologically in AusE, AusE listeners will show a tendency to categorize long and short Japanese vowels as long and short AusE counterparts, respectively. Second, should the results show stronger duration effects for certain Japanese vowels (e.g., /ee/-/e/ and /aa/-/a/), this would suggest that AusE listeners utilize duration cues only as necessitated by their native categories (e.g., /eː/-/e/ and /ɐː/-/ɐ/), supporting the segment-based framework of speech perception. Alternatively, should similar duration effects be observed across all vowel qualities, this would suggest that AusE listeners are able to extend the use of duration cues beyond their native categories, supporting the feature-based framework. These predictions will be tested in the experiment presented below.

3 While the traditional binary feature of length (i.e., [±long]) would not hold here, the obtained results are not incompatible with the feature-based view per se because features can be univalent (Gussenhoven and Jacobs, 2017); segments without a "long" feature are simply unmarked rather than having a "short" feature.

# 2. Materials and methods

## 2.1. Participants

Twenty female native AusE listeners were recruited for the experiment at Western Sydney University, Sydney, Australia. They were undergraduate or graduate students at the University between the ages of 17 and 35 (mean age = 21.4), born and raised in the greater Sydney area. All participants reported normal hearing and only very basic knowledge of any foreign language. They were compensated for their time in the form of course credit.

## 2.2. Stimuli

The stimuli were 10 Japanese vowels—five long /ii, ee, aa, oo, uu/ and five short /i, e, a, o, u/—embedded in three consonantal contexts (/bVp, dVt, gVk/) and spoken by 10 native Japanese speakers (five female, five male), for a total of 300 tokens. These are a subset of the production data reported in Yazawa and Kondo (2019). The speakers were students or graduates of universities in Japan between the ages of 21 and 27 (mean age = 23.9) who had spent most of their lives in Tokyo and surrounding areas. They read aloud the sentence */CVCe/—/CVCo/—/CVCe/ to /CVCo/ ni wa V ga aru* "/CVCe/—/CVCo/—In /CVCe/ and /CVCo/ there is V," where V is the target Japanese vowel with the lexical pitch accent.[4] Each sentence was presented in Japanese *kana* orthography, which the speakers read at a comfortable speed. The /e/ in the underlined /CVCe/ was then excised at the first positive zero crossing of the vowel to create /CVC/ stimuli in Praat (Boersma and Weenink, 2022). The utterances were recorded in an anechoic chamber at Waseda University, Tokyo, Japan, using a SONY F-780 microphone with a 44,100 Hz sampling frequency and 16-bit resolution. The volume of all stimuli was adjusted to have a peak intensity of 70 dB.

## 2.3. Procedure

Prior to the experiment, the participants signed a consent form and completed a language background questionnaire. They were informed that they would be listening to "sounds from a foreign language."

The experiment was a forced-choice task, where the participants had to categorize the vowel in the aforementioned Japanese /CVC/ stimuli presented in isolation. During the experiment, the participants were shown on a computer monitor a list containing the words in Table 2. After hearing each stimulus, the participants chose the word containing a vowel that best matched the vowel in the stimulus (e.g., [diːt] → *heed*). The words in the list all had the shape /hVd/, except for two words that had the shape /fVd/. Participants were asked to make their choice as quickly as possible. The stimuli were presented in random order

4 Although accented and unaccented vowels can have different durations in Japanese (Kozasa, 2004), this study uses accented tokens so that the stimuli conform to the default accent placement pattern for loanwords and nonwords (Kawahara, 2015).

10.3389/fpsyg.2023.1122471

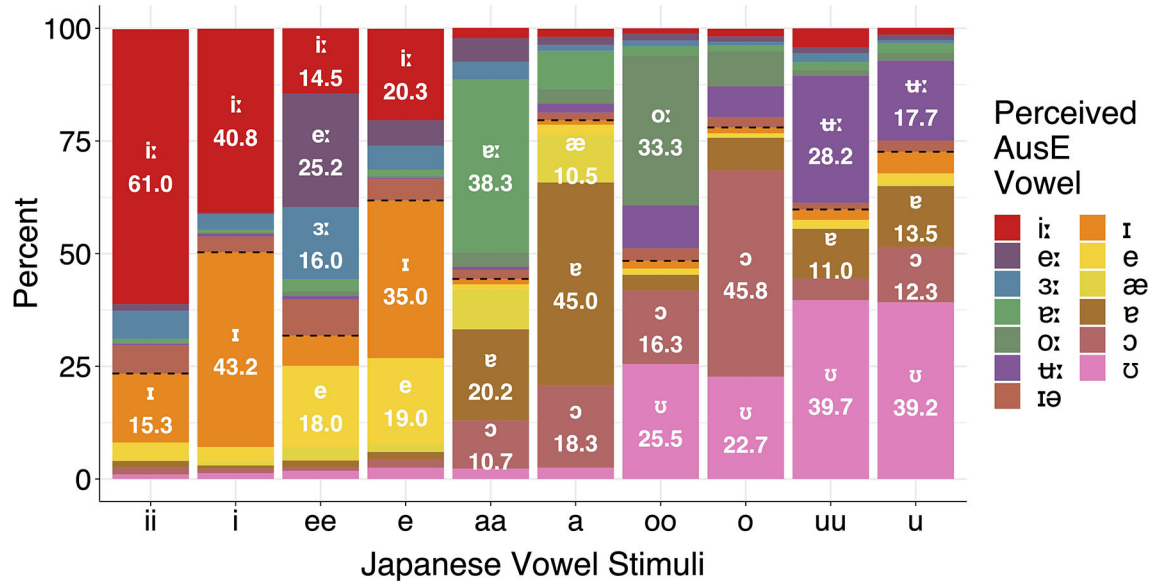



FIGURE 1
Categorization of Japanese vowels by all participants. The dotted lines show the boundaries between "long" and "short" AusE responses. Only responses >10% are labeled.

through noise-isolating headphones, and participants responded by clicking the word choice with a computer mouse. A break was programmed to occur after 150 tokens (i.e., midpoint of experiment), which ended when participants clicked the mouse. The experiment was conducted in a sound-attenuated room at Western Sydney University, using PsychoPy2 (Peirce, 2007), which recorded participants' responses and response times. Response times were measured from the end of the stimulus to the participants' mouse click.

## 2.4. Statistical analysis

All statistical analyses were performed in R (R Core Team, 2022). The *lme4* package (Bates et al., 2015) was used to build mixed statistical models, and the *lmerTest* package (Kuznetsova et al., 2017) was used to obtain *p*-values for the models. Details of the fixed effects are discussed along with the results in the following section. All models included random intercepts for listener (participant), speaker (of the stimuli), and stimulus word.

TABLE 3 GLMM analysis on effects of Japanese length and quality on AusE length categorization.

| | *β* | SE | *z* | *p*-value | |
|---|---|---|---|---|---|
| (Intercept) | −0.241 | 0.128 | −1.876 | 0.060 | . |
| long | 0.630 | 0.063 | 10.017 | $2^{-16}$ | *** |
| /i/ | 0.865 | 0.126 | 6.864 | $7^{-12}$ | *** |
| /e/ | 0.382 | 0.125 | 3.059 | 0.002 | ** |
| /a/ | −0.364 | 0.126 | −2.884 | 0.004 | ** |
| /o/ | −0.394 | 0.126 | −3.133 | 0.002 | ** |
| /u/ | −0.489 | 0.125 | −3.905 | $9^{-05}$ | *** |
| long:/i/ | 0.011 | 0.126 | 0.086 | 0.932 | |
| long:/e/ | 0.034 | 0.125 | 0.270 | 0.787 | |
| long:/a/ | 0.212 | 0.126 | 1.676 | 0.094 | . |
| long:/o/ | 0.072 | 0.126 | 0.569 | 0.569 | |
| long:/u/ | −0.328 | 0.125 | −2.618 | 0.009 | ** |

Baseline = grand mean (*** = 0.001, ** = 0.01, * = 0.05, . = 0.1).

# 3. Results

## 3.1. Overall categorization

Figure 1 presents the overall response patterns. The question going into the experiment was whether AusE listeners' categorization would be affected by Japanese vowel length and, if so, whether and how the effect would be related to Japanese vowel quality. To answer this question, both the Japanese target vowels and AusE response vowels were first collapsed by length ("long" vs. "short").[5] A generalized linear mixed model (GLMM) with a logit link function was then fitted using the *glmer()* function, with AusE vowel length (1 = "long," 0 = "short") as the outcome variable and Japanese vowel length ("long," "short"), Japanese vowel quality (/i, e, a, o, u/), and their interaction as the predictor variables. The predictors were coded with sum contrast coding so that each level of a variable is compared to the grand mean rather than a fixed reference level.

5 AusE /ɪə/ was coded as "long" for convenience.

Table 3 presents the results of the analysis. Note that the table shows the combined results of two models of the same structure but with different reference levels. This is because regression models do not return the coefficient of the reference level, and although the missing coefficient can be calculated by hand, its significance level is not examined, making a second model necessary (Clopper, 2013). Changing the reference level results in negligible changes in the coefficients for the non-reference levels, and thus for viewing convenience, Table 3 combines (a) the result of a model with Japanese /u/ as the reference level and (b) the result for Japanese /u/ obtained from a model with Japanese /i/ as the reference level.[6]

The main effect of Japanese length (i.e., "long") was statistically significant, suggesting that AusE listeners tended to choose "long" AusE response categories when the target Japanese vowel was phonologically long. The main effect of Japanese quality was also all significant, indicating that the likelihood of "long" AusE response categories being chosen differed according to the target Japanese quality. This is expected, as the number of AusE "long" and "short" vowels that correspond to a Japanese quality can vary depending on the quality, as can be seen in Figure 1. In contrast, the interaction between Japanese length and quality was significant only for /u/. This indicates that the effect of Japanese length on AusE listeners' categorization was generally independent of Japanese quality, except for /u/. The negative coefficient of the significant interaction implies that listeners tended to choose a "short" AusE vowel when the target Japanese vowel was long (i.e., /uu/).

## 3.2. By-vowel categorization

In order to explore the factors that drove the overall categorization patterns in more detail, we also performed by-vowel analyses, fitting GLMMs for AusE vowel responses that are the closest to Japanese vowels in terms of height, backness, roundedness, and length. To complement the analyses, acoustic data of the relevant AusE and Japanese vowels (Elvin et al., 2016; Yazawa and Kondo, 2019) are presented in Figures 2, 3, 4.

We start with high vowels, where the front vowels /ii, i/ showed a clear effect of duration on categorization but the back vowels /uu, u/ did not. In the case of Japanese /ii, i/, both vowels were categorized predominantly as their closest counterparts in AusE—high, front, unrounded, long/short—namely /iː/ and /ɪ/. AusE listeners showed a clear preference for AusE /iː/ when categorizing Japanese long /ii/ (61.0%), but were split between /iː, ɪ/ when categorizing Japanese short /i/ (40.8 and 43.2%, respectively). To test whether AusE listeners chose AusE categories that matched to Japanese vowels both in terms of length and quality, we first fitted a GLMM with logit link function to the whole data, with the rate of AusE /iː/ responses (1 = /iː/ chosen, 0 = /iː/ not chosen) as the outcome variable and Japanese vowel category (/ii, ee, aa, oo, uu, i, e, a, o, u/) as the predictor. Japanese /ii/ was set as the baseline, which was expected to have the highest AusE /iː/ response rates across all ten Japanese vowels. The analysis found that AusE listeners categorized Japanese /ii/ significantly more often as AusE /iː/ than Japanese /i/ ($\beta = -0.884, SE = 0.297, t = -2.974, p = 0.003$) as well as all other Japanese vowels ($ps < 0.001$). To assess whether a similar length-based difference is found for vowels categorized as AusE /ɪ/, another GLMM of the same structure was fitted for AusE /ɪ/ responses with Japanese /i/ as the baseline. The results showed that AusE listeners categorized Japanese /i/ significantly more often as AusE /ɪ/ than both Japanese /ii/ ($\beta = -1.609, SE = 0.411, t = -3.910, p < 0.001$) and all other vowels ($ps < 0.001$), with the exception of /e/ ($\beta = 0.367, SE = 0.402, t = -0.913, p = 0.361$). Since spectral differences between Japanese /ii, i/ are negligible, the two vowels should show similar categorization patterns if AusE listeners were relying primarily on quality, much like AmE listeners in Nishi et al. (2008). This is clearly not the case, where instead AusE listeners make use of the longness of Japanese /ii/ to categorize it as AusE /iː/. On the other hand, the shortness of Japanese /i/ does not seem to have an effect.

Similar to Japanese /ii, i/, AusE listeners also categorized Japanese /uu, u/ predominantly as their closest AusE vowels in terms of height and backness—high, non-front—namely AusE /ʉː, ʊ/, respectively. However, while duration did have a significant effect on how the Japanese vowels were categorized, the effect was relatively small compared to Japanese /ii, i/ in that both Japanese /uu, u/ were categorized most often as AusE /ʊ/, a short vowel. As with the responses to Japanese /ii, i/, we fitted two GLMMs, one for the rate of AusE /ʉː/ responses and another for /ʊ/ responses, to test whether AusE listeners chose AusE categories that matched to Japanese vowels in terms of length. The model for AusE /ʉː/ responses indeed showed that Japanese /uu/ was categorized more often as AusE /ʉː/ than Japanese /u/ ($\beta = -0.679, SE = 0.218, t = -3.117, p = 0.002$) and all other vowels ($ps < 0.001$). However, there was no significant difference between Japanese /uu/ and /u/ among vowels categorized as AusE /ʊ/ ($\beta = 0.026, SE = 0.200, t = 0.131, p = 0.896$), showing that both long and short Japanese vowels were equally likely to be categorized as AusE /ʊ/. A possible factor driving the observed pattern is roundedness. According to Harrington et al. (1997) and Cox (2006), AusE /ʉː/ often exhibits onglide with lowering third formant (F3), i.e., increased lip rounding toward the target. This is shown in Figure 4, where AusE /ʉː/ shows unusually low F3 that is much lower than that of AusE /ʊ/ and Japanese /uu, u/. Since both AusE /ʊ/ and Japanese /uu, u/ lack such articulation, AusE listeners might have prioritized the lack of lowered F3 itself as a cue for AusE /ʊ/ over the long duration as a cue for AusE /ʉː/.[7]

The remaining Japanese vowels all patterned more closely with Japanese /ii, i/ than with /uu, u/ in that long Japanese vowels clearly led to more categorizations as long AusE vowels. In the case of Japanese /ee, e/, AusE listeners categorized /ee/ as a long AusE vowel 68.2% of the time and /e/ as short 61.8% of the time. GLMMs were fitted for AusE /eː, e/ responses, vowels closest to Japanese /ee, e/ in terms of height, backness, and roundedness. The model for AusE /eː/ responses showed that Japanese /ee/ was significantly more likely to be categorized as the AusE vowel than

6 The main effect and interactions of "short" Japanese length are omitted from the table because the estimate magnitudes of a binary factor are identical with just the signs reversed when sum contrast coded.

7 While AusE listeners have been shown to be sensitive to such dynamic spectral cues (Williams et al., 2018), note that the Japanese stimuli in the current experiment exhibited very little VISC (Yazawa and Kondo, 2019).

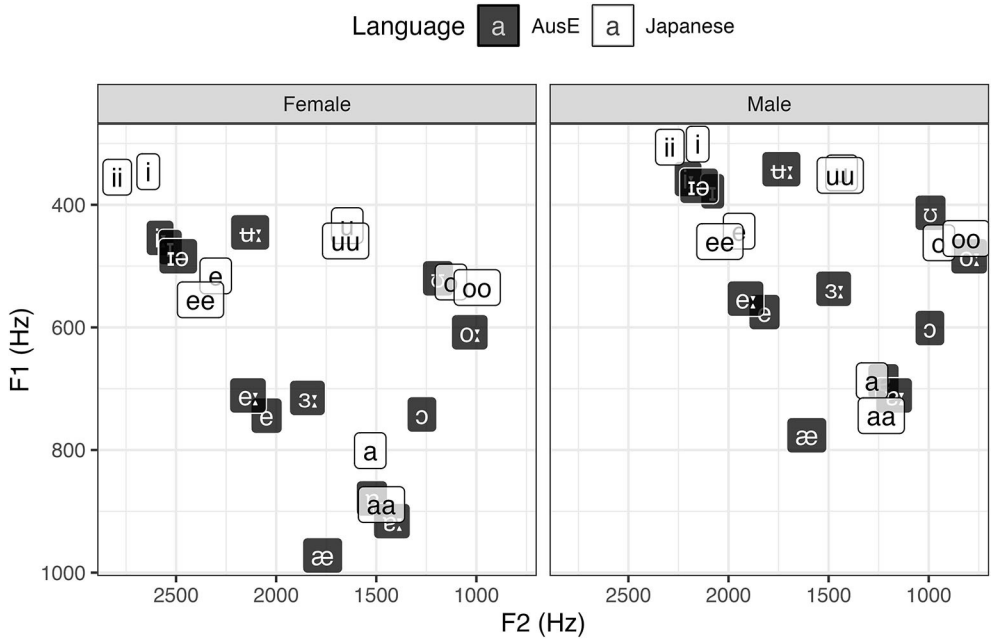



FIGURE 2
Average F1 and F2 of AusE and Japanese vowels. Adapted from Elvin et al. (2016) and Yazawa and Kondo (2019).

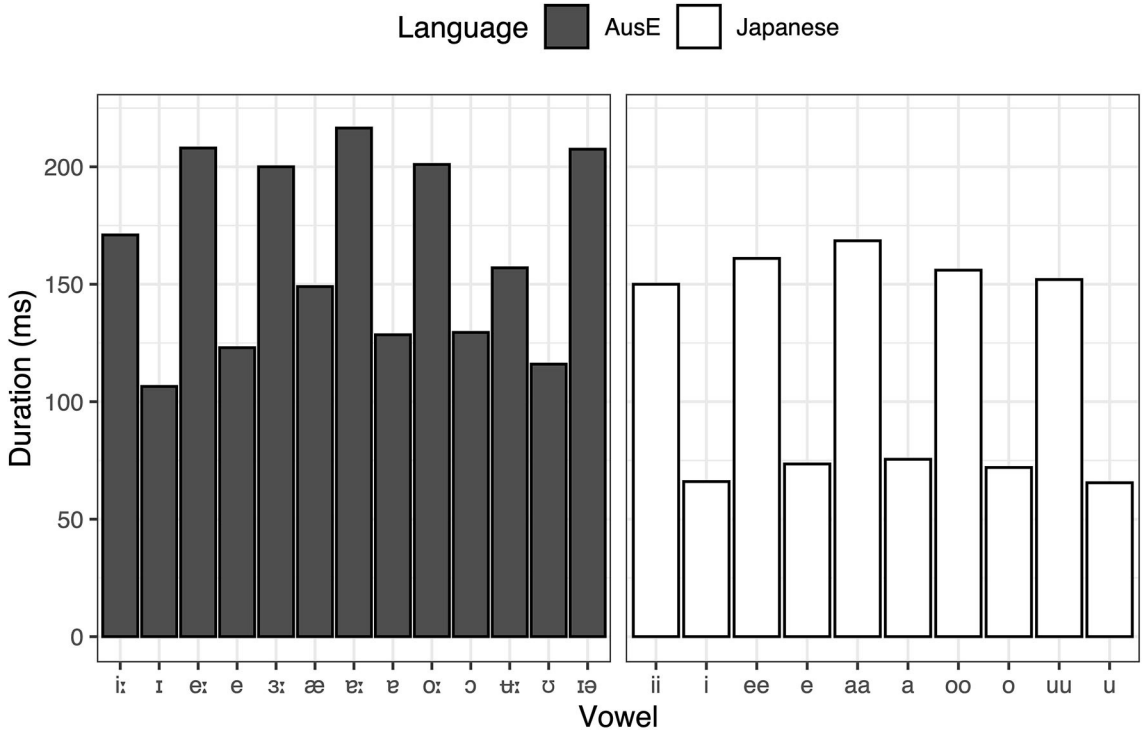



FIGURE 3
Average duration of AusE and Japanese vowels (male and female means collapsed). Adapted from Elvin et al. (2016) and Yazawa and Kondo (2019).

all other Japanese vowels ($ps < 0.001$). Likewise, the model for AusE /e/ responses also showed that Japanese /e/ was significantly more likely to be categorized as the AusE vowel than all other Japanese vowels ($ps < 0.001$), except for /ee/ ($\beta = -0.070, SE = 0.193, t = -0.361, p = 0.718$). One thing to note is that both Japanese /ee, e/ were categorized persistently as AusE /e/, a short vowel, at rates of 18.0 and 19.0%, respectively, and AusE /iː/, a long vowel, at rates of 14.5 and 20.3%, respectively. The persistence

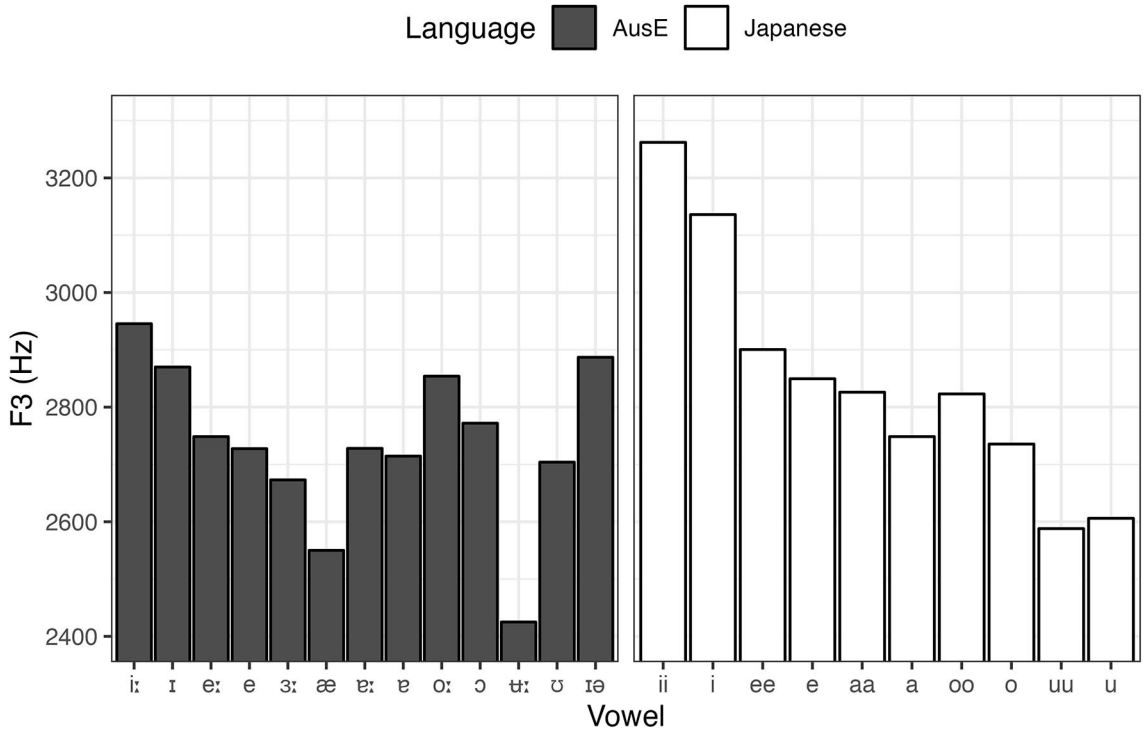



FIGURE 4
Average F3 of AusE and Japanese vowels (male and female means collapsed). Adapted from Elvin et al. (2016) and Yazawa and Kondo (2019).

of /e, iː/ categorizations for both vowels seems to reflect listeners' uncertainty regarding the spectral quality of Japanese /ee, e/, which lie between the AusE high and mid front vowels (Figure 2). This explains why the GLMM for AusE /ɪ/ responses did not yield a significant difference between Japanese /i/ and /e/ as mentioned above. Therefore, the assumption that Japanese /ee, e/ should be categorized predominantly as AusE /eː, e/, which contrast primarily in duration unlike most other AusE vowels and may thus elicit an elevated duration effect, may not necessarily hold. Despite the increased variability in categorization, however, the effect of length largely parallels the pattern observed with Japanese /ii, i/; AusE listeners use the longness of Japanese /ee/ to categorize it as an AusE long vowel and the shortness of Japanese /e/ to categorize it as an AusE short vowel.

For the low Japanese vowels, the most frequent response category for Japanese long /aa/ was AusE /ɐː/ (38.3%) and for Japanese short /a/ was AusE /ɐ/ (45.0%), showing again a duration-based preference according to height, backness, and roundedness. However, like AusE /eː, e/ discussed above, AusE /ɐː, ɐ/ are spectrally overlapping and thus contrast primarily in duration, which admittedly makes the duration effect in the categorization of Japanese /aa, a/ seem weaker than expected (since the effect is comparable to that in the categorization of Japanese /ii, i/ as AusE /iː, ɪ/ which contrast both spectrally and durationally). This is perhaps due to the fact that Japanese /aa/ is durationally ambiguous between AusE /ɐː/ and /ɐ/, as can be seen in Figure 3. This ambiguity effect is actually reflected by the fact that the second-most frequent categorization of Japanese /aa/ was AusE /ɐ/ (20.2%). Despite the seemingly weak duration effect, the results of GLMMs nonetheless showed that Japanese /aa/ and /a/ were significantly more likely to elicit AusE /ɐː/ and /ɐ/ responses, respectively, than all other Japanese vowels ($ps < 0.001$).

Lastly, the categorization pattern is similar with Japanese /oo, o/, where the vowels were significantly more likely to elicit AusE /oː/ and /ɔ/ responses, respectively, according to GLMMs ($ps < 0.001$). However, both Japanese /oo, o/ also showed rather persistent categorizations as AusE /ʊ/ regardless of length (25.5 and 22.7%, respectively). This is most likely due to the spectral uncertainty of Japanese back /oo, o/ as either AusE high back /ʊ/ or non-high back /oː/. It is noteworthy, therefore, that Japanese /o/ was most often categorized as AusE /ɔ/ despite the spectral mismatch, suggesting that durational similarity (i.e., shortness) was prioritized over spectral similarity (i.e., height).

## 3.3. Response time

The response time data were analyzed with a linear mixed-effects model (LME). The model was fitted using the *lmer()* function, with response time (in seconds) as the outcome variable and Japanese vowel category (/ii, ee, aa, oo, uu, i, e, a, o, u/) as the predictor variable. The predictor was again sum contrast coded to set the baseline of the model as the grand mean. The results are presented in Table 4, which combines (c) the result of a model with Japanese /u/ as the reference level and (d) the result for Japanese /u/ obtained from a model with Japanese /ii/ as the reference level, for the same reason as stated in Section 3.1.

TABLE 4 LME analysis comparing response times by Japanese vowel category.

| | $\beta$ | SE | $z$ | $p$-value | |
|---|---|---|---|---|---|
| (Intercept) | 3.166 | 0.224 | 14.110 | $4.52^{-12}$ | *** |
| /ii/ | −0.388 | 0.095 | −4.063 | $4.91^{-05}$ | *** |
| /i/ | −0.303 | 0.095 | −3.177 | 0.001 | ** |
| /ee/ | 0.257 | 0.095 | 2.696 | 0.007 | ** |
| /e/ | 0.400 | 0.095 | 4.193 | $2.79^{-05}$ | *** |
| /aa/ | 0.166 | 0.095 | 1.740 | 0.082 | . |
| /a/ | −0.057 | 0.095 | −0.594 | 0.553 | |
| /oo/ | 0.292 | 0.095 | 3.057 | 0.002 | ** |
| /o/ | −0.183 | 0.095 | −1.919 | 0.055 | . |
| /uu/ | −0.121 | 0.095 | −1.269 | 0.205 | |
| /u/ | −0.063 | 0.095 | −0.664 | 0.506 | |

Baseline = grand mean (*** = 0.001, ** = 0.01, * = 0.05, . = 0.1).

The results show that listeners took the shortest to categorize Japanese /ii/ at 2.778 s and Japanese /i/ at 2.863 s, which are both significantly shorter than the grand mean of 3.166 s. This suggests that the Japanese /i/ quality was relatively easy to categorize, probably because it is unambiguously high and front. In contrast, listeners took significantly longer than the grand mean to categorize Japanese /e/ at 3.566 s and Japanese /ee/ at 3.423 s. This suggests that the Japanese /e/ quality was generally difficult to categorize, most likely due to its spectral ambiguity as discussed earlier. Another Japanese vowel that took significantly longer than the grand mean was /oo/, probably due to its ambiguous quality between AusE /oː/ and /ʊ/. It is then worth noting that the response time for Japanese /o/ was marginally shorter than the grand mean (−0.183 s, $p = 0.055$), as it implies that Japanese /o/ was less ambiguous than Japanese /oo/ despite both vowels being spectrally alike, suggesting that the short duration of /o/ outweighs the spectral ambiguity.

One additional factor that is relevant to the response time data is potential lexical effects. While listeners were instructed that the stimuli were not English words, some of the Japanese tokens (e.g., /biip/) that resemble a real English word (e.g., *beep*) may have implicitly activated AusE lexical knowledge. Since listeners used word choices (e.g., *heed*) to respond, such tokens may have been processed faster than other tokens with no corresponding English word (e.g., /gaak/). To test this possibility, listeners' responses were coded as either "lexical" or "non-lexical," where "lexical" responses have a corresponding AusE lexical item. For example, if a listener chose AusE /iː/ when the target stimulus's consonantal context was /bVp/, the response was coded as "lexical" because the perceived form /biːp/ corresponds to a real AusE word *beep*. Other cases of "lexical" responses were: /ɔ/ responses to /bVp/ stimuli (i.e., *bop*), /e, ɐː, ɔ/ responses to /dVt/ stimuli (i.e., *debt, dart, dot*, respectively), and /iː, oː, ʉː/ responses for /gVk/ stimuli (i.e., *geek, gawk, gook*, respectively). The remaining responses were coded as "non-lexical," which accounted for 72.8% of all responses (4,370 of 6,000).

Adding this variable of lexicality with sum contrast coding to the aforementioned LME model significantly improved the model fit according to a likelihood ratio test [$\chi^2(1) = 14.585, p < 0.001$]. The resultant model found a significantly shorter response time for "lexical" responses than "non-lexical" ones ($\beta = -0.136, SE = 0.037, t = -3.636, p < 0.001$). Thus, it is speculated that AusE listeners recognized English words in some of the Japanese tokens, which were processed faster than the other tokens without any lexical reference. Yet, further addition of the interaction of lexicality and Japanese vowel category did not improve the model fit [$\chi^2(9) = 10.361, p = 0.322$], meaning that there were no by-category differences in the shortening effect of lexicality on response time.[8]

# 4. Discussion

## 4.1. Summary of the results

The first purpose of this study was to evaluate whether vowel duration is used phonologically in AusE, unlike most other varieties of English. Previous research had shown that AusE listeners are sensitive to acoustic-phonetic changes in vowel duration (Tsukada, 2012), but it was unclear whether they would actively utilize the duration cue for their native vowel identity. The current study therefore examined AusE listeners' categorization of Japanese long and short vowel pairs, each of which differs systematically in duration but minimally in spectral quality (cf. Figures 2, 3, 4). The analysis found a general tendency for Japanese long and short vowels to be categorized as AusE long and short vowels, respectively (Figure 1 and Table 2), indicating that vowel duration does play an important role in AusE phonology. The result contrasts with previously reported AmE listeners' categorization of the same Japanese vowels (Nishi et al., 2008), which was largely unaffected by length (Table 1).

The second purpose was to test whether the above effect of duration on AusE listeners' vowel categorization would be specific to certain Japanese categories or generalized across the board. Given that only a subset of AusE vowels contrast in duration alone (/eː/-/e/ and /ɐː/-/ɐ/) while others contrast in both duration and spectra (e.g., /iː/-/ɪ/ and /oː/-/ɔ/), AusE listeners may use duration more readily for Japanese vowels that spectrally match the former categories (e.g., /ee/-/e/ and /aa/-/a/) than those matching the latter (e.g., /ii/-/i/ and /oo/-/o/). However, the AusE listeners in the current study seem to have utilized duration for both cases regardless of spectral ambiguity (Table 4), suggesting that the effects of length and quality were generally independent of each other, with a notable exception of /uu/-/u/ (Table 3). These results have important theoretical and pedagogical implications, as discussed below.

8 Other studies have also found no item effect when nonwords that sound like real words are used in tasks that combine categorization and lexical learning (Escudero et al., 2022, p. 4).

## 4.2. Theoretical implications

As outlined in Section 1, the segment- and feature-based frameworks of speech perception predicted different results for the current experiment. On the one hand, the segment-based view predicted that the effect of duration should be stronger for certain Japanese qualities (i.e., /e, a/) than the others (i.e., /i, o, u/), as the reliance on duration cues should be specific to each native segmental category that nonnative sounds are categorized as. On the other hand, the feature-based view predicted a uniform effect of duration across all Japanese categories, assuming that a length feature plays a role in the whole phonological system. The GLMM analysis in Table 3 suggests that the observed perceptual patterns align better, but not perfectly, with the feature-based view. Vowel length had an independent effect from vowel quality, where long Japanese vowels tended to be categorized as long AusE vowels despite mismatches in quality. In this respect, the categorization tendency was largely the same between AusE vowels that contrast in both duration and spectra (e.g., /iː, ɪ/) and those that contrast exclusively in duration (e.g., /ɐː, ɐ/). The only exception was Japanese /uu, u/, which were consistently categorized as short AusE /ʊ/.

Importantly, this kind of generalization of native length to nonnative perception has been observed in other previous studies as well. Returning to Tsukada (2012)'s study, Arabic differs from Japanese in lacking the /e/ and /o/ qualities and, according to the segment-based view, native Arabic listeners should be less accurate in discriminating the length of Japanese /ee/-/e/ and /oo/-/o/ (absent segments) than /ii/-/i/, /aa/-/a/, and /uu/-/u/ (present segments). The result contrarily showed no difference in discrimination ability between present and absent qualities, which is in line with the feature-based view. The segment-based view would also have difficulty in explaining the link between vowel and consonant length found in McAllister et al. (2002) and Pajak and Levy (2014) because it would be implausible for consonant categories to assimilate to vowel categories and vice versa. Moreover, Tsukada et al. (2018) found that both L1 AusE and L1 Korean learners of L2 Japanese were generally accurate in identifying the consonantal length of Japanese as well as Italian (>80%), despite the fact that AusE does not have a singleton-geminate contrast while Korean does. This would support the view that AusE does have a vowel length feature that can transfer or extend to nonnative consonant length perception.

One caveat with the feature-based approach is that the property of the "same" feature can vary from language to language, despite the traditional belief that features are language-universal. As for the vowel length feature, what is "long" in one language is not necessarily also "long" in another language and vice versa, as the actual duration of "long" vowels can differ substantially across languages. This can be seen in Figure 3, where Japanese vowels are shorter in duration than AusE vowels in general. It follows that some tokens of "long" Japanese vowels are not sufficiently long in duration to be categorized as "long" in AusE, which likely affected the categorization patterns shown in Figure 1.[9] This explanation would align with a recent proposal that features are substance-free and emergent (Boersma et al., 2022); there is no innate phonological substance of absolute "longness" in the mind, and listeners rather learn to interpret what is meaningfully long or short in the given language based on the available linguistic input.

The exceptional categorization pattern of Japanese /uu, u/ by AusE listeners, however, poses a challenge to the feature-based view. As mentioned earlier, the result can only be explained by referring to the F3, an acoustic cue for lip rounding. One may thus hypothesize that a roundedness feature was contributing somehow, although it would still be unclear why only this feature suppressed the effect of length while other features such as height and backness did not. A possible reason lies in the multiplicity of acoustic cues or the lack thereof in the given features. While height, backness, and length features are considered to have only one corresponding acoustic cue (i.e., F1, F2, and duration, respectively), the roundedness feature is associated with multiple acoustic cues (i.e., F2 and F3). Llompart and Reinisch (2018) found that effects of selective adaptation on German vowel contrasts generalized for contrasts differing in height (F1) and those differing in backness (F2) but not for those differing in tenseness (F1, F2, and duration),[10] suggesting that acoustically complex features such as tenseness and roundedness may behave differently from acoustically simple features such as height, backness, and length in vowel perception.[11] Assuming that perceptual input is gradually abstracted and integrated into higher-level representations (Greenberg and Christiansen, 2019), it may be the case that acoustically complex features outweigh lower-level, acoustically simple features, which may explain why Japanese /uu/ without strong lip rounding would not be categorized as AusE /ʉː/ despite their similar durations.

Finally, it should be noted that the dichotomy of segment- vs. feature-based views is not an absolute one. Incorporation of these two approaches is possible, as indicated in the above explanation of gradual abstraction and integration of cues to segments via features. An example of such integration comes from the Second Language

9 Hirata and Lambacher (2004) found that Japanese listeners can misidentify a short vowel produced at a slow speaking rate as being long because its duration can overlap with that of a long vowel produced at a fast rate. Thus, speech rate is also relevant to the property of "long" and "short" vowels within a language. The current experiment used stimuli at a normal speaking rate to minimize potential rate effects.

10 The result needs to be interpreted with caution because selective adaptation effects did generalize for lax vowels; it was only for tense vowels that the effects did not generalize. See also Boersma et al. (2022, p. 664) for a possibility that "tense" and "lax" features do not have any phonetic correlate to begin with.

11 While Llompart and Reinisch (2018) advocate that acoustic cues rather than phonological features drive vowel perception, they also note that the notion of features would be compatible if they are acoustically or phonetically defined. We agree that phonological features should have a phonetic basis (Boersma and Chládková, 2011; Mesgarani et al., 2014; Chládková et al., 2015a), but would also put forward that acoustic cues cannot be substituted for such phonetically grounded phonological features. This is because a purely cue-based approach would not suffice to explain the current result and other perceptual phenomena that can be readily explained by assuming features. See Boersma and Chládková (2011) for further discussion.

Linguistic Perception (L2LP) model (Escudero and Yazawa, in press; Escudero, 2005; van Leussen and Escudero, 2015), which defines speech perception as the mapping of acoustic cues onto a linguistic representation. While the majority of studies conducted within L2LP have assumed segmental categories as the fundamental unit of perception, some studies have referred to other units including features. For example, Escudero and Boersma (2004) demonstrated that L1 Spanish listeners' over-reliance on duration in perceiving the /iː/-/ɪ/ in L2 Southern British English (SBE) can be accurately modeled by assuming that the SBE vowels are represented as /i, long/ and /i, short/ in the learners' phonological grammar, i.e., addition of a new length feature to an existing segmental category. Yazawa (2020) also proposed that Japanese listeners' perception of AmE /æ/ as a deviant, non-prototypical exemplar of Japanese /a/ or possibly /e/ (Strange et al., 1998; Shinohara et al., 2019, 2022) can be explained as a result of mismatch in height and frontness features, i.e., AmE /low, front/ (/æ/) is too front to be Japanese /low, central/ (/a/) and too low to be Japanese /mid, front/ (/e/). An important side note on these studies is that the learners' target variety (SBE or AmE) was explicitly specified, as is proposed within the L2LP model, which is essential for making accurate predictions and explanations regarding cross-linguistic perception patterns (cf. Section 4.4).

## 4.3. Pedagogical implications

Some pedagogical implications for English listeners' learning of nonnative length arise from the above discussion on features. According to the "feature" hypothesis, nonnative length contrasts would be less of a challenge for AusE listeners who has access to a vowel length feature than for AmE speakers who do not.[12] This prediction has been attested in previous studies showing that monolingual AusE listeners could already discriminate Japanese vowel length well (75% accurate; Tsukada, 2012) while AmE listeners prior to perceptual training identified Japanese vowel length poorly (39% accurate; Hirata, 2004). However, the presence of a vowel length feature in AusE does not guarantee immediate and successful learning of nonnative length contrasts because, as discussed earlier, the acoustic property of a feature is likely language-specific. For example, "long" Japanese vowels that are about 150 ms long can be ambiguous between "long" (200 ms) and "short" (100 ms) for AusE listeners, resulting in occasional misperception of "long" as "short." This kind of mismatch in featural properties would explain, at least in part, why nonnatives perform consistently worse than natives in vowel and consonant length perception even when they share the "same" feature of length (Tsukada, 2012; Tsukada et al., 2018). The learning task for AusE learners of Japanese, therefore, is to shift the boundary between "long" and "short" vowels to match that of Japanese (which L2LP calls a "perceptual task"). On the other hand, AmE learners of Japanese have an additional task to establish a new length feature in their phonological grammar (a "representational task" in L2LP's term), similar to Spanish learners of SBE as mentioned above. Thus, the presence or absence of a vowel length feature in the two varieties of English leads to different kinds of learning tasks.

The remaining question, then, is how the learning process can be facilitated in such a way that is appropriate for each language variety (Elvin and Escudero, 2019). On the one hand, AusE learners of Japanese may be able to shift their perceptual boundary via simple distributional learning (i.e., abundant exposure to Japanese long and short vowels), perhaps aided by artificially enhanced durational distributions (Escudero et al., 2011), to achieve immediate and long-lasting learning effects (Escudero and Williams, 2014). On the other hand, AmE learners of Japanese may need to be directed to the presence of vowel length more explicitly, as acquiring a new feature seems more problematic than shifting an existing boundary (Chládková et al., 2022). Hirata (2004)'s success in training AmE listeners on Japanese length contrasts may be attributed to the unique training procedure, where AmE participants were instructed to count the number of morae in each training token, e.g., /ii/ "good" (= 2 morae) and /i/ "stomach" (= 1 mora). The participants were thus made aware of Japanese length throughout the training period of 3.5 weeks, potentially resulting in efficient and robust learning. Moreover, the training also involved consonant length, e.g., /kata/ "shoulder" (= 2 morae) and /katta/ "won" (= 3 morae). Given that vowel and consonant length seems interrelated, perhaps the training on vowels and consonants interacted with each other, further enhancing the learning efficacy.[13] This kind of explicit training could be useful for teaching nonnative length to native listeners of AmE and other varieties of English. What still needs testing is whether the learning task of boundary shift for AusE listeners is really as easy as expected and, if so, how long and to what extent the learning effect can be maintained. To this end, implicit training paradigms such as cross-situational word learning (CSWL) can be useful (Escudero et al., 2022, 2023; Escudero and Yazawa, in press).

## 4.4. Future directions

The current study has demonstrated that AusE listeners systematically utilize duration cues for vowel identity, despite the common belief that length is not phonemic in English. The result contrasts with previous studies on AmE listeners (Hillenbrand et al., 2000; McAllister et al., 2002; Hirata, 2004; Dietrich et al., 2007; Kondaurova and Francis, 2008; Mugitani et al., 2009), especially that of Nishi et al. (2008) who examined the categorization of Japanese long and short vowels. However, the current result cannot be directly compared with that of Nishi et al. (2008) due to methodological differences. For example, while Nishi et al. (2008) used /hVba/ disyllables as stimuli, the current study used varying consonantal contexts (/bVp, gVk, dVt/), but without /hVb/.

12 While the original SLM (Flege, 1995) was in favor of the "feature" hypothesis, its recent revision (SLM-r; Flege and Bohn, 2021) has replaced it with the "full access" hypothesis, which claims that L2 learners can gain full access to non-L1 features.

13 Unlike Hirata (2004), Tajima et al. (2008) found limited effects of training on Canadian English speakers' perception of Japanese vowel and consonant length. This can be due to the differences in training procedure (word identification vs. mora counting) and training period (5 days vs. 3.5 weeks) between the two studies.

Based on the large body of previous research reviewed earlier, we can assume with some confidence that AmE listeners would show a similar perceptual pattern to Table 1 with our stimuli and procedure, but an additional parallel data collection in the US would be ideal to allow for a more direct comparison.

The results of the current study also highlight the necessity to further investigate non-AmE varieties of English. Although Karpinska et al. (2015) found similar perceptual trends in high front vowel identification across several varieties of English (except for AusE), these varieties may actually show some variability in perceptual patterns. For example, Escudero and Boersma (2004) showed that SBE listeners rely systematically more on duration than Scottish English listeners when perceiving synthetic high front tense and lax vowels (commonly referred to as "long" and "short" vowels in British English). Moreover, it is unclear whether similar perceptual tendencies would be observed for non-high-front vowels as well, as AmE listeners' reliance on duration seems to somewhat differ between high and non-high vowel contrasts (Hillenbrand et al., 2000), which is reflected in their perception of nonnative Japanese long and short vowels (Nishi et al., 2008). Thus, cross-examining listeners of different varieties of English in their categorization of Japanese length in future studies—preferably with AmE and AusE as references—would shed further light on whether and to what extent the segment- and feature-based approaches are capable of explaining cross-variety similarities and differences.

## 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 present study involving human participants was part of a larger study reviewed and approved by Western Sydney University Human Research Ethics Committee (approval number: H11022). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

## Author contributions

KY and PE contributed to the conception and design of the study. KY collected the production data and wrote the first draft of the manuscript. JW supervised the collection of the perception data and was responsible for the data analysis. JW and PE wrote parts of the manuscript. All authors contributed to the article and approved the submitted version.

## Funding

KY's work, visit to the MARCS Institute for Brain Behaviour and Development, and data collection were supported by JSPS Overseas Challenge Program for Young Researchers (grant number: 201708171), the Australian Research Council (ARC) Centre of Excellence for the Dynamics of Language where PE was Chief Investigator (grant number: CE140100041), and PE's ARC Future Fellowship grant (grant number: FT160100514). JW and PE work was supported by the two ARC grants. Article Processing Charges were covered by JSPS Grant-in-Aid for Scientific Research (grant number: 21H00533) and the Support Program for Publication of Research Results in English, University of Tsukuba.

## Acknowledgments

The authors would like to thank Jaydene Elvin for providing the AusE acoustic data and Nicole Traynor for helping with the perception data collection.

## 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

Asu, E. L., and Teras, P. (2009). Estonian. *J. Int. Phon. Assoc.* 39, 367–372. doi: 10.1017/S002510030999017X

Bates, D., Mächler, M., Bolker, B. M., and Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Best, C. T. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed W. Strange (Timonium, MD: York Press), 171–204.

Best, C. T., and Tyler, M. (2007). "Nonnative and second-language speech perception: commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, eds O.-S. Bohn, and M. J. Munro (Amsterdam: John Benjamins), 13–34. doi: 10.1075/lllt.17.07bes

Boersma, P., and Weenink, D. (2022). *Praat: Doing Phonetics by Computer*. Amsterdam: University of Amsterdam.

Boersma, P., and Chládková, K. (2011). "Asymmetries between speech perception and production reveal phonological structure," in *Proceedings of the 17th International*

*Congress of Phonetic Sciences*, eds W.-S. Lee, and E. Zee (Hong Kong: The University of Hong Kong), 328–331.

Boersma, P., Chládková, K., and Benders, T. (2022). Phonological features emerge substance-freely from the phonetics and the morphology. *Can. J. Linguist.* 67, 611–669. doi: 10.1017/cnj.2022.39

Chen, H., Rattanasone, N. X., and Cox, F. (2014). "Perception and production of phonemic vowel length in Australian English-learning 18-month-olds," in *Proceedings of the 14th Conference on Laboratory Phonology* (Tokyo: National Institute for Japanese Language and Linguistics), 42.

Chládková, K., Boersma, P., and Benders, T. (2015a). "The perceptual basis of the feature vowel height," in *Proceedings of the 18th International Congress of Phonetic Sciences*, eds The Scottish Consortium for ICPhS 2015 (Glasgow: The University of Glasgow), 711.

Chládková, K., Boersma, P., and Escudero, P. (2022). Unattended distributional training can shift phoneme boundaries. *Biling. Lang. Cogn.* 25, 1–14. doi: 10.1017/S1366728922000086

Chládková, K., Escudero, P., and Lipski, S. C. (2013). Pre-attentive sensitivity to vowel duration reveals native phonology and predicts learning of second-language sounds. *Brain Lang.* 126, 243–252. doi: 10.1016/j.bandl.2013.05.020

Chládková, K., Escudero, P., and Lipski, S. C. (2015b). When "AA" is long but "A" is not short: speakers who distinguish short and long vowels in production do not necessarily encode a short long contrast in their phonological lexicon. *Front. Psychol.* 6, 438. doi: 10.3389/fpsyg.2015.00438

Clopper, C. G. (2013). "Modeling multi-level factors using linear mixed effects," in *Proceedings of Meetings on Acoustics*, Volume 19 (Melville, NY: The Acoustical Society of America), 060028. doi: 10.1121/1.4799729

Cox, F. (2006). The Acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Aust. J. Linguist.* 26, 147–179. doi: 10.1080/07268600600885494

Cox, F., and Palethorpe, S. (2007). Australian English. *J. Int. Phon. Assoc.* 37, 341–350. doi: 10.1017/S0025100307003192

Dietrich, C., Swingley, D., and Werker, J. F. (2007). Native language governs interpretation of salient speech sound differences at 18 months. *Proc. Nat. Acad. Sci.* 104, 16027–16031. doi: 10.1073/pnas.0705270104

Elvin, J., and Escudero, P. (2019). "Cross-linguistic influence in second language speech: implications for learning and teaching," in *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*, eds M. J. Gutierrez-Mangado, M. Martínez-Adrián, and F. Gallardo-del-Puerto (New York, NY: Springer International Publishing), 1–20.

Elvin, J., Williams, D., and Escudero, P. (2016). Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *J. Acoust. Soc. Am.*, 140, 576–581. doi: 10.1121/1.4952387

Escudero, P. (2005). *Linguistic Perception and Second Language Acquisition: Explaining the Attainment of Optimal Phonological Categorization* [PhD thesis]. Utercht: Utercht University.

Escudero, P., Benders, T., and Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: the case of Dutch, German, and Spanish listeners. *J. Phon.* 37, 452–465. doi: 10.1016/j.wocn.2009.07.006

Escudero, P., Benders, T., and Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *J. Acoust. Soc. Am.* 130, EL206–EL212. doi: 10.1121/1.3629144

Escudero, P., and Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Stud. Second Lang. Acquis.* 26, 551–585. doi: 10.1017/S0272263104040021

Escudero, P., Simon, E., and Mitterer, H. (2012). The perception of English front vowels by North Holland and Flemish listeners: acoustic similarity predicts and explains cross-linguistic and L2 perception. *J. Phon.* 40, 280–288. doi: 10.1016/j.wocn.2011.11.004

Escudero, P., Smit, E. A., and Angwin, A. J. (2023). Investigating orthographic versus auditory cross-situational word learning with online and laboratory-based testing. *Lang. Learn.* 73, 543–577. doi: 10.1111/lang.12550

Escudero, P., Smit, E. A., and Mulak, K. E. (2022). Explaining L2 lexical learning in multiple scenarios: cross-situational word learning in L1 Mandarin L2 English speakers. *Brain Sci.* 12, 1618. doi: 10.3390/brainsci12121618

Escudero, P., and Williams, D. (2012). Native dialect influences second-language vowel perception: Peruvian versus Iberian Spanish learners of Dutch. *J. Acoust. Soc. Am.* 131, EL406–EL412. doi: 10.1121/1.3701708

Escudero, P., and Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition* 133, 408–413. doi: 10.1016/j.cognition.2014.07.002

Escudero, P., and Yazawa, K. (in press). "The Second Language Linguistic Perception Model (L2LP)," in *The Cambridge Handbook of Bilingual Phonetics and Phonology*, ed M. Amengual (Cambridge: Cambridge University Press).

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

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. doi: 10.1017/9781108886901.002

Greenberg, S., and Christiansen, T. U. (2019). The perceptual flow of phonetic information. *Atten. Percept. Psychophys.* 81, 884–896. doi: 10.3758/s13414-019-01666-y

Gussenhoven, C., and Jacobs, H. (2017). *Understanding Phonology*. London: Routledge. doi: 10.4324/9781315267982

Harrington, J., Cox, F., and Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Aust. J. Linguist.* 17, 155–184. doi: 10.1080/07268609708599550

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (2000). Some effects of duration on vowel recognition. *J. Acoust. Soc. Am.* 108, 3013–3022. doi: 10.1121/1.1323463

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *J. Acoust. Soc. Am.* 116, 2384–2394. doi: 10.1121/1.1783351

Hirata, Y., and Lambacher, S. G. (2004). Role of word-external contexts in native speakers' identification of vowel length in Japanese. *Phonetica* 61, 177–200. doi: 10.1159/000084157

International Phonetic Association (1999). *Handbook of the International Phonetic Association: A Guide to Use the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Jacquemot, C., and Pallier, C. LeBihan, D., Dehaene, S., Dupoux, E. (2003). Phonological grammar shapes the auditory cortex: a functional magnetic resonance imaging study. *J. Neurosci.* 23, 9541–9546. doi: 10.1523/JNEUROSCI.23-29-09541.2003

Karpinska, M., Uchida, S., and Grenon, I. (2015). "Vowel perception by listeners from different English dialects," in *Proceedings of the 18th International Congress of Phonetic Sciences* (Glasgow: The University of Glasgow), 264.

Kawahara, S. (2015). "The phonology of Japanese accent," in *Handbook of Japanese Phonetics and Phonology*, ed. H. Kubozono (Berlin: De Gruyter Mouton), 445–492. doi: 10.1515/9781614511984.445

Keating, P. A., and Huffman, M. K. (1984). Vowel variation in Japanese. *Phonetica* 41, 191–207. doi: 10.1159/000261726

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. doi: 10.1121/1.2999341

Kozasa, T. (2004). "The interaction of duration and pitch in Japanese long vowels," in *Proceedings of BLS 30: General Session and Parasession on Conceptual Structure and Cognition in Grammatical Theory* (Berkeley, CA), 211–222. doi: 10.3765/bls.v30i1.960

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

Lipski, S. C., Escudero, P., and Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: an event-related potential study. *Psychophysiology* 49, 638–650. doi: 10.1111/j.1469-8986.2011.01347.x

Llompart, M., and Reinisch, E. (2018). Acoustic cues, not phonological features, drive vowel perception: evidence from height, position and tenseness contrasts in German vowels. *J. Phon.* 67, 34–48. doi: 10.1016/j.wocn.2017.12.001

Mazuka, R., Cao, Y., Dupoux, E., and Christophe, A. (2011). The development of a phonological illusion: a cross-linguistic study with Japanese and French infants. *Dev. Sci.* 14, 693–699. doi: 10.1111/j.1467-7687.2010.01015.x

McAllister, R., Flege, J. E., and Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *J. Phon.* 30, 229–258. doi: 10.1006/jpho.2002.0174

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science* 343, 1006–1010. doi: 10.1126/science.1245994

Mugitani, R., Pons, F., Fais, L., Dietrich, C., Werker, J. F., Amano, S., et al. (2009). Perception of vowel length by Japanese- and English-learning infants. *Dev. Psychol.* 45, 236–247. doi: 10.1037/a0014043

Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., and Trent-Brown, S. A. (2008). Acoustic and perceptual similarity of Japanese and American English vowels. *J. Acoust. Soc. Am.* 124, 576–588. doi: 10.1121/1.2931949

Pajak, B., and Levy, R. (2014). The role of abstraction in non-native speech perception. *J. Phon.* 46, 147–160. doi: 10.1016/j.wocn.2014.07.001

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *J. Neurosci. Methods* 162, 8–13. doi: 10.1016/j.jneumeth.2006.11.017

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Ratko, L., Proctor, M., and Cox, F. (2022). Articulation of vowel length contrasts in Australian English. *J. Int. Phon. Assoc.* 1–30. doi: 10.1017/S0025100322000068

Shinohara, Y., Han, C., and Hestvik, A. (2022). Discriminability and prototypicality of nonnative vowels. *Stud. Second Lang. Acquis.* 44, 1260–1278. doi: 10.1017/S0272263121000978

Shinohara, Y., Han, C., and Hestvik, A. (2019). "Effects of perceptual assimilation: the perception of English /æ/, /ʌ/, and /ɑ/ by Japanese speakers," in *Proceedings of the 19th International Congress of Phonetic Sciences*, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Melbourne, VIC: Australasian Speech Science and Technology Association Inc), 2344–2348.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent-Brown, S. A., Nishi, K., Jenkins, J. J., et al. (1998). Perceptual assimilation of American English vowels by Japanese listeners. *J. Phon.* 26, 311–344. doi: 10.1006/jpho.1998.0078

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., and Munhall, K. G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *J. Acoust. Soc. Am.* 123, 397–413. doi: 10.1121/1.2804942

Tsukada, K. (2012). Comparison of native versus nonnative perception of vowel length contrasts in Arabic and Japanese. *Appl. Psycholinguist.* 33, 501–516. doi: 10.1017/S0142716411000452

Tsukada, K., Cox, F., Hajek, J., and Hirata, Y. (2018). Non-native Japanese learners' perception of consonant length in Japanese and Italian. *Second Lang. Res.* 34, 179–200. doi: 10.1177/0267658317719494

van Leussen, J.-W., and Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Front. Psychol.* 6, 1000. doi: 10.3389/fpsyg.2015.01000

Williams, D., and Escudero, P. (2014). Influences of listeners' native and other dialects on cross-language vowel perception. *Front. Psychol.* 5, 1065. doi: 10.3389/fpsyg.2014.01065

Williams, D., Escudero, P., and Gafos, A. (2018). Spectral change and duration as cues in Australian English listeners' front vowel categorization. *J. Acoust. Soc. Am.* 144, EL215–EL221. doi: 10.1121/1.5055019

Yazawa, K. (2020). *Testing Second Language Linguistic Perception: A Case Study of Japanese, American English, and Australian English Vowels* [PhD thesis]. Tokyo: Waseda University.

Yazawa, K., and Kondo, M. (2019). "Acoustic characteristics of Japanese short and long vowels: formant displacement effect revisited," in *Proceedings of the 19th International Congress of Phonetic Sciences*, eds S. Calhoun, P. Escudero, M. Tabain, and P. Warren (Canberra, ACT: Australasian Speech Science and Technology Association Inc), 671–675.

Yazawa, K., Whang, J., Kondo, M., and Escudero, P. (2020). Language-dependent cue weighting: an investigation of perception modes in L2 learning. *Second Lang. Res.* 36, 557–581. doi: 10.1177/0267658319832645

# Frontiers in Psychology

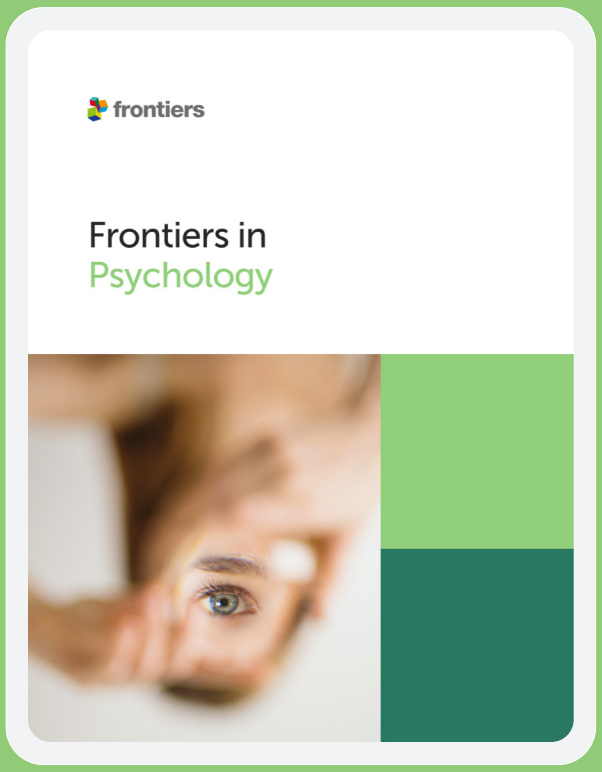
**Paving the way for a greater understanding of human behavior**

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