Bilingualism: Language and Cognition

cambridge.org/bil

Exploring how L2 utterance fluency relates to cognitive fluency in monologic and dialogic speaking

Jianmin Gao[1,2] and Peijian Paul Sun[3]

[1]College of Foreign Languages, Zhejiang University of Technology, Zhejiang, China; [2]Department of Applied Foreign Language Studies, Nanjing University, Jiangsu, China and [3]Department of Linguistics, Zhejiang University, Hangzhou, China

Research Article

Cite this article: Gao, J. and Sun, P.P. (2026). Exploring how L2 utterance fluency relates to cognitive fluency in monologic and dialogic speaking. *Bilingualism: Language and Cognition* **29**, 1050–1063. https://doi.org/10.1017/S1366728925100564

Received: 31 July 2024
Revised: 1 August 2025
Accepted: 26 August 2025
First published online: 7 October 2025

Keywords:
L2 speaking; L2 utterance fluency; L2 cognitive fluency; monologic speaking; dialogic speaking

Corresponding author:
Peijian Paul Sun;
Email: luapnus@zju.edu.cn

Abstract

We explored the relationships between L2 utterance fluency and cognitive fluency in monologic and dialogic tasks. The study involved 136 Chinese university-level English learners. Utterance fluency was measured through speed, breakdown, and repair fluency aspects. Cognitive fluency was indicated by L2 lexical and syntactic processing efficiency measures. Stepwise regression models, including metrics of L2-specific cognitive fluency, L2 knowledge, and L1 utterance fluency as predictors, targeted L2 utterance fluency as the dependent variable. We found that L2 cognitive fluency predicted limited variance in utterance fluency, with its influence more evident in monologues. L2 lexical processing efficiency paralleled syntactic processing efficiency's importance in the monologic task but surpassed it in dialogues. Moreover, L2 processing speed had a more significant impact on utterance fluency than processing stability across both contexts. We suggest that cognitive fluency is not the sole determinant of utterance fluency; L2 knowledge and L1 utterance fluency play non-negligible roles.

Highlights

- L2 cognitive fluency predicts limited variance in L2 utterance fluency
- L2 cognitive fluency's influence on utterance fluency is more evident in monologic speaking
- L2 lexical processing efficiency is more crucial than syntactic processing efficiency for utterance fluency
- L2 processing speed impacts utterance fluency more than processing stability

1. Introduction

Oral fluency is a core indicator of second language (L2) proficiency (Ginther et al., 2010) and is commonly viewed as a key learning goal in L2 teaching and learning (Lintunen et al., 2020; Yan, 2015). A comprehensive understanding of fluency is thus essential for promoting effective L2 speech production and guiding instructional practices. Specifically, insights into how cognitive mechanisms and temporal speech characteristics interact can inform teaching design and help learners achieve higher fluency (Suzuki & Kormos, 2023).

In the cognitive approach, L2 fluency is a multidimensional construct, encompassing cognitive fluency (CF), utterance fluency (UF), and perceived fluency (Segalowitz, 2010, 2016). While CF refers to the efficiency of cognitive processes underlying speech production, UF represents the temporal features of speech, such as speed and hesitation, that reflect these processes (Segalowitz, 2010, 2016; Suzuki & Kormos, 2023). Recent studies have modeled the UF-CF relationship in monologic tasks, shedding light on the cognitive underpinnings of UF (De Jong et al., 2013; Kahng, 2020; Suzuki & Kormos, 2023). However, these studies often neglect the potential effect of L1 UF and general cognitive abilities (e.g., domain-general information processing ability) on L2 UF and CF (De Jong et al., 2015; Gao & Sun, 2023; Gao & Sun, 2024; Kahng, 2020), and little is known about how these factors operate in dialogic contexts where interactional demands such as turn-taking can significantly influence speech production (McCarthy, 2010; Peltonen, 2017). Given the importance of dialogic communication in real-world contexts (Tavakoli, 2016), investigating the UF-CF link across both monologic and dialogic tasks is crucial for a more nuanced understanding of L2 fluency. Therefore, the overarching aim of the study is to examine the UF-CF relationship in both monologic and dialogic speaking, providing insights into the cognitive mechanisms underlying L2 speech production and offering practical implications for language teaching, learning, and assessment.

2. Literature review

2.1. L2 utterance fluency

In L2 fluency research, L2 UF refers to the temporal characteristics of speech that reflect CF (Segalowitz, 2010, 2016). This conceptualization of L2 UF aligns with Lennon (1990)'s narrow

© The Author(s), 2025. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

sense of L2 fluency, which distinguishes fluency from other aspects of speaking performance, such as lexical diversity and syntactic complexity. According to Lennon, while the broad sense of fluency encompasses overall speaking proficiency, fluency in its narrow sense is a performance phenomenon rather than linguistic knowledge that can be mentally stored. Building on this narrow sense of fluency, Skehan (2003) classified L2 UF measures into three dimensions: speed, breakdown, and repair fluency. Tavakoli and Skehan (2005) further substantiated this conceptual model through factor analysis, confirming the three-dimension nature of L2 UF. Speed fluency, often measured by articulation rate (number of syllables/speaking time excluding silent pauses), reflects delivery speed. Breakdown fluency is measured by the frequency and duration of filled and silent pauses. Filled pauses are identified as non-lexicalized pauses (e.g., *uh* or *um* in English), and silent pauses are identified as silence longer than 250 ms (De Jong & Bosker, 2013) or 200 ms (De Jong & Wempe, 2009; Gao et al., 2025). Repair fluency is exhibited by a range of repair behaviors, such as repetitions, corrections, and false starts.

2.2. L2 utterance fluency and speech production

Some researchers have resorted to the models of L2 speech production to explain L2 UF from a cognitive perspective (e.g., Gao & Sun, 2023, Kahng, 2014). According to Kormos' (2006) model, speech production goes through several major stages, including conceptualization, formulation, articulation, and monitoring. During conceptualization, speakers form an intention and organize relevant conceptual information for expression, resulting in a preverbal message. In the formulation stage, the preverbal plan undergoes lexico-grammatical, morpho-phonological, and phonetic encoding, culminating in internal speech, which is then executed at the articulation stage. Throughout the speech production process, monitoring loops continuously check the alignment of each output with the communicative purpose. If misalignment occurs, the process can be restarted or repaired. Kormos claimed that L2 UF is primarily related to the automatization of lexical, syntactic, morphological, and phonological encoding processes. Segalowitz (2010) further conceptualized the efficiency of the cognitive processes during L2 speech production as L2 CF. According to Segalowitz, L2 UF reflects L2 CF, and processing difficulties might occur at each stage of speech production, leading to dysfluencies.

2.3. L2 cognitive fluency

L2 CF can be operationalized as the efficiency of linguistic encoding processes, including lexical and syntactic processing speed and stability (Segalowitz, 2010, 2016). Segalowitz emphasized that CF involves more than faster processing; stability in processing is equally critical. To address this, he proposed the coefficient of variation (CV) as another measure of CF, reflecting reaction time (RT) variability normalized by processing speed. Calculated as the standard deviation (SD) of RT divided by the mean RT across trials (SD/RT), a lower CV indicates greater processing stability.

While some studies have focused on measures of lexical and/or syntactic processing efficiency (speed and stability) to indicate L2 CF (e.g., Olkkonen et al., 2024; Segalowitz & Freed, 2004), other studies have included L2 knowledge as an additional component of CF (De Jong et al., 2013; Kahng, 2020; Suzuki & Kormos, 2023). These two approaches, according to Suzuki and Kormos (2023), reflect narrow and broad conceptualizations of L2 CF, similar to Lennon's (1990) distinction between narrow and broad senses of L2 fluency. In the broad sense, L2 CF encompasses linguistic knowledge and processing speed that are essential for fluent speech production. The narrow sense aligns with Segalowitz's (2010, 2016) original conceptualization that L2 CF is a performance measure of rapidity and fluidity in mobilizing the complex cognitive processes (utterance planning, assembling, and executing) underlying L2 speech production.

Following Segalowitz's (2016) argument that L2 knowledge base does not directly indicate how fluent the L2 speech production process is (i.e., the core concern of L2 cognitive fluency), we chose to operationalize and interpret L2 CF in its narrow sense. This approach is firmly grounded in the theory of automaticity (Kormos, 2006; Segalowitz, 2010, 2016), wherein rapid and smooth speech reflects the automatization of encoding processes during L2 speech production. Although we did not include L2 knowledge as a component of L2 CF, we still analyzed its potential influence on L2 UF, as prior research (e.g., De Jong et al., 2013; Kahng, 2020; Suzuki & Kormos, 2023) suggests it may significantly affect UF. By doing so, it can help further clarify the relationship between L2 UF and CF, while ensuring comparability with previous studies.

It should be noted that L2 CF measures (processing speed and stability) are related to general-purpose cognitive control processes (Segalowitz, 2010, 2016). A recent study by Olkkonen et al. (2024) has found a strong relationship between L1-L2 cognitive traits, such as efficiency in lexical access (measured by accuracy in a rapid word recognition task, r = .59) and monitoring in attention control (measured by repair frequency in a Stroop task, r = .70). Olkkonen et al. further suggested that the L1 and L2 cognitive measures likely tap into shared general cognitive processes. This indicates that a general cognitive trait may govern both L1 and L2 cognitive functions. To obtain a purer measure of L2 processing speed and stability (referred to as the *L2-specific* measure by Segalowitz, 2010, 2016), a commonly used approach is to regress L2 measures on measures of general cognitive processing efficiency (Feng, 2022; Kahng, 2020; Segalowitz, 2010, 2016). However, previous studies, except for Kahng (2020), did not attempt to measure L2-specific cognitive processing efficiency.

2.4. The relationship between L2 utterance fluency and cognitive fluency

According to Segalowitz (2010, 2016), L2 CF plays a pivotal role in shaping UF. Investigating the links between different dimensions of CF (lexical and syntactic processing efficiency) and UF (speed, breakdown, and repair fluency) offers insights into the cognitive mechanisms driving fluency. While some studies have provided indirect evidence that different L2 UF measures may have varying cognitive underpinnings (e.g., Kahng, 2014; Tavakoli et al., 2020; Yan et al., 2021), much remains unclear. These studies found that the frequency of mid-clause silent pauses had stronger predictive power for L2 speaking proficiency than end-clause pauses. They hypothesized that silent pauses within clauses were related to the L2 encoding process during formulation, while end-clause pauses were linked to conceptualization. However, these assumptions have not been verified by examining the specific links between different UF and CF measures. The available empirical findings on the link remain extremely limited.

De Jong et al. (2013) made an early attempt to unravel the relationship between L2 UF and CF. Using a corpus consisting of 179 L2 Dutch learners' speech on eight role-play monologic tasks,

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

they explored the contribution of L2 knowledge and processing speed (CF in the broad sense) to UF through linear mixed modeling. They found that all UF measures were linked to one or more measures of L2 knowledge and processing speed. Specifically, the UF measure that was best explained was mean syllable duration (the inverse articulation rate), with 50% of the variance explained.

Following De Jong et al. (2013), Kahng (2020) also measured L2 CF in its broad sense. Kahng investigated the contribution of L2 knowledge, L2-specific processing speed, and L1 UF to L2 UF of 44 Chinese English learners across two monologic tasks. Through multiple regressions, Kahng found that the number of mid-clause silent pauses was predicted by L2 knowledge and processing speed, while end-clause silent pauses showed no such effects. This finding advanced the understanding of the cognitive basis of L2 UF and highlighted the distinct nature of mid-clause and end-clause silent pauses.

Suzuki and Kormos (2023) also conceptualized L2 CF in its broad sense, and modeled it as a two-factor construct, including linguistic resources (L2 knowledge) and processing speed. Through the structural equation modeling approach, they explored the connections between L2 UF and CF dimensions. The data were collected from 128 Japanese learners of English performing four monologic tasks. Their findings showed that speed fluency was primarily predicted by processing speed across all tasks, while breakdown fluency was influenced by both processing speed and linguistic resources. Repair fluency, however, was mainly predicted by linguistic resources in three out of four tasks, with processing speed having no significant effect. Although their study revealed contributions of L2 CF to different UF dimensions, more intricate relationships within these dimensions were not fully examined. This is particularly important since both the processing speed and linguistic resources dimensions encompass multiple sub-dimensions that may function differently in shaping L2 UF.

The varying statistical approaches and research focuses make it difficult to compare findings across these studies. While De Jong et al. (2013) and Kahng (2020) examined how L2 knowledge and processing speed jointly accounted for different L2 UF dimensions, Suzuki and Kormos (2023) focused on their separate contributions. Overall, previous studies agreed that most L2 UF measures could be predicted by L2 knowledge and processing speed, with repair fluency showing weaker predictability than speed and breakdown fluency.

2.5. Influence of L1 utterance fluency on L2 utterance fluency

As suggested by Segalowitz (2010, 2016), individual differences in L1 fluency should be considered when investigating the relationship between L2 UF and CF, an issue that has not been fully addressed in previous studies (e.g., Suzuki & Kormos, 2023). Research on L2 UF has increasingly recognized the impact of learners' L1 fluency on their L2 (e.g., De Jong et al., 2015; Derwing et al., 2009; Gao & Sun, 2023). These studies have shown a strong correlation between L1 and L2 UF measures, with L1 UF influencing all L2 UF dimensions. A meta-analysis of 16 empirical studies by Gao and Sun (2024) found that L1 UF significantly correlated with L2 UF, particularly in breakdown fluency (*r*s ranged from .58 to .62), followed by speed (r = .46) and repair fluency (*r*s ranged from .26 to .34). Given that L1 UF is a powerful predictor of L2 UF, it is crucial to account for its effect when examining the relationship between L2 UF and CF (De Jong et al., 2013). Including L1 UF as a predictor also provides clearer insights into how L1 UF and L2 CF shape L2 speech fluency.

2.6. L2 fluency in monologic versus dialogic tasks

Despite insights from previous studies, there is a clear knowledge gap regarding the relationship between L2 UF and CF in the dialogic context. Given the widely documented differences in the nature of monologues and dialogues (e.g., Michel, 2011; Tavakoli, 2016, 2018; Witton-Davies, 2014), it is anticipated that the cognitive underpinnings of UF vary between the two contexts. A key difference lies in the cognitive demands. While monologic tasks require sustained and uninterrupted speech production by one individual, dialogic tasks are characterized by turn-taking and interaction, which fundamentally alters the cognitive processing involved. Specifically, without interactional support in monologic tasks, speakers have to depend predominantly on their internal cognitive resources (e.g., lexical access and syntactic processing) to manage all stages (conceptualization, formulation, articulation, and monitoring) of speech production. In contrast, the dialogic turn-taking allows speakers to plan and organize their own speech during their interlocutors' turns, reducing cognitive strain (Michel, 2011; Tavakoli, 2016). Additionally, dialogic speaking inherently requires shorter utterances within turns to ensure the effective exchange of information, thus reducing the chance to produce complex utterances that typically rely more on L2 processing (Cameron, 2001; Tavakoli & Wright, 2020).

An additional factor that may differentially impact cognitive processes in monologic and dialogic tasks is the influence of interlocutors' cognitive abilities. According to Feng (2022), a speaker's UF in L2 dialogues is not solely determined by their own cognitive resources but is also shaped by those of their interlocutor. As interlocutors engage in a dialogue, they tend to synchronize and align with each other (Pickering & Garrod, 2021). This synchronization allows one speaker's cognitive abilities to affect the other's UF performance indirectly through their utterances (Tavakoli & Wright, 2020; Pickering & Garrod, 2021). Considering the differences in speech production between monologic and dialogic tasks, it is imperative to extend the investigation of the UF-CF relationship to dialogic contexts, which will contribute to a more nuanced and comprehensive understanding of this link.

Motivated by the methodological and knowledge gaps and built upon Segalowitz's (2010, 2016) cognitive framework of L2 fluency, the current study explores the relationships between L2 UF and CF in both monologic and dialogic contexts. Drawing on insights from previous research, we also considered the contributions of L2 knowledge and L1 UF to L2 UF and employed refined measures of L2 CF that minimize the confounding effects of general cognitive ability when investigating the relationship. Specifically, the study was guided by the following two research questions (RQs):

RQ1: What are the relationships between L2 UF and CF in monologic speaking, considering the influence of L2 knowledge and L1 UF?

RQ2: What are the relationships between L2 UF and CF in dialogic speaking, considering the influence of L2 knowledge and L1 UF?

3. Method

3.1. Participants

Through convenience sampling, a total of 136 university English learners from a public university in Southeastern China participated in this study. The sample comprised non-English majors, 2nd-year English majors, and 4th-year English majors. Participants' English levels were estimated to range from the lower

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

intermediate to advanced. Before the experiment, we collected all participants' consent forms and background questionnaires. Only participants who provided a complete set of data were included in each analysis. Participants with missing values on any relevant measures were excluded from the corresponding analyses. For RQ1, data from 108 participants (92 females; M_{age} = 19.87, SD = 1.34) were analyzed. For RQ2, data from 103 participants (93 females; M_{age} = 20.08, SD = 1.37) were analyzed. Of these, 81 participants contributed to both RQ1 and RQ2.

3.2. EIT for measuring L2 proficiency

The L2 elicited imitation task (EIT) is an effective measure of L2 proficiency (Wu et al., 2022). The task requires participants to listen to sentences in the target language and accurately repeat them. A meta-analysis by Yan et al. (2016) indicates that the EIT score can effectively distinguish L2 learners' proficiency levels (Hedges' g = 1.34). We adopted the English EIT developed by Ortega et al. (2002) and validated by Wu et al. (2022). To elicit balanced conversations in the dialogic task, participants with similar EIT scores were paired, as mismatched pairs (i.e., low-high proficiency) can result in asymmetrical interaction, with one speaker dominating the conversation (Davis, 2009).

The task included 30 English sentences ranging from seven to 19 syllables. Participants listened to stimuli ordered by syllable count, starting with the shortest sentence. The current study adopted the same time parameters as those used in Ortega et al. (2002). Specifically, a ring tone prompted participants to start their repetition 2.5 s after each sentence ended. Participants were allowed an extra 2 s to repeat the sentence beyond the time taken by the native speaker in Ortega et al.'s study to articulate the sentence. An extra 0.5 s was added for each additional syllable beyond the seventh. A practice session of four Chinese sentences at the beginning was used to ensure participants understood the procedure.

3.3. L1 and L2 monologic and dialogic speaking tasks

Four speaking tasks were developed for eliciting L1 and L2 monologic and dialogic speech. We adapted the topic-given speaking tasks in College English Test-Spoken English Test (CET-SET) Band 4 to elicit participants' L2 speech. CET-SET is a national English test for university students in China. In the monologic task, participants shared their personal opinions and explained their reasoning about the importance of a particular behavior, such as protecting the environment. In the dialogic task, participants were invited to work in pairs to discuss how to organize a public lecture on the same topic they addressed in the monologic task (e.g., a lecture on environmental protection). They engaged in a discussion regarding the choice of lecture venue, speakers to be invited, and shared their own suggestions on the topic.

In collaboration with an experienced university English teacher, two task sets were designed to elicit L1 and L2 speech. Each set included a monologic and a dialogic task. The task set about environmental protection was used to elicit L1 speech, and the task about keeping healthy was used to elicit L2 speech (see Supplemental File 1 for the prompts). Following CET-SET guidelines, participants had 45 s for preparation and 1 min of speaking time for the monologic task. For the dialogic task, they had 1 min for preparation and 3 min for speaking. To mitigate potential time pressure, participants were informed they could exceed the designated speaking time. In the instructions for dialogic tasks, participants were reminded to provide their partners with equal opportunities for a balanced conversation. The L1 and L2 tasks were the same in task implementation conditions (see more details in the *Procedure* section). The decision to use the same task format to elicit L1 and L2 speech was made to focus on the influence of L1 UF as a predictor for L2 UF, rather than introducing variability caused by task differences. By using different topics for the L1 and L2 tasks, we aimed to minimize practice effects.

3.4. L2 cognitive fluency tasks

3.4.1. L2 lexical processing task

We adopted an L2 picture-naming task to capture lexical processing speed and stability. Following previous research (De Jong et al., 2013; Kahng, 2020; Suzuki & Kormos, 2023), we selected 47 pictures from Snodgrass and Vanderwart's (1980) pool of 260 black-and-white line drawings. Five pictures were used in practice trials, and 42 in the formal experiment. The selection criteria, based on the normative data from Liu et al. (2011) and Johnston et al. (2010), included: (a) the picture should be easily recognizable by Chinese English learners (concept agreement ≥95%), (b) the concept depicted by the picture should be familiar to Chinese English learners (concept familiarity rating ≥4 on a 5-point Likert scale), (c) the picture itself should not cause ambiguity in English naming (name agreement ≥95%). Picture attributes are detailed in Supplemental File 2.

The task was designed using PsychoPy v.2021.2.3 (Peirce et al., 2019), a software for creating psychology experiments. Consistent with De Jong et al. (2013) and Kahng (2020), each trial commenced with a 1500 ms fixation cross, followed by a 2000 ms pictorial stimulus and a 500 ms blank screen. Picture order was randomized for each participant, who was instructed to promptly and accurately name each picture. The trial process was illustrated in Figure 1 (see *Procedure* section for details in data collection).

3.4.2. L2 syntactic processing task

Following Suzuki and Kormos (2023), a maze task was employed to elicit L2 syntactic processing, which involves sequential forced choice between a legitimate continuation of a sentence and a syntactically inappropriate distractor. We employed the maze task developed by Suzuki and Sunada (2018) and modified by Gao and Sun (2024) specifically for university-level Chinese English learners. The task required participants to complete 32 English sentences of four types of syntactic structures, including declarative sentences, wh-questions, relative clauses, and indirect questions, with eight sentences for each type (see Supplemental File 3 for the sentences).

The maze task was designed using PsychoPy v.2021.2.3 (Peirce et al., 2019). As depicted in Figure 2, the first screen served as a prompt to initiate a trial, displaying the initial word of the sentence on the left and cross signs on the right to guide word selection on subsequent screens. This selection process continued until the sentence was fully constructed. Following Suzuki and Kormos (2023), each prompt screen lasted 2000 ms, and participants had 4300 ms to make a choice. If an incorrect option was selected, the trial would be immediately terminated, with the remaining part skipped. We randomized the sentence order for participants and instructed them to make the keyboard response as fast and accurately as possible, pressing either "F" (left word) or "J" (right word). A practice session of four English sentences preceded the formal experiment, which required 297 keyboard responses, with the keys evenly split between "F" (51%) and "J" (49%).

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

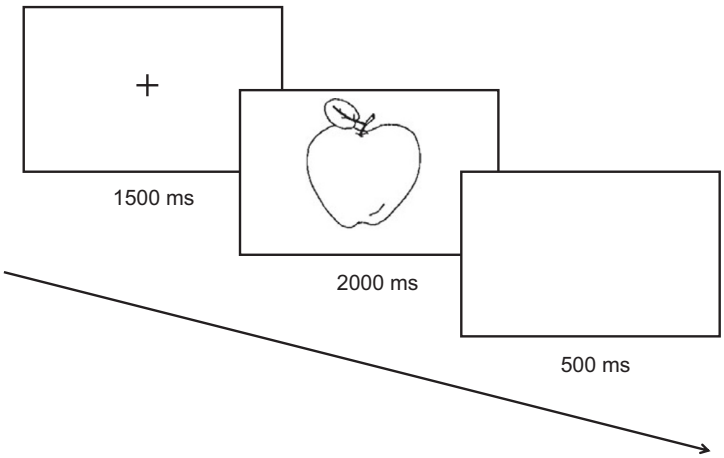


Figure 1. The procedure of the picture-naming task.

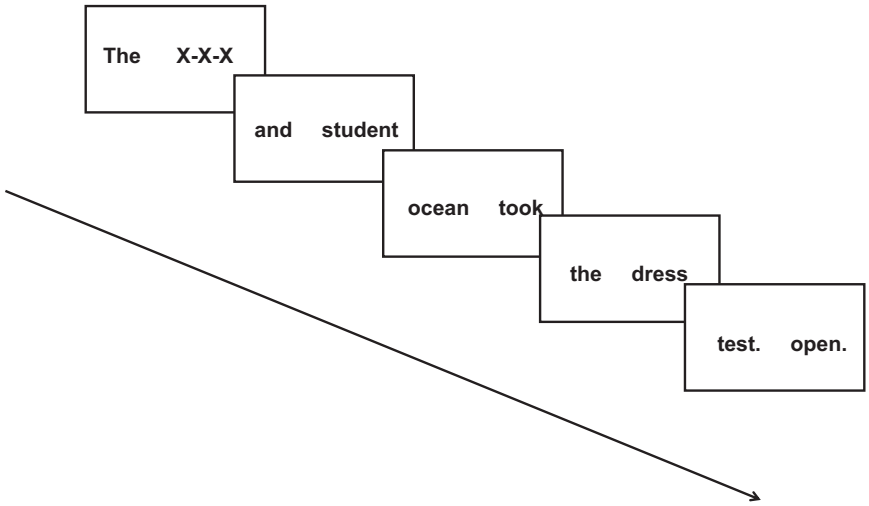


Figure 2. The procedure of the maze task.

3.5. Domain-general information processing task

The computerized Digit Symbol Substitution Test (c-DSST) was employed to measure the speed and stability of domain-general information processing. Given that participants' individual differences in the ability to process non-verbal information may compound the measurement of their L2 processing efficiency (Segalowitz, 2010), it is necessary to include the task to control such effects. Additionally, the syntactic processing task required keyboard responses, which might lead to measurement errors arising from individual differences in the motor speed of keyboard pressing. Using the c-DSST as a baseline for data correction on the maze task may help mitigate measurement errors.

In light of Chen et al. (2020), we designed five blocks of c-DSST, each consisting of 18 trials, resulting in a total of 90 trials. Each trial involved the visual presentation of nine digit-symbol pairs as references and a single digit-symbol probe (see Figure 3). Participants were instructed to promptly determine whether the probe matches any of the nine digit-symbol pairs. If a match was identified, participants were required to press "F"; otherwise, they pressed "J." The stimulus lasted 4000 ms until the participants pressed one of the designated keys. Between each trial, there was a 1500 ms blank screen. To minimize practice effects, the nine reference pairs and the digit-symbol probes varied across the five blocks. Each block had an equal distribution of nine matching and non-matching probes. The order of blocks and trials was randomized. Participants completed five practice trials to familiarize themselves with the procedure.

3.6. Tests of L2 knowledge

Following previous research (Kahng, 2020; Suzuki & Kormos, 2023), this study focused on L2 knowledge of vocabulary and

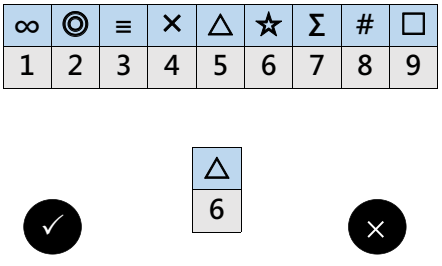


Figure 3. The procedure of the c-DSST.

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

grammar. Specifically, L2 vocabulary size test, L2 phrasal vocabulary size test, and L2 test of vocabulary and grammar were employed for measuring participants' L2 knowledge.

3.6.1. L2 vocabulary size test

Participants' vocabulary size was estimated by the Vocabulary Size Test (VST) developed by Nation & Beglar (2007). The VST "was developed to provide a reliable, accurate, and comprehensive measure of a learner's vocabulary size from the first 1000 to the 14th 1000 word families of English" (Nation & Beglar, 2007, p. 9). The original test comprises 140 items, evenly distributed across 14 frequency levels of words. It employs a meaning-recognition format, requiring test takers to select the correct option from one target answer and three distractors. Each word being tested is embedded in a non-defining context.

We used a bilingual version of the VST, as recommended by Nguyen and Nation (2011). The original English-Chinese version of the VST was modified by Zhao and Ji (2016) to better suit Chinese English learners. Modifications included adjusting the wording of certain Chinese options for clarity and grammatical accuracy. As suggested by Nation and Beglar (2007) and Beglar (2010), to reduce test duration and fatigue, a shortened version of the VST by Zhao and Ji (2016) was employed. Considering participants' education and proficiency levels, items from the third to eighth 1000-word frequency levels were selected. Five items were randomly chosen from the 10 items at each of these levels, resulting in a 30-item test (see Supplemental File 4 for the VST used in the study). Consistent with Nation and Beglar (2007), one point was awarded for each correct answer, with no points for incorrect answers.

3.6.2. L2 phrasal vocabulary size test

The Phrasal Vocabulary Size Test (PVST) developed by Martinez (2011) was used to measure participants' knowledge of phrasal expressions. Phrasal expressions, as defined by Martinez, refer to fixed or semi-fixed sequences of two or more words that co-occur but may not necessarily appear together consecutively. These expressions possess cohesive meanings or functions that cannot be readily deduced by decoding the individual words.

Martinez (2011) developed the PHRASE List (also referred to as Martinez & Schmitt, 2012), consisting of the 505 most frequent multiword expressions based on the British National Corpus. It includes 32, 84, 129, 157, and 103 phrases corresponding to the first 1000 frequency level to the fifth 1000, respectively. The PVST, based on the PHRASE List, measures the breadth of knowledge of phrasal expressions. It consists of 50 items, with 10 items at each of the five frequency levels. The format of PVST is consistent with the VST (Nation & Beglar, 2007), requiring participants to choose from three distractors and one key option that explains the meaning of the target phrase. The options were in English, using words or phrases at the same frequency level as the target phrase. Given the absence of a validated bilingual version of the test, the original PVST was used. The scoring adhered to Martinez's (2011) frequency-sensitive method, with each item's score weighted according to its frequency level. For example, the first 10 items, representing the first 1000 frequency level of phrases, were assigned a weight of 3.2. The test has a total score of 505, mirroring the complete set of phrasal expressions in the PHRASE List.

3.6.3. L2 test of vocabulary and grammar

The vocabulary and grammar test used is part of the DIALANG online test batteries, designed to diagnose L2 learners' skills in listening, writing, reading, grammar, and vocabulary use. The test has been validated as an effective tool for measuring L2 proficiency (Alderson, 2005; Alderson & Huhta, 2005). The vocabulary section assesses the ability to accurately use vocabulary in context. It measures both productive and receptive vocabulary knowledge through tasks such as multiple-choice, filling-in-the-blank, and short-answer questions. Unlike the VST, it taps into a far more comprehensive knowledge of vocabulary, including denotative meaning, semantic relations, combinations, and word formation, than merely the form-meaning link. The grammar section measures the ability to understand and use morphology and syntax. The items encompass a variety of task types, allowing individuals to demonstrate their ability to comprehend and produce relevant grammatical structures.

Alderson (2005) noted the challenge of drawing a clear distinction between grammar and vocabulary, as numerous linguistic features intersect these areas. Considering that both the sections focus on language use in contexts, scores from these sections were aggregated in this study to form a unitary variable capturing comprehensive knowledge of vocabulary and grammar. There are three versions of the test with different difficulty levels (easy, medium, and difficult). The choice to administer the medium-difficulty version of the test was made given participants' proficiency levels, which range from the lower intermediate to advanced. Using the medium-difficulty version ensures that the test is neither too easy nor overly challenging for participants. There were 60 items, including 30 items from the vocabulary section and 30 from the grammar section. Each item was scored as one point for correct and zero for incorrect or missing answers.

3.7. Procedure

The data collection process was implemented consistently across different participant groups (non-English majors, 2nd-year English majors, and f4th-year English majors). For each group, data were collected within 2 weeks, comprising four data collection sessions (see Figure 4).

In the first session, participants were situated in a multimedia classroom. The EIT was administered first, lasting approximately 8 min. After a 10-min break, participants completed the VST, PVST, and a test of comprehensive knowledge of vocabulary and grammar. Although the time to complete the L2 knowledge tests varied from participants, all adhered to the maximum allotted time of 40 min. EIT performances were rated immediately after this session. In reference to the specific rating guideline developed by Ortega et al. (2002), the first author rated the EIT performance on a scale of 0 to 4 based on the completeness of the repetition. To examine rating reliability, a research assistant independently rated 30 randomly selected speech samples (22.1% of the total 136), using the rating guideline. A high inter-rater agreement (the intraclass correlation coefficient, ICC) of .98, 95%CI [.97, .99] was achieved.

The second session lasted 1 week. Participants individually visited the first author or a research assistant in a quiet classroom to complete the picture-naming task, followed by the maze task and the c-DSST. All tasks were delivered using the computer. RT data were automatically calculated by the PsychoPy program. In the picture-naming task, RT was the duration between the onset of the picture and the oral response. In the maze task and c-DSST, it was the duration between the onset of word options or digit-symbol probes and the keyboard response. Each task was separated by an optional 5-min break, and the whole session lasted approximately

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

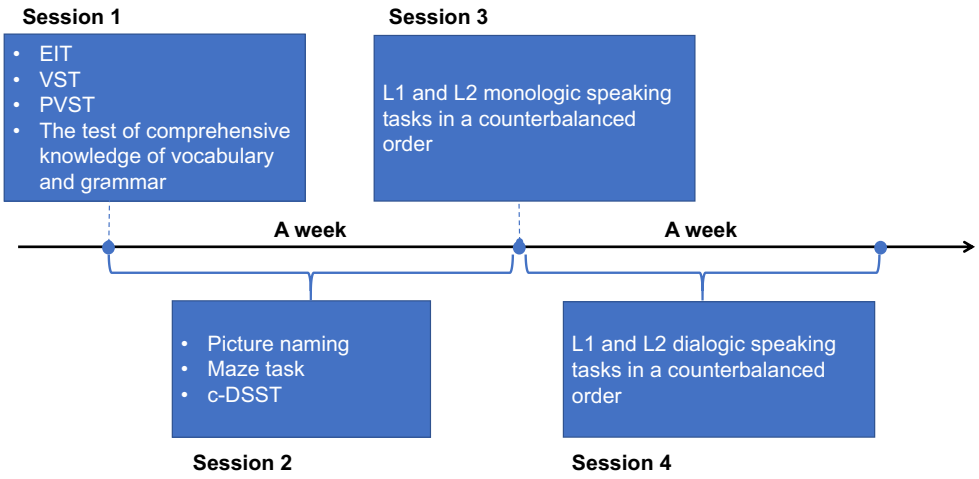


Figure 4. Timeline of data collection.

25 min per participant. RT data were excluded for incorrect responses, and data from participants with accuracy rate below 60% were excluded.

In the third session, participants completed L1 and L2 monologic tasks in a multimedia classroom. The order of L1 and L2 tasks were counterbalanced across participants. The session lasted 15 min per participant.

In the final session, participants completed L1 and L2 dialogic tasks, paired based on their EIT scores. Most pairs (86.8%, $N = 46$) had a score difference of 0 to 4, indicating a close match in L2 proficiency. A smaller proportion (11%, $N = 6$) had a score difference of 5 to 8, and only one pair had a larger score difference of 18, with both participants having low EIT scores (39 and 21). These tasks were conducted in a quiet classroom, with each pair attending individually and the task order counterbalanced.

3.8. Data analysis

3.8.1. Acoustic analysis

The present study adopted a set of well-established measures to capture L2 UF across three dimensions: speed, breakdown, and repair fluency (see Table 1 for the specific measures). To clarify the specific links between dimensions of L2 UF and CF, the composite fluency indicator that is related to both speed and breakdown fluency dimensions, such as speech rate (number of syllables/speaking time including silent pauses), was not used to measure L2 UF. In measuring breakdown fluency, silent pauses were analyzed based on their location. The Analysis of Speech Unit (ASU), proposed by Foster et al. (2000), was employed as the syntactic unit for speech analysis instead of the clause-based approach, as evidenced by its suitability in analyzing spoken language (Tavakoli & Wright, 2020). An ASU is defined as "an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000, p. 365). Therefore, some clause-boundary pauses were coded as mid-ASU pauses using this approach. The frequency and duration of silent pauses at the middle and end of an ASU were calculated separately.

For L2 dialogic speech samples, following Tavakoli (2016), conversations were examined to ensure even distribution of turns and balanced participation between interlocutors. Conversations with significant imbalances, such as one speaker dominating over 70% of the time or the other remaining silent for such periods, were excluded. Specifically, seven conversations (14 out of the 136 participants) were excluded. The data were also scrutinized to ensure that each participant had at least two turns within the 3-min dialogue. Turn pauses (the silence between speaker changes, Trouvain & Werner, 2022) were differentiated from the silent pauses within individual turns and then excluded from data analysis as the primary focus was on individual UF performance in both monogic and dialogic contexts.

For L1 UF measurement, we followed previous research on the influence of L1 on L2 UF (e.g., De Jong et al., 2015; Derwing et al., 2009; Gao & Sun, 2023) and employed the same set of measures used for L2 UF to ensure comparability across languages. However, given that no threshold of silent pauses has been established or validated for L1 Chinese speech, we adopted a conservative approach in this study. To elaborate, we used the measures unaffected by the silent pause threshold to measure L1 Chinese UF. Since speed fluency measures (e.g., articulation rate) and silent pause measures require identification of silent pauses based on a threshold, they were not used in the L1 sample analysis. Instead, L1 speech rate, which was related to both speed and breakdown fluency and unaffected by the silent pause threshold, was employed as a control variable when examining the link between L2 speed/breakdown fluency measures and L2 CF. Other measures used to measure L1 UF included filled pause rate, repetition rate, correction rate, and false start rate. Table 1 presents the detailed descriptions of the measures of L1 and L2 UF.

Speech samples were automatically transcribed on the Feishu platform (https://www.feishu.cn/en/) and manually cross-checked for verbatim accuracy. Repetitions, corrections, and false starts were manually coded. The first author coded the transcriptions. To examine coding reliability, a total of 80 transcriptions encompassing 20 transcriptions each for L1 monologic speech, L1 dialogic speech, L2 monologic speech, and L2 dialogic speech were randomly selected and re-coded by a research assistant. Inter-coder agreement was used to establish coding reliability (see Table 2).

Praat v. 6.1.36 (Boersma & Weenink, 2020) was used to segment sounded and silent intervals. Using a script developed by De Jong and Wempe (2009), silent pauses of at least 200 ms were detected within the L2 monologic and dialogic samples. To ensure accuracy, the automatically annotated boundaries of silences were manually

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

Table 1. Utterance fluency measures adopted in the current study

L1/L2	Measures	Description	Reference
Speed fluency			
L2	articulation rate (AR)	Number of syllables/speaking time excluding silent pauses	Suzuki & Kormos (2023)
Breakdown fluency			
L2	mid-ASU silent pause rate (MASPR)	Number of mid-ASU silent pauses/100 syllables	Gao & Sun (2023)
L2	end-ASU silent pause rate (EASPR)	Number of end-ASU silent pauses/100 syllables	Gao & Sun (2023)
L2	mid-ASU silent pause duration (MASPD)	Duration of mid-ASU silent pauses/number of mid-ASU silent pauses	Huensch & Tracy-Ventura (2017)
L2	end-ASU silent pause duration (EASPD)	Duration of end-ASU silent pauses/number of end-ASU silent pauses	Huensch & Tracy-Ventura (2017)
L1 and L2	filled pause rate (FPR)	Number of filled pauses (e.g., *uh* or *um*)/100 syllables	Suzuki & Kormos (2023)
Repair fluency			
L1 and L2	repetition rate (RR)	Number of partial and complete repetitions/100 syllables	Suzuki & Kormos (2023)
L1 and L2	correction rate (CR)	Number of corrections/100 syllables	Williams & Korko (2019)
L1 and L2	false start rate (FSR)	Number of false starts/100 syllables	Williams & Korko (2019)
Composite measure			
L1	speech rate (SR)	Number of syllables/speaking time including silent pauses	Suzuki & Kormos (2023)

Table 2. Percentages of agreement in coding repetitions, corrections, and false starts

	L1 monologic speech (%)	L1 dialogic speech (%)	L2 monologic speech (%)	L2 dialogic speech (%)
Repetition	91.1	81.6	88.0	89.2
Correction	94.1	83.3	87.9	86.7
False start	85.4	82.4	85.9	85.0

verified and adjusted by cross-referencing with the corresponding waveform and spectrogram. Filled pauses and the position of each silent pause were manually annotated in Praat's TextGrid files.

3.8.2. Statistical analysis

L2 CF measures (Mean RT and CV for L2 lexical and syntactic processing tasks) were first regressed on the corresponding measures of domain-general information processing efficiency. Following prior research (De Jong et al., 2015; Gao & Sun, 2023; Kahng, 2020), the resulting residuals were retained as the L2-specific CF measures.

Stepwise multiple regressions were conducted, where L2 UF measures were outcome variables, and measures of L2 CF in its narrow sense (L2 lexical and syntactic processing efficiency measures) were used as predictors. To control for L2 CF in its broad sense and gain a comprehensive understanding of the multiple cognitive factors underlying L2 UF performance, the model also included L2 knowledge measures and the L1 UF measures as predictors. The assumptions for linear regressions, including the normality of residuals, extreme outliers, homoscedasticity, and multicollinearity, were checked by detecting the distribution pattern of residuals, Q-Q plot of residuals, Crook distance plot, and variance inflation factor (VIF) values. All models met these assumptions.

4. Results

4.1. L2-specific measures of cognitive fluency

Pearson correlation coefficients were first computed between measures of domain-general information processing efficiency and L2 CF in its narrow sense (lexical and syntactic processing efficiency). The results exhibited that domain-general information processing speed (M = 1.5, SD = .21) was significantly correlated with both L2 lexical (M = 1.1, SD = .47) and syntactic (M = 1.2, SD = .22) processing speed. Based on Plonsky and Oswald's (2014) benchmarks for effect sizes (r = |.25|, small; r = |.40|, medium; r = |.60|, large), the correlation strength was weak with L2 lexical processing speed (r = .28, p = .001), whereas the correlation with L2 syntactic processing speed was stronger (r = .31, p < .001). No significant correlations were found between domain-general information processing stability (M = .33, SD = .06) and L2 lexical (M = .56, SD = .18) (r = −.02, p = .784) or syntactic processing stability (M = .42, SD = .07) (r = .05, p = .606). The measures of L2-specific lexical and syntactic processing speed were computed by regressing the data obtained from L2 picture-naming and maze tasks on domain-general information processing speed, saving the residuals for the following analyses.

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

Table 3. Results of the stepwise regressions predicting L2 utterance fluency using L2 cognitive fluency, L2 knowledge, and L1 utterance fluency in monologic speaking (N = 108)

Dependent variables	Predictors	adjusted R^2	Δ adjusted R^2	β	t	VIF	F
L2AR	L2 lexical processing stability	.06	.06	−.27**	−2.65	1.00	7.05
L2MASPR	L2 lexical processing speed	.14	.14	.19*	1.98	1.25	16.46
	+ L1SR	.19	.05	−.30**	−3.28	1.10	12.30
	+ L2 syntactic processing stability	.25	.06	.26**	2.94	1.04	11.30
	+ L2 comprehensive knowledge of vocabulary and grammar	.30	.05	−.26**	−2.79	1.14	11.05
L2MASPD	L1SR	.16	.16	−.33**	−3.69	1.08	18.39
	+ L2 comprehensive knowledge of vocabulary and grammar	.26	.10	−.27**	−2.88	1.14	17.47
	+ L2 lexical processing speed	.28	.02	.20*	2.03	1.22	13.41
L2EASPD	L2 lexical processing speed	.06	.06	.26*	2.60	1.00	6.74
L2FPR	L1FPR	.42	.42	.66**	8.37	1.00	70.11
L2RR	L2 comprehensive knowledge of vocabulary and grammar	.18	.18	−.40**	−4.58	1.08	21.77
	+ L2 syntactic processing speed	.27	.09	.22*	2.41	1.21	18.45
	+ L2 syntactic processing stability	.31	.04	.19*	2.22	1.03	14.93
	+ L1RR	.33	.02	.19*	2.13	1.17	12.77
L2CR	L2 syntactic processing speed	.05	.05	.24*	2.36	1.00	5.58

Note: AR = articulation rate; MASPR = mid-ASU silent pause rate; MASPD = mid-ASU silent pause duration; EASPD = end-ASU silent pause duration; FPR = filled pause rate; RR = repetition rate; CR = correction rate; SR = speech rate.
*$p < .05$.
**$p < .01$.

4.2. Utterance fluency-cognitive fluency link in the monologic task

Table 3 presents the results of stepwise multiple regressions in the monologic task (see Supplemental File 5 for descriptive statistics). For mid-ASU silent pause rate (MASPR), mid-ASU silent pause duration (MASPD), and repetition rate (RR), a combination of L2 CF measures, comprehensive knowledge of vocabulary and grammar, and the corresponding L1 UF measures emerged as significant predictors. For the specific relationships between L2 UF and CF, MASPR was predicted by L2 lexical processing speed and L2 syntactic processing stability, accounting for 20% of the variance. MASPD was predicted only by L2 lexical processing speed, explaining 2% of the variance, while RR was predicted by L2 syntactic processing speed and stability, accounting for 13% of the variance.

Additionally, articulation rate (AR), end-ASU silent pause duration (EASPD), and correction rate (CR) showed predictability merely based on a single L2 CF measure among the set of predictors. L2 lexical processing stability was the only predictor for AR, explaining 6% of the variance, and L2 lexical processing speed was the predictor for EASPD, also explaining 6% of the variance. Similarly, CR could only be predicted by syntactic processing speed, accounting for 5% of the variance. It was noteworthy that L2 filled pause rate (FPR) could only be predicted by L1 FPR, with a substantial amount of variance (42%) explained. Please note that the outcomes of the model predicting end-ASU silent pause rate (EASPR) and false start rate (FSR) are not presented, as neither UF measure could be significantly predicted by any predictors.

4.3. Utterance fluency-cognitive fluency link in the dialogic task

Table 4 presents the results of stepwise multiple regressions in the dialogic task. Similar to the findings in monologic speaking, FSR could not be significantly predicted by any of the measures. In the dialogic task, less variance in L2 UF measures was predicted by L2 CF, L2 knowledge, and L1 UF. Only four L2 UF measures (MASPR, MASPD, EASPD, and CR) were significantly predicted by L2 CF measures. MASPR and EASPD could be predicted by L2 lexical processing speed, accounting for 7% and 8% of the variance, respectively. MASPD could be predicted by L2 syntactic processing speed, explaining 4% of the variance. CR was predicted by L2 lexical processing stability, which accounted for 7% of the variance.

Among the other L2 UF measures, AR and EASPR were only predicted by L2 comprehensive knowledge of vocabulary and grammar, explaining 9% and 7% of the variance, respectively. L2 FPR not only showed predictability based on L2 comprehensive knowledge of vocabulary and grammar (6%) but also displayed predictability from L1 FPR, which explained a larger amount of variance (19%). Additionally, L2 RR in the dialogic task was solely predicted by the equivalent L1 measure, accounting for 7% of the variance.

5. Discussion

5.1. Overall relationships in monologic speaking

In the monologic task, except for EASPR and FSR, which could not be predicted by any measure, and FPR, which could only be predicted by its L1 counterpart, all six other L2 UF measures were predictable by L2 CF. AR, EASPD, and CR could only be predicted by L2 CF, with 6%, 6%, and 5% of the variance predicted. MASPR, RR, and MASPD could be predicted by a combination of measures of L2 CF, L2 knowledge, and L1 UF. MASPR and RR were more affected by L2 CF factors than by L2 knowledge and L1 UF, with 20% and 13% variance explained by L2 CF, respectively. On the other hand, the CF measure only predicted an additional 2% of the variance in MASPD, with most of its variance predicted by L2 knowledge and L1 UF.

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

Table 4. Results of the stepwise regressions predicting L2 utterance fluency using L2 cognitive fluency, L2 knowledge, and L1 utterance fluency in dialogic speaking (N = 103)

Dependent variables	Predictors	adjusted R^2	Δ adjusted R^2	β	t	VIF	F
L2AR	L2 comprehensive knowledge of vocabulary and grammar	.09	.09	.31**	3.31	1.00	10.92
L2MASPR	L2 comprehensive knowledge of vocabulary and grammar	.13	.13	−.30**	3.35	1.06	16.39
	+ L2 lexical processing speed	.21	.07	.30**	3.29	1.06	14.40
L2EASPR	L2 comprehensive knowledge of vocabulary and grammar	.07	.07	−.28**	−2.98	1.00	8.87
L2MASPD	L2 comprehensive knowledge of vocabulary and grammar	.09	.09	−.25*	−2.64	1.08	11.05
	+ L2 syntactic processing speed	.13	.04	.23*	2.38	1.08	8.60
L2EASPD	L2 lexical processing speed	.08	.08	.29 **	3.07	1.00	9.40
L2FPR	L1FPR	.19	.19	.45**	5.29	1.00	24.23
	+ L2 comprehensive knowledge of vocabulary and grammar	.25	.06	−.27**	−3.19	1.00	18.29
L2RR	L1RR	.07	.07	.29**	3.00	1.00	9.00
L2CR	L2 lexical processing stability	.07	.07	.27**	2.85	1.00	8.11

Note: AR = articulation rate; MASPR = mid-ASU silent pause rate; EASPR = end-ASU silent pause rate; MASPD = mid-ASU silent pause duration; EASPD = end-ASU silent pause duration; FPR = filled pause rate; RR = repetition rate; CR = correction rate.
*$p < .05$.
**$p < .01$.

The importance of L2 CF in predicting monologic UF features echoes with previous studies. For example, Kahng (2020) found L2 lexical and syntactic processing speed (L2 CF in the current study) contributed to most L2 UF measures expect for the number of filled pauses and repetitions, with the explained variance ranging from 7% to 30%. De Jong et al. (2013) also found that lexical and syntactic processing speed were significantly, though weakly, correlated with most L2 UF measures. Additionally, Suzuki and Kormos (2023) revealed that processing speed exhibited stable predictive effects on speed and breakdown fluency (but not on repair fluency) in four different monologic tasks. Taken together, L2 CF has a prevailing predictive effect on L2 UF in the monologic task. However, the importance of L2 CF is still relatively marginal, as suggested by the findings that the largest explained variance by CF measures was below 20% in this study and below 30% in Kahng's study. In other words, at least in the sense of L2 processing speed and stability, L2 CF alone is not sufficient to determine L2 UF in monologic speaking. The contributions of L2 knowledge and L1 UF to L2 UF are further discussed in the following sections.

5.2. Overall relationships in dialogic speaking

L2 CF in the dialogic task exhibited weaker predictive power for L2 UF than in the monologic task. Specifically, L2 CF only showed limited predictability on four out of the nine UF measures, including MASPR, MASPD, EASPD, and CR. For MASPR and MASPD, the predictive effect of L2 CF was less than that of L2 knowledge. Based on the prediction of L2 knowledge, an additional 7% and 4% of the variance in MASPR and MASPD could be predicted by the CF measure. EASPD and CR could only be predicted by CF, with 8% and 7% of the variance predicted. These results suggest that L2 UF in the dialogic task may be overall less reliant on L2 processing ability than in the monologic task. A possible reason is that dialogic speaking inherently puts fewer cognitive demands on speech production and thus is less demanding on L2 processing ability. This assumption garners support from prior findings that learner's L2 performance is more fluent in dialogues than monologues, both in cross-sectional (Michel, 2011; Tavakoli, 2016) and longitudinal studies (Witton-Davies, 2014). In dialogic tasks, speakers are able to plan and assemble their utterances during their interlocutors' turns (Tavakoli, 2016). Additionally, interlocutors may experience synchronization, such as lexical alignment (Shen & Wang, 2025), which facilitates automatic activation and retrieval of lexical items (Pickering & Garrod, 2004), thereby reducing the dependence on L2 CF to some extent.

5.3. Comparing the effects of L2 lexical and syntactic processing efficiency

Lexical and syntactic processing efficiency appeared to be of equal importance in predicting L2 UF in the monologic task. Specifically, L2 lexical processing efficiency (either processing speed or stability) predicted AR (6%), MASPR (14%), and EASPD (6%) and also contributed to predicting an additional 2% of the variance in MASPD based on other predictors. L2 syntactic processing efficiency predicted 5% of the variance in CR and also contributed to predicting additional variance in MASPR (6%) and RR (13%). This is consistent with Kahng's (2020) finding that lexical and syntactic processing speed had an even importance in contributing to L2 UF in monologic speaking.

However, in dialogic speaking, most of the measures that could be predicted by L2 CF (MASPR, EASPD, and CR) were simply predictable by lexical processing speed or stability. This reinforces the argument that, due to the inherently interactive nature, dialogues necessitate less elaborate syntactic structures in utterances compared to monologues (Tavakoli & Wright, 2020). Additionally, these findings highlight the essential role of lexical access in L2 speech production, regardless of monologic and dialogic contexts. As Segalowitz (2010) claimed, the fundamental aspect of language use lies in the ability to establish connections between word forms and their meanings.

5.4. Comparing the effects of L2 processing speed and stability

We found that in both monologic and dialogic tasks, processing stability showed less predictive power compared to processing

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

speed. In the monologic task, lexical and syntactic processing speed had significant predictive effects on a range of UF measures, predicting 14%, 6%, and 5% of the variance in MASPR, EASPD, and CR, and also predicting an additional 2% and 9% of the variance in MASPD and RR based on other factors. However, the processing stability measures could only significantly predict AR with 6% of the variance predicted, as well as MASPR and RR, with an additional 6% and 4% of the variance predicted. In the dialogic task, only processing speed could predict L2 UF, while no processing stability measure showed significant effects.

These results suggest that L2 processing speed may be enough to indicate L2 CF. The limited predictive power of L2 processing stability may also be attributed to the processing tasks used in the study, which might have placed relatively low demands on cognitive processing. In this case, individual differences would be more pronounced in the speed of cognitive processes than in how stable the processing speed is. Future studies should explore other possible measures of L2 processing stability and examine its role in predicting UF. They may examine the effect using other processing tasks, such as speeded lexical decision tasks employed in the study of Akamatsu (2008) and th sentence verification task used by Lim and Godfroid (2015).

Another important point to note is the current level of development in L2 processing among participants in this study. According to Segalowitz (2010) and studies validating CV as a measure of processing efficiency, the correlation between CV and RT can indicate automatization development (e.g., Lim & Godfroid, 2015). As automatization occurs, subcomponents that previously required more attentional effort are restructured, leading to greater organizational efficiency. This further results in a more significant reduction in the SD of processing time compared to the proportional reduction of SD to the reduction in RT, and thus a decrease in CV. In other words, both CV and RT decrease during automatization, producing a positive correlation between the two. In the current study, for participants providing data for RQ1, the Pearson correlation between mean RT and CV (corrected based on domain-general processing RT and CV) for the lexical processing task was −.01 (p = .946), and for the syntactic processing task was .14 (p = .142). For participants providing data for RQ2, the correlation between mean RT and CV for the lexical processing task was .12 (p = .222), and for the syntactic processing task was .04 (p = .666). The pattern of data is insufficient to support that there were skill differences reflecting automatization in lexical and syntactic processing. In other words, participants may not have reached a level where automatization has started to become relevant to individual differences.

5.5. Contribution of L2 knowledge to L2 utterance fluency

The importance of L2 knowledge in predicting L2 UF is non-negligible. In both monologic and dialogic tasks, only the comprehensive knowledge of vocabulary and grammar played a significant role, with vocabulary size and phrasal expression size having no significant effects. This suggests that the role of knowing how many words and phrasal expressions differ from that of knowing how to use these words, phrases, and grammar in context. The vocabulary depth knowledge and knowledge of grammar use are more important than simply size of vocabulary and phrases for fluent speech production. In Kahng's (2020) study, L2 knowledge measures included vocabulary depth knowledge, grammar knowledge, and phrasal expression size, among which phrasal expression size was the only significant predictor of L2 UF. In the current study, measures of the depth of vocabulary knowledge and grammar knowledge were combined into a single measure to avoid inaccurate delineation between the two. This practice might lead to the detection of a far more pronounced predictive effect of the comprehensive knowledge of vocabulary and grammar than phrasal expression size.

In the monologic task, the knowledge measure predicted an additional 5% of the variance in MASPR, an additional 10% of the variance in MASPD, and 18% of the variance in RR. It should be noted that more variance in MASPD and RR was predicted by the knowledge measure than CF. In the dialogic task, L2 knowledge also had a noteworthy effect on UF. Five of nine UF measures, including AR, MASPR, EASPR, MASPD, and FPR, could be predicted by the L2 knowledge measure. Among these measures, AR and EASPR had the L2 knowledge measure as the only significant predictor, with 9% and 7% of the variance predicted. For the measures that could be predicted by both L2 knowledge and CF, including MASPR and MASPD, the L2 knowledge measure contributed more than CF measures, with 13% and 9% of the variance predicted. These results suggest that L2 comprehensive knowledge of vocabulary and grammar is indispensable for fluent speaking performance. In this study, participants' UF during dialogic speaking was even more reliant on their L2 knowledge than on L2 CF. A possible explanation is that while both L2 CF and L2 knowledge contribute to L2 UF, dialogic speaking has lower demands for automatic L2 processing (Cameron, 2001; Michel, 2011; Tavakoli, 2016), thus reducing the predictive power of L2 CF while making the effect of L2 knowledge more pronounced.

5.6. Contribution of L1 utterance fluency to L2 utterance fluency

In addition to L2 CF and knowledge, some L2 UF features were also predicted by L1 UF. In the monologic task, L1 UF could predict MASPR, MASPD, FPR, and RR. In the dialogic task, L1 UF continued to predict FPR and RR. It was noteworthy that in both speaking contexts, L1 FPR contributed to most of the variance in L2 FPR (42% of the variance predicted in the monologic task; 19% of the variance predicted in the dialogic task).

These results partly echo the finding that L1 and L2 UF are strongly correlated with each other, especially in breakdown fluency (e.g., Gao & Sun, 2024). However, in the present study, after accounting for the influence of L2 CF and knowledge, the L1 influence was overall weak on L2 UF except for FPR, regardless of speaking contexts. This suggests that while there is an L1 impact on L2 UF, the L1 influence is generally weaker than the impact of L2 CF and L2 knowledge. One possible explanation for the findings on FPR is that filled pauses are heavily influenced by individual speaking habits and strategies. From a problem-solving perspective on speech production, filled pauses, along with other speech features like fillers, drawls, and repetitions, are regarded as stalling mechanisms that help speakers avoid silence while planning speech (Dörnyei & Kormos, 1998; Peltonen, 2017). According to Peltonen (2018), the use of stalling mechanisms follows highly idiosyncratic patterns in both L1 and L2, which may explain why L2 FPR in the present study was largely influenced by L1 rather than L2 CF or knowledge. This pattern was not observed for L2 RR, likely due to the lower frequency of repetitions compared to filled pauses.

6. Conclusion

The study investigated the relationships between L2 UF and CF in L2 monologic and dialogic tasks. In both speaking contexts, L2 CF was a significant factor underlying UF, though the predictive power

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

was relatively limited. It had a prevalent impact on different L2 UF measures, especially in the monologic task. A detailed inspection of the relationships revealed that L2 lexical processing efficiency was as important as syntactic processing efficiency in the monologic task but more important in the dialogic task for L2 UF. In both contexts, L2 processing speed had a more significant impact on UF than L2 processing stability. L2 knowledge and L1 UF were also non-negligible factors that influenced L2 UF.

The study has several limitations. First, we adopted the ASU-based approach defined by Foster et al. (2000) to code silent pause positions. While this approach has been widely adopted, it may not fully capture the differences between clause boundary pauses and those within clauses in relation to speech production processes. Future studies are encouraged to explore the effects of applying an ASU-based versus a clause-based approach on fluency measures and the findings of this study. Second, we employed a series of tasks to measure participants' L2 lexical, syntactic, and domain-general information processing efficiency. While trials within each task were randomized, the potential effect of task order was not accounted for. Future studies should control for this order effect. Third, participants' pronunciation knowledge was not assessed or included as a variable, which should be addressed in future research. Fourth, we did not find a significant correlation between L2-specific RT (processing speed) and CV (processing stability) measures. This may indicate that participants in our sample had not yet achieved a level of L2 proficiency where an association between UF and CF could be detected underlying individual differences. In other words, they may still be in earlier stages of acquisition, where such a relationship has yet to be established. Future studies may benefit from using this correlation as a screening indicator when selecting more advanced L2 learners. Finally, the study did not include a statistical comparison of the variance in the L2 UF-CF link between the monologic and dialogic tasks, due to the limited number of overlapping participants and differences in task characteristics – the monologic task being open-ended and the dialogic task more structured with specific prompts. Future studies should aim to make direct comparisons by using a larger sample size and more carefully designed tasks.

Theoretically, the study extends the investigation of the link between L2 UF and CF, situating it in an underexplored speaking context, dialogic speaking. Based on the current study's findings, several adjustments and expansions can be proposed regarding the relationship between L2 UF and CF, as assumed by Segalowitz (2010, 2016): (a) the relationship varies between monologic and dialogic speaking due to the differences in contextual characteristics, (b) the impact of L2 CF is notable in shaping L2 UF, yet its magnitude of influence remains constrained. Given that in both speaking contexts, L2 knowledge and L1 UF also had significant impacts on certain L2 UF measures, the study posits that while L2 CF holds importance, it does not stand as the determinant factor underlying UF performance.

Practically, the study unravels the cognitive and L1 factors shaping L2 UF, holding implications for L2 speaking training. The study suggests that lexical processing speed is crucial for fluent monologic and dialogic performance. Accordingly, the study emphasizes the importance of deliberate training on lexical processing through practicing vocabulary tasks involving active engagement from learners (see Akamatsu, 2008; Fukkink et al., 2005; Pellicer-Sánchez, 2015 for details). Additionally, among the components of L2 knowledge, the comprehensive knowledge of vocabulary and grammar was found to be the most important. Therefore, L2 instruction should shift its focus to enhancing learners' vocabulary depth and improving their ability to understand and use grammar in context. Finally, L1 UF was found to influence certain L2 UF features, particularly evident in the monologic task. Although the influence was overall limited and only evident for L2 filled pauses, we suggest that future studies explore the possible benefits of L1 fluency training on L2 performance.

Supplementary material. Visit https://osf.io/5m46e/ for all supplemental files.

Data availability statement. Data are available from the authors upon reasonable request.

Acknowledgments. We thank members of the Speech, Evaluation, and Acquisition Lab at Zhejiang University.

Funding statement. This study is supported by the National Social Science Foundation of China (Grant no. 19CYY008) and the Fundamental Research Funds for the Central Universities in China.

Competing interests. The authors declare none.

Ethical standard. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

References

Akamatsu, N. (2008). The effects of training on automatization of word recognition in English as a foreign language. *Applied PsychoLinguistics*, **29** (2), 175–193. https://doi.org/10.1017/S0142716408080089.

Alderson, J. C. (2005). Diagnosing foreign language proficiency: The interface between learning and assessment. *Continuum*. https://doi.org/10.4324/9780429431463-6.

Alderson, J. C., & **Huhta, A.** (2005). The development of a suite of computer-based diagnostic tests based on the common European framework. *Language Testing*, **22** (3), 301–320. https://doi.org/10.1191/0265532205lt310oa.

Beglar, D. (2010). A Rasch-based validation of the vocabulary size test. *Language Testing*, **27**(1), 101–118. https://doi.org/10.1177/0265532209340194.

Boersma, P., & **Weenink, D.** (2020). *Praat: Doing phonetics by computer* (Version 6.1.36) [Computer software]. https://www.praat.org/

Cameron, D. (2001). *Working with spoken discourse*. Sage.

Chen, X., **Hu, N.**, **Wang, Y.**, & **Gao, X.** (2020). Validation of a brain-computer interface version of the digit symbol substitution test in healthy subjects. *Computers in Biology and Medicine*, **120**, 103729. https://doi.org/10.1016/j.compbiomed.2020.103729.

Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, **26** (3), 367–396. https://doi.org/10.1177/0265532209104667.

Derwing, T. M., **Munro, M. J.**, **Thomson, R. I.**, & **Rossiter, M. J.** (2009). The relationship between L1 fluency and L2 fluency Development. *Studies in Second Language Acquisition*, **31**(4), 533–557. https://doi.org/10.1017/s0272263109990015.

De Jong, N. H., & **Bosker, H. R.** (2013). Choosing a threshold for silent pauses to measure second language fluency. *Proceedings of DiSS* **2013**, 17–20. https://www.isca-speech.org/archive/pdfs/diss_2013/jong13_diss.pdf

De Jong, N. H., **Groenhout, R.**, **Schoonen, R.**, & **Hulstijn, J. H.** (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied PsychoLinguistics*, **36** (2), 223–243. https://doi.org/10.1017/S0142716413000210.

De Jong, N. H., **Steinel, M. P.**, **Florijn, A.**, **Schoonen, R.**, & **Hulstijn, J. H.** (2013). Linguistic skills and speaking fluency in a second language. *Applied PsychoLinguistics*, **34** (5), 893–916. https://doi.org/10.1017/S0142716412000069.

De Jong, N. H., & **Wempe, T.** (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, **41**, 385–390. https://doi.org/10.3758/BRM.41.2.385.

Dörnyei, Z., & **Kormos, J.** (1998). Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in Second Language Acquisition*, **20** (3), 349–385. https://doi.org/10.1017/S0272263198003039.

Feng, R. (2022). Cognitive factors influencing utterance fluency in L2 dialogues: Monadic and non-monadic perspectives. *Frontiers in Psychology*, **13**, 926367. https://doi.org/10.3389/fpsyg.2022.926367.

Foster, P., **Tonkyn, A.**, & **Wigglesworth, G.** (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, **21**, 354–375. https://doi.org/10.1093/applin/21.3.354.

Fukkink, R. G., **Hulstijn, J.**, & **Simis, A.** (2005). Does training in second-language word recognition skills affect reading comprehension? An experimental study. *The Modern Language Journal*, **89** (1), 54–75. https://doi.org/10.1111/j.0026-7902.2005.00265.x.

Gao, J., & **Sun, P. P.** (2023). To correct or not: The role of L1 fluency in understanding and measuring L2 fluency. *TESOL Quarterly*, **57** (2), 643–655. https://doi.org/10.1002/tesq.3198.

Gao, J., & **Sun, P. P.** (2024). How does learners' L2 utterance fluency relate to their L1? A meta-analysis. *International Journal of Applied Linguistics*, **34** (1), 276–291. https://doi.org/10.1111/ijal.12493.

Gao, J., **Sun, P. P.**, & **Li, C.** (2025). Exploring the optimal thresholds of silent pauses for measuring second language utterance fluency in monologic and dialogic speaking. *Language Testing*, **42**(3), 283–311. https://doi.org/10.1177/02655322251315792.

Ginther, A., **Dimova, S.**, & **Yang, R.** (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, **27** (3), 379–399. https://doi.org/10.1177/0265532210364407.

Huensch, A., & **Tracy-Ventura, N.** (2017). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycho-Linguistics*, **38** (4), 755–785. https://doi.org/10.1017/S0142716416000424.

Johnston, R. A., **Dent, K.**, **Humphreys, G. W.**, & **Barry, C.** (2010). British-English norms and naming times for a set of 539 pictures: The role of age of acquisition. *Behavior Research Methods*, **42** (2), 461–469. https://doi.org/10.3758/BRM.42.2.461.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, **64** (4), 809–854. https://doi.org/10.1111/lang.12084.

Kahng, J. (2020). Explaining second language utterance fluency: Contribution of cognitive fluency and first language utterance fluency. *Applied PsychoLinguistics*, **41** (2), 457–480. https://doi.org/10.1017/S0142716420000065.

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, **40**, 387–417. https://doi.org/10.1111/j.1467-1770.1990.tb00669.x.

Lim, H., & **Godfroid, A.** (2015). Automatization in second language sentence processing: A partial, conceptual replication of Hulstijn, Van Gelderen, and Schoonen's 2009 study. *Applied PsychoLinguistics*, **36** (5), 1247–1282. https://doi.org/10.1017/S0142716414000137.

Lintunen, P., **Mutta, M.**, & **Peltonen, P.** (2020). *Fluency in L2 learning and use*. Multilingual Matters.

Liu, Y., **Hao, M.**, **Li, P.**, & **Shu, H.** (2011). Timed picture naming norms for mandarin Chinese. *PLoS One*, **6** (1), e16505. https://doi.org/10.1371/journal.pone.0016505.

Martinez, R. (2011). *The development of a corpus-informed list of formulaic sequences for language pedagogy*. [Doctoral dissertation, University of Nottingham]. https://eprints.nottingham.ac.uk/12963/1/555398.pdf

Martinez, R., & **Schmitt, N.** (2012). A phrasal expressions list. *Applied Linguistics*, **33**(3), 299–320. https://doi.org/10.1093/applin/ams010.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, **1**, e4. https://doi.org/10.1017/s2041536210000012.

Michel, M. (2011). Effects of task complexity and interaction in L2 performance. In P. Robinson (Ed.), *Second language task complexity: Researching the cognition hypothesis of language learning and performance* (pp. 141–174). John Benjamins. https://doi.org/10.1075/tblt.2.12ch6

Nation, P. & **Beglar, D.** (2007). A vocabulary size test. *The Language Teacher*, **31**(7), 9–13. http://www.jalt-publications.org/archive/tlt/2007/07_2007TLT.pdf

Nguyen, L. T. C., & **Nation, P.** (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal*, **42** (1), 86–99. https://doi.org/10.1177/0033688210390264.

Olkkonen, S., **Snellings, P.**, **Veivo, O.**, & **Lintunen, P.** (2024). Cognitive fluency in L2: The effect of automatic and controlled lexical processing on speech rate. *Journal of Psycholinguistic Research*, **53**, 66. https://doi.org/10.1007/s10936-024-10099-0.

Ortega, L., **Iwashita, N.**, **Norris, J. M.**, & **Rabie, S.** (2002, October). *An investigation of elicited imitation tasks in crosslinguistic SLA research*. [Paper presentation]. The 20th Second Language Research Forum, Toronto, Canada. https://www.iris-database.org/details/HAPZ2-KRwXl

Peirce, J., **Gray, J. R.**, **Simpson, S.**, **MacAskill, M.**, **Höchenberger, R.**, **Sogo, H.**, **Kastman, E.**, & **Lindeløv, J. K.** (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, **51** (1), 195–203. https://doi.org/10.3758/s13428-018-01193-y.

Pellicer-Sánchez, A. (2015). Developing automaticity and speed of lexical access: The effects of incidental and explicit teaching approaches. *Journal of Spanish Language Teaching*, **2** (2), 126–139. https://doi.org/10.1080/23247797.2015.1104029.

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

Peltonen, P. (2018). Exploring connections between first and second language fluency: A mixed methods approach. *The Modern Language Journal*, **102** (4), 676–692. https://doi.org/10.1111/modl.12516.

Pickering, M. J., & **Garrod, S.** (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Science*, **27**, 169–226. https://doi.org/10.1017/S0140525X04000056.

Pickering, M. J., & **Garrod, S.** (2021). *Understanding dialogue: Language use and social interaction*. Cambridge University Press.

Plonsky, L., & **Oswald, F. L.** (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, **64** (4), 878–912. https://doi.org/10.1111/lang.12079.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching*, **54** (2), 79–95. https://doi.org/10.1515/iral-2016-9991.

Segalowitz, N., & **Freed, B. F.** (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition*, **26** (2), 173–199. https://doi.org/10.1017/S0272263104262027.

Shen, H., & **Wang, M.** (2025). Effects of interlocutors' linguistic competence on L2 speakers' lexical alignment. *Bilingualism: Language and Cognition* Advance online publication. https://doi.org/10.1017/S1366728924000725.

Skehan, P. (2003). Task-based instruction. *Language Teaching*, **36** (1), 1–14. https://doi.org/10.1017/S026144480200188X.

Snodgrass, J. G., & **Vanderwart, M.** (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, **6** (2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174.

Suzuki, S., & **Kormos, J.** (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, **45** (1), 38–64. https://doi.org/10.1017/S0272263121000899.

Suzuki, Y., & **Sunada, M.** (2018). Automatization in second language sentence processing: Relationship between elicited imitation and maze tasks. *Bilingualism: Language and Cognition*, **21** (1), 32–46. https://doi.org/10.1017/S1366728916000857.

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching*, **54** (2), 133–150. https://doi.org/10.1515/iral-2016-9994.

Tavakoli, P. (2018). L2 development in an intensive study abroad EAP context. *System*, **72**, 62–74. https://doi.org/10.1016/j.system.2017.10.009.

Tavakoli, P., & **Skehan, P.** (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance* (pp. 239–273). John Benjamins.

Tavakoli, P., & **Wright, C.** (2020). *Second language speech fluency: From research to practice*. Cambridge University Press.

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press

Trouvain, J., & **Werner, R.** (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. In C. Schwarze, & S. Grawunder (Eds.), *Transkription und annotation gesprochener sprache und multimodaler interaktion: Konzepte, probleme, lösungen* [] (pp. 55–73). Gunter Narr Verlag.

Williams, S. A., & **Korko, M.** (2019). Pause behavior within reformulations and the proficiency level of second language learners of English. *Applied Psycholinguistics*, **40**(3), 723–742. https://doi.org/10.1017/S0142716418000802.

Witton-Davies, G. (2014). *The study of fluency and its development in monologue and dialogue* [Doctoral dissertation, University of Lancaster]. https://taiwan.academia.edu/GilesWittonDavies

Wu, S. L., **Tio, Y. P.**, & **Ortega, L.** (2022). Elicited imitation as a measure of L2 proficiency. *Studies in Second Language Acquisition*, **44** (1), 271–300. https://doi.org/10.1017/S0272263121000103.

Yan, X. (2015). *The processing of formulaic language on elicited imitation tasks by second language speakers* [Doctoral dissertation, Purdue University] https://docs.lib.purdue.edu/open_access_dissertations/597

Yan, X., **Kim, H. R.**, & **Kim, J. Y.** (2021). Dimensionality of speech fluency: Examining the relationships among complexity, accuracy, and fluency (CAF) features of speaking performances on the Aptis test. *Language Testing*, **38** (4), 485–510. https://doi.org/10.1177/0265532220951508.

Yan, X., **Maeda, Y.**, **Lv, J.**, & **Ginther, A.** (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, **33** (4), 497–528. https://doi.org/10.1177/0265532215594643.

Zhao, P., & **Ji, X.** (2016). A validity study on VST: Based on classical test theory and item response theory. *Foreign Language Testing and Teaching*, **2**, 39–46 59.

https://doi.org/10.1017/S1366728925100564 Published online by Cambridge University Press