Research Methods in Applied Linguistics 4 (2025) 100176

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Introducing fluency measures to the elicited imitation task

Hui Sun [a,b], Dagmar Divjak [b,c], Petar Milin [b,*]

[a] School of English, Communication and Philosophy, Cardiff University, Cardiff, UK
[b] Department of Modern Languages, University of Birmingham, Birmingham, UK
[c] Department of Linguistics and Communication, University of Birmingham, Birmingham, UK

ARTICLE INFO

Keywords:
Written elicited imitation
Writing fluency
Keystroke logging
Automatization
L2 grammar

ABSTRACT

The elicited imitation (EI) task has been widely used as a measure of automatized L2 knowledge. However, the scoring of the task has relied exclusively on product-based measures (i.e., accuracy of L2 production), without considering any process-based indices of automatization, such as fluency. To fill this gap, our study develops a written version of the EI task and innovatively draws on keystroke logging techniques to introduce new measures of fluency in EI production. To test whether the addition of fluency measures improves task sensitivity, we examined the degree to which fluency and accuracy predicted L2 proficiency among 40 L1 Polish speakers of English, living in the UK (Mage = 31, 20–60). The participants were late learners of English at intermediate-to-advanced level (CEFR B1–C2) with varying lengths of residence (0.5–18 years). Their L2 proficiency was measured through self-evaluation according to CEFR scales and through test-evaluation by DIALANG English grammar and vocabulary tests. Their written production of English article and tense-aspect target structures was coded for grammatical accuracy, speed and pausing fluency, and consistency in fluency. Generalized Additive Modelling revealed a nonlinear interaction between speed and pausing fluency and grammatical accuracy as predictors of self-evaluated (but not test-evaluated) proficiency, which suggests that participants tended to be more accurate and fluent from low to average proficiency, after which their accuracy plateaued while fluency continued to improve. The results support the importance of assessing fluency to maintain the sensitivity of the EI task especially among advanced learners.

Introduction

In second language (L2) research, the elicited imitation (EI) task is considered a reliable measure of global oral proficiency and implicit grammatical knowledge (for meta-analyses, see Kostromitina & Plonsky, 2022; Yan et al., 2016). Although the task design varies, EI usually involves having participants listen to a sentence and repeat it orally as accurately as possible. It is assumed that, to reproduce the sentence, participants need to draw on implicit linguistic knowledge to decode the audio stimuli and comprehend the meaning. This assumption is supported by empirical evidence (Suzuki & DeKeyser, 2015), but debate remains as to whether it is implicit (without awareness) or explicit but automatized knowledge (with awareness), or a combination of both, that is engaged at the reproduction stage (Ellis, 2005; Erlam, 2006; Granena, 2016; Spada et al., 2015; Suzuki & DeKeyser, 2015; Suzuki et al., 2023). Nevertheless, it is agreed that EI tasks assess an individual's ability to process and reconstruct linguistic structures in an automatic (fast and effortless) manner, which suggests some degree of automatization in the process. The existing scoring methods of EI tasks,

[*] Corresponding author at: University of Birmingham, B15 2TT, UK.
E-mail address: P.Milin@bham.ac.uk (P. Milin).

https://doi.org/10.1016/j.rmal.2024.100176
Received 3 March 2024; Received in revised form 27 November 2024; Accepted 30 November 2024
Available online 12 December 2024

however, primarily focus on the accuracy of the final product, without considering the fluency (e.g., speed, pausing) which is a crucial process-based index of automatization (Suzuki & Révész, 2023).

Meta-analyses have demonstrated that the sensitivity of EI tasks is influenced by the scoring method (Kostromitina & Plonsky, 2022; Yan et al., 2016). Specifically, ordinal rating scales (e.g., assessments of overall sentence repetition quality; Gaillard & Tremblay, 2016; Tracy-Ventura et al., 2014) tend to predict proficiency levels more effectively than binary scales (e.g., evaluations of correctness in repeating specific structures; Erlam, 2006). This effect could be two-fold, prompted by the nature of the scale and/or the nature of the measured construct (global vs. specific). Ordinal rating scales are generally employed to evaluate global proficiency, whereas binary coding methods target specific constructs. Although Kostromitina and Plonsky (2022) confirmed that the nature of the measured construct does not significantly affect task sensitivity, several studies have shown that, generally speaking, binary variables can inflate correlation coefficients compared to ordinal or continuous variables (e.g., Agresti, 2013; Cohen, 1983), and thus are statistically less efficient or accurate. This inflation can, in turn, distort conclusions about the strength of relationships with other variables under investigation.

We investigate whether fluency could play a role here: it is possible that the ordinal rating of oral reproduction based on audio recording (e.g., Gaillard & Tremblay, 2016) reflects fluency to some degree, something that is missed by the cruder binary coding approach. This study takes first steps towards introducing fluency measures (e.g., speed and pausing) into the EI task, by adapting sentence reproduction to written format and integrating the keystroke logging technique to record time-related data (Miller et al., 2008). We then examine the extent to which fluency in performing the EI task, alongside accuracy, improves sensitivity to individual variation within traditional proficiency bands by more precisely accounting for factors such as individual differences and task-specific performance.

EI tasks as a measure of automatized grammatical knowledge

A series of validation studies has confirmed the reconstructive nature (vs. rote repetition) of the EI task and its capacity to measure automatized (explicit or implicit) knowledge of specific grammatical structures (Ellis, 2005; Erlam, 2006; Granena, 2016; Spada et al., 2015; Suzuki & DeKeyser, 2015). The reconstructive nature of the task is supported by certain task features. For example, it is possible to imitate a sequence of novel sound strings if the strings are short enough and can be repeated immediately, but with a 3-second delay this becomes rather difficult, unless the meaning of the strings is already understood (McDade et al., 1982). To prevent rote imitation, the length of sentence stimuli in EI tasks are set to at least 7 syllables or 2 s, which is thought to exceed the average short-term memory span (Perkins et al., 1986). Additionally, a delay in response is typically required. This delay is enforced through a 3-second period of silence or an interruptive task, such as counting down from three out loud (e.g., Suzuki & Sunada, 2018), or a True/False judgement question (e.g., Erlam, 2006; Granena, 2016). Thus, for participants to accurately repeat a sentence they have never heard after a delay, they need to firstly process and comprehend its meaning in real time drawing on their internal language system, and then reconstruct the sentence based on the meaning they have retained.

The reconstructive nature of EI tasks can be further enhanced by directing participants' attention to the meaning of the sentences they are asked to reproduce. A meaning-focused instruction of the task, such as describing it as a True/False judgement task (Erlam, 2006), would prevent participants from focusing their attention on the form of the target structures. The aforementioned validation studies also used ungrammatical stimuli to elicit spontaneous correction of grammatical errors, which provide evidence against the reliance on rote repetition. At the reproduction stage, in some studies time pressure was exerted to trigger greater reliance on automatized grammatical knowledge (Ellis, 2005; Erlam, 2006; Spada et al., 2015).

Indeed, with the task design described above, the accuracy scores of oral EI have been found to correlate with other time-pressured tasks or spontaneous proficiency tests which tap into automatized knowledge (e.g., oral narrative task, IELTS test, timed grammatical judgement test/GJT), but not with untimed tests of metalinguistic knowledge which tap into explicit knowledge (e.g., untimed GJT, error correction task) or working/short-term memory tests (Ellis, 2005; Erlam, 2006; Granena, 2016; Kim et al., 2016; Spada et al., 2015; but see Suzuki & DeKeyser, 2015 for evidence of EI task as a measure of automatized explicit knowledge, associated with metalinguistic knowledge). Even though the EI task is designed to measure automatized grammatical knowledge (i.e., retrieved and processed rapidly, stably, and effortlessly), process-based indices of automatization (processing fluency and consistency) have not yet been taken into account. The next section will overview measures of automatization in L2 production.

Automatization in L2 production

The construct of L2 automatization, suggesting reduced cognitive efforts and/or attentional demands, is often measured through processing fluency (speed and pausing) and processing consistency (coefficient of variation) (Segalowitz & Segalowitz, 1993; Suzuki & Révész, 2023). That is, automatized L2 processing is characterized by being fast and effortless as well as stable and consistent.

In their synthesis of 19 primary studies, Suzuki and Révész (2023) identified key process-based measures for utterance and writing fluency. For utterance fluency, they are speed (higher articulation rate), breakdown (fewer silent pauses), and composite fluency (higher speech rate or mean length of run). Writing fluency, on the other hand, is indicated by speed (more words/characters per minute), pause duration (shorter pause length between sentences), and pause frequency (smaller number of pauses between words). Compared with commonly used spontaneous writing tasks (e.g., argumentative essays; Révész et al., 2022), the written EI task is more restricted due to the sentence stimuli. However, as participants do not know the sentences they would be exposed to in advance while performing in real time (see García-Amaya & Cintrón-Valentín, 2021 for using written EI task for a similar purpose), the EI production also taps into automatized L2 knowledge and is thus appropriate for measuring writing fluency.

In exploring whether processing consistency can also indicate automatization, Segalowitz and Segalowitz (1993) introduced the use of the coefficient of variation (CV). This metric serves to distinguish between general *speed-up* and *true automatization*, where the former is indicated by a similar rate of decrease in the mean and deviation of reaction time latencies (RT), while the latter shows a steeper decrease in the deviation (see Hulstijn et al., 2009 for a thorough discussion). The use of CV, which is often called relative standard deviation being a simple ratio of standard deviation and the mean, provides a measure of normalized deviation and helps comparisons of decrease in average speed vs. deviation in speed: a positive correlation between the two measures indicates true automatization. Although the findings about the link between reduction in CV and progression in proficiency/experience levels have been mixed (Hulstijn et al., 2009; Lim & Godfroid, 2015; McManus & Marsden, 2019; Suzuki & Sunada, 2018), we included this measure in our study given its conceptual relevance.

The current study

To summarize, the EI task is designed to and has been widely used to assess automatized L2 grammatical knowledge. However, this type of knowledge is unlikely to be fully captured by the conventional accuracy measures (correctness of the final product/responses) alone. Therefore, this study proposes to add process-based fluency measures to the EI task by including processing fluency and consistency of responses. To this end, a written version of the English EI task was developed to record temporal data efficiently using keystroke logging software (i.e., InputLog; Leijten & Van Waes, 2013).

Using this task, we aim to test if a combination of accuracy and fluency measures can offer a more detailed and nuanced account of L2 proficiency. In line with the meta-analyses (Kostromitina & Plonsky, 2022; Yan et al., 2016), we examined the correlations between EI measures and other established English proficiency measures as an index of task sensitivity. Thus, the better the EI measures predict proficiency levels assessed by other proficiency tests, the higher the sensitivity of the EI task.

More specifically, the research question of this study is: To what extent do accuracy and fluency measures of the EI task predict proficiency levels? It is hypothesized that participants with better accuracy scores from the EI task would be of higher proficiency levels, as reported in previous EI studies; they would be expected to demonstrate greater fluency and consistency on the task at hand.

Methods

Participants

Participants in this study were 40 Polish L1 speakers of L2/LX English (33 women, 7 men; $M_{age} = 31$, $SD_{age} = 8.7$), recruited in Birmingham, UK via posters, emails, and social media (e.g., Facebook, Twitter). They had moved to the UK after the age of 15 ($M = 22.7$, $SD = 5.7$) to study, work or join their families, and their length of residence ranged from 0.5 to 18 years ($M = 8.5$, $SD = 5.5$). Overall, they had attained intermediate to advanced level of English proficiency. Before moving to the UK, all participants had studied English at school in Poland for at least 4 years ($M = 10.7$, $SD = 3.6$) where they had learned grammar through formal instruction; 13 of them had also taken English classes in the UK.

Procedures and tasks

This study is part of a large project where participants completed a series of language and cognitive tasks; participation was compensated with Amazon vouchers. The tasks relevant to this study include a written EI task which was administered in the lab, alongside a self-assessment task for English proficiency (Council of Europe, 2001), the DIALANG 2.0 language proficiency tests (Alderson & Huhta, 2005) and a nonword recognition task (O'Brien et al., 2006) which were completed online.

Written elicited imitation task

We designed a written elicited imitation task in line with recommendations made by Erlam (2006). Our task contained both grammatical and ungrammatical stimuli and required participants to reproduce the stimuli in correct English rather than verbatim. Out of the 17 gmatical structures covered in Erlam's study, we focused on English article and tense-aspect structures specifically, as they are among the most difficult ones (i.e., low correctness rates) and are normally acquired at the intermediate-late stage of learning (cf. Bardovi-Harlig, 2000; Crosthwaite, 2016; Divjak et al., 2023), but rarely ever mastered to perfection in learners whose L1 does not have a similar system. Participants were instructed to listen to a sentence, rate its ease of understanding (i.e., comprehensibility), and then type in the sentence they had heard using InputLog software (Leijten & Van Waes, 2013) under a time limit. Although the EI production task was in written format which makes it easier for participants to monitor their output, by following Erlam's (2006) task design in terms of length of stimuli, delay in response, time pressure, and meaning-oriented instruction, it is reasonable to assume that our written EI task elicited automatized grammatical knowledge. The design of the stimuli, data collection procedure and scoring of the EI task is outlined below.

Stimuli. Each sentence stimulus had one target structure, containing one of three types of articles (indefinite a/an, definite the, zero ø) or one of twelve tense-aspect combinations (present/past/future + simple/progressive/perfect/perfect progressive). To cover different kinds of structures, a pool of 240 text stimuli was created based on 120 sentences extracted from the British National Corpus (BNC; Leech, 1992), the Internet, and Erlam (2006). Each grammatical stimulus had its ungrammatical counterpart—56 stimuli with an article error and 64 stimuli with a tense-aspect error. Two pairs of examples are given below, where a) contains the grammatical stimulus and b) the ungrammatical one. The full list of 240 text stimuli is deposited in the UBIRA repository (https://doi.org/10.25500/edata.bham.00001200).

(1) a. Though she had long retired, she was **a** wonderful teacher. b. *Though she had long retired, she was **ø** wonderful teacher.

(2) a. Frank is optimistic that we **will have sorted** it by today. b. *Frank is optimistic that we **sort** it by today.

The stimuli varied in length, ranging from 8 to 25 syllables, as EI tasks with varying sentence length have been found to differentiate proficiency levels better (Yan et al., 2016), especially considering that most of the participants in our study rated themselves as Advanced. The stimuli were then pseudo-randomly divided into four sets (a1, a2, b1, b2), each containing 60 stimuli counter-balanced in length, grammaticality and type of target structure. The grammatical stimuli in Set a1/b1 had all their ungrammatical counterparts in Set a2/b2, and vice versa. The audio stimuli were recorded on a phone by a male L1 speaker of British English, who read the sentences out loud at a natural speed in a quiet room. The 4 different sets of the EI task were assigned to participants randomly.

Procedure. Participants were instructed to complete two subtasks after listening to each sentence: first, to rate ease of understanding on a 4-point Likert scale (1 = very difficult, 2 = somewhat difficult, 3 = somewhat easy, 4 = very easy); then, to type in the stimulus in correct English. They had four practice trials including two grammatical stimuli and two stimuli with other types of grammatical errors (i.e., 3rd person –s and comparatives, from Erlam, 2006); they were given model responses to indicate how ungrammatical stimuli could be corrected. The audio stimuli were presented using Experiment Builder software by SR Research on Computer 1, through a set of Bose QuietComfort Noise Cancelling QC35 II Over-Ear headphones. The typed responses were collected by InputLog 7, a keystroke-logging program that records all keyboard activities during typing (Leijten & Van Waes, 2013) on Computer 2, with a high-precision Apple A1048 wired keyboard.

The rating task was presented as the main task to direct participants' attention to the meaning of the sentences instead of to the form; this design increased the possibility of eliciting automatized grammatical knowledge. Participants had to wait for at least three seconds before continuing onto the typing task, which created a delay between hearing the sentence and reproducing it, preventing rote repetition of the sound strings. Then, the instructions on screen directed them to Computer 2 on their left, where an empty InputLog document was set up, ready for them to type the sentences they had heard. If they turned earlier, the researcher would stop them. After typing the response, participants turned back to Computer 1 and pressed a button to hear the next sentence.

To minimize reliance on explicit knowledge, participants were told that (a) the sentence should be typed as fast as possible, and (b) once the sentence is complete, it should not be edited. Also, (c) there was a time limit depending on sentence lengths (2 s per syllable, 20 s minimum), and typing should stop immediately once a bell sound was heard. The next trial was set to start automatically five seconds after the bell sound. Participants were encouraged to produce a complete sentence that would make sense and were also asked to number the sentences from 1 to 64 to make it easier for scoring. Two computers were used because switching to InputLog on the same computer would interrupt the operation of Experiment Builder, and the body movement of participants could help the researcher monitor the typing phase. The EI task took about 45 min to complete.

Scoring of accuracy. The accuracy of sentence repetition was scored using a binary coding approach according to three categories, following the method in Erlam (2006, p. 479).

obligatory occasion created – supplied;
obligatory occasion created – not supplied;
no obligatory occasion created.

A correct response required the obligatory occasion for the target structure and the correct use of the target structure (Category 1). If the target structure was not supplied correctly (Category 2), or the essential obligatory occasion failed to be created (Category 3), the response was deemed as incorrect, as it suggests that the participant had not internalized the target structures.

The article target structure included the article, the noun and the adjective between them (if any). If the article was part of a phrase (e.g., "at the time"), the whole phrase was the target structure. This approach allows fluency measures to be generated for target structures with zero articles. The tense-aspect target structure comprised the main verb, its auxiliary verbs, and the adverbs between them (if any). In Example 1-a above, the obligatory occasion and target structure were both "a wonderful teacher"; in Example 2-a, the target structure was "will have sorted" and the obligatory occasion also included "by today". The target structure in each stimulus is annotated within ⟨⟩ in the stimuli list deposited in the UBIRA repository (https://doi.org/10.25500/edata.bham.00001200).

When the stimulus was ungrammatical (Example 2-b), any acceptable correction made the response correct (Example 3), as it showed that the participant had noticed the error and corrected it. When the stimulus was grammatical (Example 2-a), any change would make the response incorrect, even if it was acceptable (Example 4), as it indicated that the participant had not recognized the correct use of the target structure. Lexical accuracy was not considered in the scoring.

(3) Frank is optimistic that it **will be sorted** by today. / Frank is optimistic that we **will sort** it by today. (correct)

(4) Frank is optimistic that he **would have sorted** it **out** by today. (incorrect)

Responses of two participants from each of the four task versions (20 % of total scoring) were scored by two coders, an L1 English speaker and the first author of this study who is an advanced-level L2 English speaker. After the scoring procedure was calibrated, the first coder continued to score the rest of the responses; the scoring was checked and corrected by the second coder.

The proportion of correct responses was calculated for each participant, in terms of all 60 items, 30 error-free items, 30 erroneous items, 28 items targeting article structures, and 32 items targeting tense-aspect structures respectively. In a few cases where the sentence was edited after completion, only the first attempts were considered.

Coding of fluency. For responses that fell in accuracy Category 1 and 2 (i.e., the obligatory occasion was created), fluency was measured in terms of processing fluency and consistency (Segalowitz & Segalowitz, 1993). Drawing on the keystroke data collected, we used a combined measure of speed and pausing fluency at key level[1]—the mean length of the inter-key interval (IKI) during the reproduction of a target structure (if available, correct or not). The general analysis output from InputLog recorded a timestamp in milliseconds for each key being pressed and released (labelled as event_startTime_E_ and event_endTime_E_). The IKI was calculated as the time difference between release of the current key and that of the previous key. Then, for each participant, the mean length of the IKI was averaged across all responses to indicate their processing fluency. This measure reflects the variance in pause duration between and within words and (reversely) the production rate of characters per minute. This measure is very sensitive as it considers pauses shorter than the commonly used 2-second threshold (cf. Baaijen et al., 2012) and is suitable for capturing L2 automatization in the production of target structures, since it also records pausing at lower textual units, which has been found to relate more to lower-order linguistic processes such as lexical retrieval and morphosyntactic encoding (Révész et al., 2019; Suzuki & Révész, 2023).To measure processing consistency, the CV in the length of IKI ($CV_{key} = SD_{key} / Mean_{key}$) within each target structure was calculated and then averaged across all responses.[2]

Copy Task. To control for individual differences in typing skills, participants were asked to take the 5-minute Copy task available on InputLog (Van Waes et al., 2019) before the written EI task. The Copy task consists of seven subtasks where different lengths of text (e.g., keys, phrases, sentences) must be typed for which specific bigrams characteristics are controlled (e.g., bigram frequency, hand combination, key adjacency). The analysis output which can be exported from InputLog summarizes overall typing speed and accuracy. The speed is calculated as the theoretical number of characters per minute (CPM), based on the mean IKI of the targeted bigrams (i.e., 60,000/mean IKI); accuracy is calculated as the proportion of correctly typed targeted bigrams (i.e., correctness). The Copy Task score was computed as the overall CPM times aggregated correctness, as suggested on the InputLog website (https://inputlog-analysis.uantwerpen.be).

Self-assessment task for English proficiency

Participants self-rated their English listening, reading, speaking (spoken production), and writing proficiency by selecting from six can-do statements on the Common European Framework of Reference for Languages (CEFR) Self-assessment Grid developed by the Council of Europe (2001, pp. 26–27). Each statement represents a different CEFR level (A1, A2, B1, B2, C1, C2) for each language skill. Such type of L2 self-assessment has been found to show a moderate correlation with external measures of the same construct (see Li & Zhang, 2021 for a meta-analysis). While 27 participants self-rated all four skills as advanced (C1–C2), the remaining13 of them rated at least one skill as intermediate (B1–B2).

DIALANG language proficiency test

Besides the self-rated evaluations, participants' English proficiency levels were also assessed through the vocabulary and grammar subtests of DIALANG 2.0 (Alderson & Huhta, 2005), an online diagnostic test (https://dialangweb.lancaster.ac.uk) that implements the Common European Framework of Reference for Languages (CEFR). For vocabulary proficiency, participants were presented with a list of 75 verbs (50 real words, 25 non-words) and asked to judge whether each word was real or not. On completion, they received a score between 0 and 1000, which maps onto a specific CEFR level in the background. Then, their knowledge of a range of grammatical structures (e.g., articles, tense and aspect, third person -s, prepositions, superlatives) was assessed at the recommended CEFR level via 30 questions (e.g., multiple choice, gap filling, word ordering); there was no time pressure. Participants' CEFR scores for grammar proficiency were recorded—5 participants scored B1; 12 scored B2; 20 scored C1; and 3 scored C2.

Nonword recognition task

To prevent rote repetition of sound strings in the EI task, the length of the stimuli was manipulated to exceed short-term memory capacity, and a delay was inserted before the sentence repetition. However, phonological short-term memory (pSTM), the capacity to hold phonological information temporarily, has been found to predict the performance of individuals with less L2 experience, in the EI task with a similar design (Park et al., 2020). Although the participants in this study were mostly long-term residents ($M = 8.5$ years), we measured the pSTM to verify the validity of the EI task.

We created an online version of the nonword recognition task used in O'Brien et al.'s (2006) on Qualtrics, an online survey platform. Participants listened to a pair of nonword lists in each trial and judged whether they were identical or not (i.e., same nonwords presented in the same or different order). They first practiced with four pairs of 4-item lists, two identical and two different. When all responses were correct, they progressed to the main test where eight pairs of 5-item, 6-item, and 7-item lists were presented. The audio files of the 160 nonwords were generated via the text-to-speech service of Microsoft Word, using the built-in "Kate" female British voice on a Macintosh computer. Each nonword list was presented at a rate of one item per 750 ms, with a 1.5-s gap between two

[1] Due to technical issues with InputLog 7, we could not export the summary and pause analysis reports where separate measures of speed and pausing would be available.

[2] Another way could be to calculate the CV for the mean IKI of each target structure across all responses, but the first method would better capture the stability in typing each key of the target structure.

lists. Some of the nonword text (see appendices in Gathercole et al., 2001) was modified to produce more precise computer-generated speech (e.g., “charn” to “charne”). The pSTM score was capped at 24, with one point for each trial. If participants relied on rote memory, the pSTM should be related to their performance in the EI task.

Results

The dataset comprised 2400 data points, with 23 % (557) containing missing values for typing fluency and consistency (Mean$_{\mathrm{key}}$, CV$_{\mathrm{key}}$) due to missing obligatory occasions in written production. The dataset available for analysis consisted of 1843 complete data points. The data and R scripts reported here are available from the UBIRA eData repository (https://doi.org/10.25500/edata.bham.00001200) and on GitHub (https://github.com/ooominds/FluencyMeasuresInElicitedImitationTask), respectively.

Statistical analyses were conducted within the **R** environment for statistical computing (R Core Team, 2023), utilizing the following packages: **car** (v. 3.1–2, Fox & Weisberg, 2019), **psych** (v. 2.3.6, Revelle, 2023), **randomForest** (v. 4.7–1.1, Liaw & Wiener, 2002), **mgcv** (v. 1.9–0, Wood, 2011, 2017), and **itsadug** (v. 2.4.1, Van Rij et al., 2022).

Data preparation and validation

The focus of the data analysis was to address the primary research question concerning the predictive capacity of EI accuracy and fluency for English proficiency measures. Consequently, the analysis centered on examining the relationships among these three groups of variables. Recall that English proficiency was assessed through various measures, distinguishing between (a) self-evaluated proficiency in four fundamental language skills: receptive listening and reading, and productive speaking and writing; and (b) test-evaluated proficiency in vocabulary and grammar. The self-evaluation measures and grammar proficiency were categorized according to standard CEFR ratings (B1, B2, C1, C2), and subsequently treated as ordinal variable (ranks 1–4). Vocabulary scores ranged from 0 to 1000 and were standardized (scaled or z-transformed) for subsequent analyses. The descriptive statistics of these dependent variables are listed in Table 1.

In the initial phase of our analysis, we scrutinized the metric characteristics of the English proficiency measures. Collectively, the six proficiency measures exhibited robust reliability, as indicated by Cronbach’s alpha ($\alpha = 0.78$) and Guttman’s lambda-6 ($\lambda 6 = 0.83$). Furthermore, McDonald’s omega revealed a strong general component with $\omega\_Total = 0.87$, providing evidence for the content and construct validity of the chosen proficiency measures (for details on using omega scores, see Hayes & Coutts, 2020). Importantly, a two-dimensional structure was suggested, distinctly demarcating self-evaluated and test-evaluated components of proficiency, as visually summarized in Fig. 1, with detailed results presented in Appendix A, Table 8.

Subsequently, Principal Component Analysis (PCA) was employed, specifying two components that were rotated orthogonally using varimax rotation. Table 2 succinctly presents the structure of these two extracted components, which jointly accounted for 76 % of the total variance. Participants’ scores on these two components served as our response (dependent) variables: self-evaluated proficiency (Proficiency.PC1) and test-evaluated proficiency (Proficiency.PC2). The correlation between the two scores was confirmed to be $r = 0.0$.

In the EI task, two main predictor variables were identified: accuracy and fluency. We evaluated accuracy through five measures, each expressed as a proportion of correct responses. These included four item-specific accuracy metrics—focusing on grammatical items, ungrammatical items, and items assessing knowledge of English articles and tense and aspect—alongside an overall accuracy measure. For fluency, we employed two metrics: the mean IKI fluency (MeanFluency) and the coefficient of variation in IKI fluency (CVFluency). Additionally, we incorporated three control variables to account for potential confounders: (1) participants’ phonological short-term memory capacity (pSTM), (2) basic typing skills, measured by the Copy Task (CTScore), and (3) the duration of participants’ immersion in English-speaking environments (ImmersionYears). The descriptive statistics of these independent variables are summarized in Table 3.

As the four item-specific accuracy measures exhibited significant intercorrelations, with the minimum correlation coefficient being 0.65, a PCA was run on these measures. The result revealed a dominant first component, which accounted for 87 % of the total variance and had individual loadings exceeding 0.90 (see Table 4), which suggests that all possible accuracy indicators contributed equally to the component. Notably, the correlation between the overall accuracy scores and the first principal component was near perfect ($r = 0.9998,\ p < 0.00001$). In further analyses, we utilized the first principal component scores from the accuracy measures (Accuracy.PC1) as a predictor variable.

Table 1
Descriptive summary of dependent variables.

Proficiency variable	M / Mdn	SD / MAD	Range	95 % interval	Cronbach’s alpha if deleted
Self-eval.					
Listening	3.58 / 4.0	0.78 / 0.0	1–4	[3.33, 3.82]	0.72
Writing	3.28 / 4.0	0.99 / 0.0	1–4	[2.97, 3.58]	0.71
Speaking	3.38 / 4.0	0.93 / 0.0	1–4	[3.09, 3.66]	0.73
Reading	3.38 / 4.0	0.93 / 0.0	1–4	[3.09, 3.66]	0.68
Test-eval.					
Grammar	2.53 / 3.0	0.82 / 0.74	1–4	[2.27, 2.78]	0.80
Vocabulary	783.63 / 806.0	132.48 / 161.6	523–1000	[742.57, 824.68]	0.80

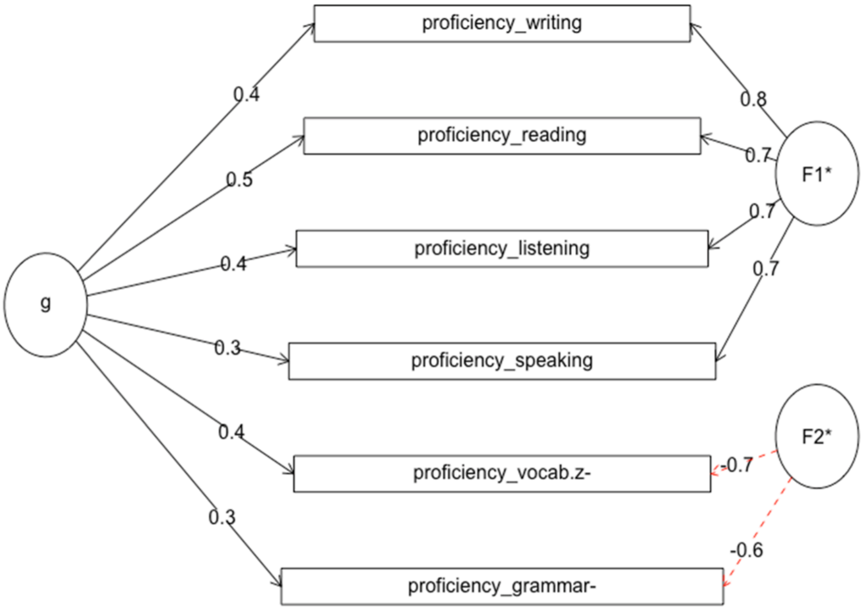

Fig. 1. Omega diagram for six measures of English language proficiency.

Table 2
Principal Component Analysis on measures of English language proficiency, featuring component loadings (PC1, PC2), communality (h2), and uniqueness (u2). Component loadings below an absolute value of 0.3 are not presented.

Proficiency variable	PC1	PC2	h2	u2
Listening	0.84		0.71	0.29
Writing	0.90		0.81	0.19
Speaking	0.83		0.69	0.31
Reading	0.87		0.80	0.20
Grammar		0.87	0.76	0.24
Vocabulary (scaled)		0.87	0.76	0.24

Table 3
Descriptive summary of independent variables.

Independent variables	M	SD	Range	95 % interval
Accuracy				
Grammatical items	0.62	0.17	0.13–0.90	[0.57, 0.67]
Ungrammatical items	0.39	0.14	0.10–0.73	[0.35, 0.44]
Article items	0.55	0.16	0.07–0.79	[0.50, 0.60]
Tense-aspect items	0.47	0.14	0.16–0.75	[0.43, 0.51]
All items	0.51	0.14	0.12–0.77	[0.46, 0.55]
Fluency				
Mean IKI (in milliseconds)	260.58	54.20	159.81–358.86	[243.78, 277.37]
CV in IKI	0.71	0.10	0.47–0.95	[0.68, 0.74]
Control variables				
pSTM	16.55	3.10	9–24	[15.59, 17.51]
Copy Task	246.42	53.58	164.25–395.04	[229.82, 263.03]
Immersion years	8.53	5.45	0.5–18	[6.84, 10.22]

In the next step, we analyzed the distributional properties of all response variables (Proficiency.PC1, Proficiency.PC2), predictor variables (Accuracy.PC1, MeanFluency, CVFluency), and control variables (CTScore, pSTM, ImmersionYears). Where necessary, we applied appropriate transformations to achieve linearity and constant conditional variation (cf. Spanos, 2019). Although the chosen modelling techniques are robust and resilient to non-normality in small or moderate datasets, uneven density can suggest artifactual nonlinearity. Therefore, to facilitate statistical modelling, we followed the approach outlined by Baayen and Milin (2010)). The selection of appropriate transformations was first discussed by Box and Cox (1964), with practical guidelines detailed by Fox and Weisberg (2019). More potent alternatives, such as rank transformations, were discussed by Johnson (1949) and applied practically by Karssen et al. (2016). Thus, we began with probing for power transformation and, if necessary, continued with rank transformation.

Table 4
Principal Component Analysis on accuracy measures, featuring component loadings (PC1), communality (h2), and uniqueness (u2).

Accuracy variable	PC1	h2	u2
Grammatical items	0.92	0.85	0.15
Ungrammatical items	0.90	0.81	0.19
Article items	0.95	0.90	0.10
Tense-aspect items	0.95	0.90	0.10

The analysis revealed that several variables deviated notably from the desired structure, indicating non-normality. These included Proficiency.PC1, Proficiency.PC2, Accuracy.PC1, CTScore, and ImmersionYears. For CTScore, the Box-Cox test (1964) suggested an inverse transformation ($\lambda \approx -1.0$). We applied scaling to the negative inverse (scale(−1/CTScore)), ensuring that (a) the directionality of the transformed variable aligned with the original CTScore, and (b) the scaling expanded the range of values. For Proficiency.PC1, Proficiency.PC2, Accuracy.PC1, and ImmersionYears, the power transformation was ineffective, but a rank transformation proved suitable. Density distribution comparison plots, pre- and post-transformation, are included in Appendix B, Fig. 3. Visual inspection shows that normality was achieved in all cases, ensuring the desired characteristics of the data and reducing the risk of artifacts.

Recall that the full set of test items was randomly divided into four sub-sets. Each set contained an equal number of grammatical and ungrammatical items, as well as tense-aspect and article items. Participants were randomly assigned to one of these sets. To ensure the absence of set-effects, we conducted ANOVAs to test for any significant differences among the four experimental sets across all critical variables. These tests uniformly indicated no significant differences ($p > 0.1$), validating the random assignment method.

Main analyses

Following the initial data preparation and validation steps, we continued to the modelling phase of our analysis. As our study is exploratory in nature, we used Generalized Additive Modelling (GAM; cf. Wood, 2017) which allows for nonlinear relations, including nonlinear interactions, to achieve maximally faithful insights about the nature of the chosen predictors. This approach was applied in two separate analyses, each focusing on a different response variable: self-evaluated proficiency (Proficiency.PC1) and test-evaluated proficiency (Proficiency.PC2). The GAM framework allowed us to effectively capture and interpret complex relationships within our dataset. We also utilized Random Forest (RF; Breiman, 2001; Genuer & Poggi, 2020) to corroborate the relative importance of the variables under consideration, as this statistical technique is particularly resilient to multicollinearity issues.

In the GAM analysis of self-evaluated proficiency (Proficiency.PC1), MeanFluency and Accuracy.PC1 emerged as significant predictors. Notably, these two key predictors also demonstrated a significant nonlinear interaction within the model. To validate the model's reliability, we engaged in model criticism as proposed by Baayen and Milin (2010)). This process involved excluding residuals outside the 95 % confidence interval and refitting the model. The exclusion of the one outlying data point in this process affirmed the robustness of the MeanFluency and Accuracy.PC1 interaction. As summarized in Tables 5 and 6, the full model explained 52.9 % deviance, showing $F = 13.48$ ($p < 0.00001$), and the critically assessed (trimmed) model explained 55 % deviance, $F = 14.28$ ($p < 0.00001$).

To determine the importance of each predictor, we compared AIC values for models with and without that predictor. Firstly, the AIC value for the model with the said predictor must be lower; secondly, the extent to which the model with the smaller AIC provides a more precise approximation of the data is summarized with Evidence Ratio (cf. Milin et al., 2017). In the case of our final GAM model, the Evidence Ratio suggests that it is 224.19 times better with MeanFluency and 8.14 times better with Accuracy.PC1, which suggests that MeanFluency is a stronger predictor of Proficiency.PC1. The results were corroborated by the Random Forest analysis, albeit pertaining to different statistical rationales and being on different scales: the same two variables were identified as the most influential predictors, with MeanFluency showing the greatest predictive importance. More details of the RF analysis are provided in Appendix C, Table 9.

To bring out the nonlinear interaction between MeanFluency and Accuracy.PC1 better, we make use of a visual representation in Fig. 2.

Fig. 2 illustrates the interaction pattern between MeanFluency and Accuracy.PC1. It is characterized by a plane of smoothly bent isolines and low values for the associated degrees of freedom of the smooth term ($edf = 3.0$, $Ref.df = 3.0$), suggesting a rather regular nonlinear (tensor) effect of the two predictors on Proficiency.PC1. The isolines, as indicators of self-evaluated proficiency, increase from, roughly, −2 (darker green, lower-right quadrant) to +1 (bright orange, upper-left quadrant), with 0.5 standardized increments.

Table 5
Outputs of the full Generalized Additive Model of the association between self-evaluated proficiency (Proficiency.PC1) and Predictors: R-sq. (adj) = 0.490; Deviance explained = 52.9 %; AIC = 92.078.

A. Parametric coefficients	Estimate	Std. error	t-value	p-value
Intercept	−0.0003	0.113	−0.003	0.998
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
Accuracy.PC1* Mean Fluency	3.000	3.000	13.48	< 0.0001

Table 6
Outputs of the critically assessed (trimmed) Generalized Additive Model of the association between self-evaluated proficiency (Proficiency.PC1) and Predictors: R-sq. (adj) = 0.512; Deviance explained = 55 %; AIC = 83.639.

A. Parametric coefficients	Estimate	Std. error	t-value	p-value
Intercept	−0.058	0.105	−0.549	0.586
B. Smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
Accuracy.PC1* Mean Fluency	3.000	3.000	14.28	< 0.0001

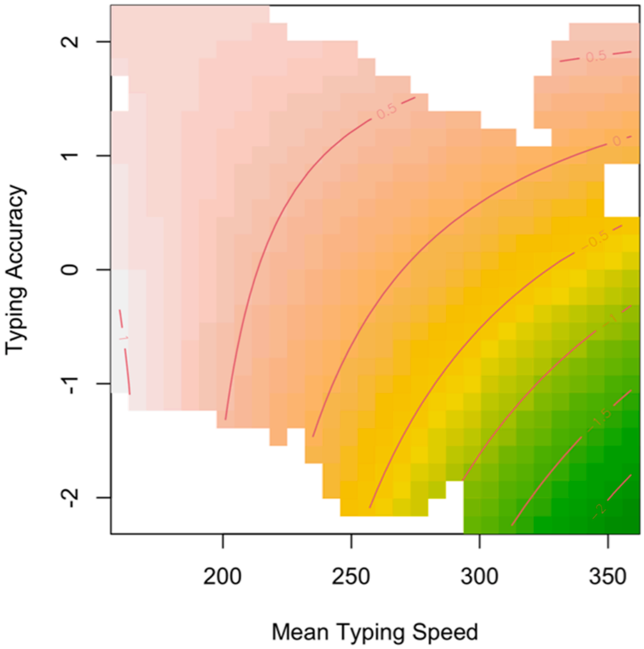

Fig. 2. Tensor products of MeanFluency by Accuracy.PC1 predicting Proficiency.PC1.

Assuming a roughly normal distribution of proficiency scores, the isoline at −2 suggests a low proficiency better than only 2.3 % of the whole sample, while the isoline at +1 suggests a high proficiency better than 84 % of the whole sample.

Notably, the green-colored area reveals a significant trend: participants with lower typing accuracy (lower Accuracy.PC1 scores) and slower mean typing speeds (higher MeanFluency value, indexing longer IKI) tended to assign themselves lower proficiency scores (Proficiency.PC1). Conversely, in the bright-orange section, participants showing high accuracy combined with faster typing speeds have the highest self-evaluated English language proficiency. In the section between 0 and −1 of accuracy, the variation in color from left to right indicates a great variability in proficiency, even though participants had similar accuracy scores. It is interesting to note

Table 7
Outputs of three progressively simpler Generalized Additive Models of Proficiency.PC2.

1. All predictors and controls as smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
Years of Immersion	1.000	1.000	0.203	0.656
pSTM	1.000	1.000	0.000	0.999
CT Score	3.518	4.349	2.499	0.065
Accuracy.PC1	1.000	1.000	2.745	0.111
Mean Fluency	1.000	1.000	0.001	0.975
CV Fluency	7.859	8.597	3.591	0.008
2. CV Fluency and CT Score as smooth terms	edf	Ref.df	F-value	p-value
CT Score	2.970	3.725	2.006	0.128
CV Fluency	7.209	8.151	2.234	0.055
CV Fluency* CT Score	1.000	1.000	0.198	0.660
3. CV Fluency as smooth terms	**edf**	**Ref.df**	**F-value**	**p-value**
CV Fluency	3.041	3.825	1.274	0.274

that our sample does not include participants with both low accuracy and high fluency (but a low MeanFluency value); this absence is represented by a white patch in the lower-left corner of Fig. 2. This gap in the data suggests the unlikely development of high speed and low accuracy among participants in this study.

As for test-evaluated proficiency (Proficiency.PC2), the GAM analysis indicates no significant effects from predictors and control variables. Initially, the GAM analysis with smooth terms for all variables revealed a significant and highly nonlinear effect of CV in fluency (CVFluency) as shown in the first model in Table 7: $edf = 7.86$, $Ref.df = 8.60$, $F = 3.59$, $p = 0.008$. In the same model, the Copy Task Score (CTScore) exhibited a marginally significant effect: $edf = 3.52$, $Ref.df = 4.35$, $F = 2.50$, $p = 0.065$.

In a subsequent analysis, a more parsimonious model retaining only CVFluency and CTScore (i.e., the second model in Table 7), along with their interaction, demonstrated a borderline significant effect for CVFluency ($p = 0.055$), while the effect of CTScore and the interaction of CVFluency by CTScore did not reach statistical significance. Further simplification, with a model incorporating only the smooth term for CVFluency (i.e., the third model in Table 7), revealed that this variable alone does not exert a significant effect ($p = 0.27$). The results of GAM models are confirmed by the Random Forest analysis which demonstrated generally low importance for all examined variables. More details of the RF results are summarized in Appendix C.

Synthesizing insights from these progressively simpler models, it becomes apparent that the statistical significance observed in the more complex models is not genuine and might be due to suppression effects.

Discussion

Although the EI task is used to assess automatized L2 knowledge, the scoring has focused on the accuracy of EI production exclusively, neglecting the process-based indices of automatization. Our study took the first step towards exploring the role of fluency in the EI task and investigated whether and to what degree the sensitivity of the EI task could be improved by adding fluency measures to the conventional accuracy measures. To this end, we designed a written version of the EI task and embedded it in a keystroke logging paradigm, in line with previous studies (Yan et al., 2016), task sensitivity was examined via correlations between the EI measures and L2 proficiency evaluated by other means. To answer the main research question about the degree to which accuracy and fluency measures predict proficiency levels, a written EI task was completed by 40 Polish speakers of English at varying proficiency levels (CEFR B1–C2). Keystroke data was elicited and analyzed for fluency (speed and pausing, coefficient of variation) and accuracy in the written EI production of English grammatical structures, specifically in terms of article and tense-aspect. L2 proficiency was evaluated via self-rating and the vocabulary and grammar tests from DIALANG 2.0.

The results of statistical modelling support our hypothesis showing that both fluency and accuracy of EI production jointly predict proficiency to provide a more fine-grained measure of language ability. These findings provide first evidence that the sensitivity of the EI task can be improved by adding fluency measures. Compared with the overall effect size of EI tasks using binary scoring ($r = 0.53$; equal to $R\text{-}sq. = 0.28$) reported in Kostromitina and Plonsky (2022), adding fluency measure (speed and pausing) increases the effect size of the EI task in this study ($R\text{-}sq. = 0.490$ or 0.512) to the overall level reported in studies using ordinal scales ($r = 0.70$; equal to $R\text{-}sq. = 0.49$). Specifically, participants who rated themselves as more proficient in English speaking (i.e., self-evaluated proficiency) demonstrated a higher speed (i.e., shorter inter-key intervals) and/or higher accuracy (i.e., higher proportion of correct responses) when reproducing the English article and tense-aspect target structures in the EI task. The fluency measure contributed more than accuracy in explaining the variance in proficiency levels, which further supports the importance of considering fluency in EI tasks.

Starting with accuracy, the results suggest that the effectiveness of the accuracy measures of oral EI tasks (Yan et al., 2016) in differentiating proficiency levels can be extended to written EI tasks. Given that our written EI task adopted all the key task features that determine the capacity to elicit automatized knowledge (e.g., length of stimuli, delay in response, time pressure, meaning-oriented instruction) and that only the first attempt at response was evaluated to limit self-monitoring, it is reasonable to assume that the task measured automatized grammatical knowledge. This is partially supported by the evidence that the accuracy measure was not correlated with test-evaluated proficiency measured via untimed tests of explicit L2 knowledge (cf. Ellis, 2005), and that the association between accuracy and self-evaluated proficiency was not moderated by phonological short-term memory (cf. Granena, 2016). As for the self-evaluated proficiency, although it was assessed by responding to can-do statements instead of real-time behavioral tests, the specific real-life condition in descriptors does concern automatized L2 use. For example, the descriptor for B2 level of Speaking (production) is "I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options" (Council of Europe, 2001, p. 27). This is likely why the accuracy of the written EI task was associated with self-evaluated proficiency.

With individual differences in typing skills controlled for, fluency (as measured by inter-key intervals) largely reflected pauses at lower-level text units (between and within words). Recall that EI measures were generated for the target structures only, where participants needed to reconstruct the article and tense-aspect grammatical structures and the content words in the sentence stimuli they heard. Thus, our finding suggests that higher proficiency, and presumably more automatized L2 knowledge, was manifested by faster lower-order writing processes such as morphosyntactic encoding and lexical retrieval (cf. Révész et al., 2019). This is in line with previous research on associations between proficiency and speed and pausing fluency in less controlled spontaneous writing (e.g., argumentative writing; Révész et al., 2022; Xu & Xia, 2021). The result thus confirms the effectiveness of using speed and pausing measures for fluency in written EI tasks. The consistency in speed fluency (i.e., coefficient of variation in inter-key intervals), however, was not correlated with proficiency, which is in line with the results in Hulstijn et al. (2009) and Suzuki and Sunada (2018).

Crucially, a non-linear interaction was found between accuracy and fluency predictors. Our results suggest a general trend in the development of EI performance across proficiency levels. When participants develop from low to average proficiency level, they perform better in both accuracy and fluency. After reaching an average proficiency level, their accuracy in EI production seems to plateau, but their fluency continues to improve. It is not surprising that participants in this study demonstrated a considerable degree of automatization, given their extensive immersion experience (8.5 years of residence in the UK on average). The result highlights a greater capacity of fluency in differentiating proficiency levels, compared with the conventional accuracy. It is especially significant to include fluency measures to maintain the sensitivity of EI tasks among advanced learners. Future studies could take a longitudinal approach to investigate individuals' developmental trajectories and to reveal how L2 knowledge is automatized through immersion.

Taking a closer look at participants at the same proficiency level, especially those around the average level, we observe substantial variability in their accuracy and fluency in EI production. There seems to be wiggle room in what participants prioritize in the task given the available cognitive resources: participants could be more accurate in responses without speeding up, or they could respond faster while retaining the same accuracy level, or they could be somewhere in between these two options. Therefore, both fluency and accuracy need to be considered in scoring the EI task to gauge proficiency more comprehensively.

Future studies could investigate how individuals prioritize these two dimensions (accuracy vs. fluency) considering the particularities of a downstream language task. For example, the learning experience or learning context could play a role and L2 users with more immersion experience may prioritize fluency which is crucial in their daily communication, whereas L2 learners with more training in foreign classrooms may emphasize accuracy. Furthermore, while our study focused on L2 users in an immersive context with mixed lengths of residence and formal instruction, future studies could compare L2 learner groups from different contexts (e.g., second language vs. foreign language), ideally with a larger sample size. Additionally, individual traits such as personality traits and/or emotional state (e.g., anxiety) could also affect both fluency and accuracy: previous research has found that individuals who are more extraverted or open to experience tend to be more fluent in L2 oral production (Dewaele & Furnham, 2000; Gaffney, 2021) and that those who are more anxious tend to make more errors (Oya et al., 2004). Finally, the design of the EI task should also be considered, including the time limit, the emphasis on grammatical correctness or general accuracy in the task instructions, the difficulty of the target structures or the complexity of sentence stimuli (cf. trade-off effects in Robinson, 2001; Skehan, 1998).

It is noteworthy that the EI task is a useful instrument to elicit rich productive language data for the examination of different dimensions of language acquisition (e.g., oral vs. written, morphosyntactic vs. lexical vs. phonological), all while maintaining a high degree of experimental control. Compared with other spontaneous L2 production tasks (e.g., narrative oral task, argumentative essay), the EI task allows us to test specific linguistic structures by manipulating the sentence stimuli. The current study demonstrates how keystroke data can be used with EI tasks to gain precise insights into the processes of the written production of English article and tense-aspect grammatical structures, which opens the door for a variety of further investigations. Future studies could focus on different linguistic features and generate more refined fluency measures (e.g., speed, pausing, revision) using the latest InputLog 9 or other keystroke logging software to look into more detailed cognitive processes underlying L2 written production. Additionally, emerging research on automated speech assessment tools has shown potential for efficient ways to measure utterance fluency (Al-Ghezi et al., 2023; de Jong et al., 2021; Isbell et al., 2023).

Although we focused on English articles and tense-aspect structures – known to differ dramatically between our participants' L1 and L2 – instead of including a wide variety of grammatical structures, the distribution of correct response proportions closely mirrors those reported in Erlam's (2006) study targeting 17 different structures. Specifically, the rates for repeating grammatical stimuli (0.62 vs. 0.61) and correcting ungrammatical stimuli (0.39 vs. 0.35) are remarkably similar. It suggests that our EI task could reflect participants' proficiency level to a similar degree as EI tasks measuring global L2 grammatical knowledge. Having said that, we should be tentative in generalizing the results to EI tasks measuring global oral proficiency.

Another limitation of this study is that no other timed or spontaneous proficiency tests (e.g., timed grammatical judgement test, narrative oral task; Ellis, 2005; Spada et al., 2015) were included apart from the EI task. This makes it hard to compare our study with previous studies that assess the correlations between EI measures and other timed or spontaneous proficiency measures. Similarly, the design of the interruptive task embedded in our EI task differs from previous studies, i.e., participants were asked to rate how easy it was to understand the sentence they had heard before typing it to prevent rote memorization of the stimulus as the interruption lasted for at least three seconds. While we did not observe an effect of phonological short-term memory, it remains unclear to what degree participants paid attention to the rating. Future studies could use interruptive tasks where the correctness of the responses can be judged (e.g., True/False questions, math questions).

Conclusion

This study introduces a novel implementation of the EI task, accompanied by an innovative scoring method designed to capture significant interindividual variability that is often obscured when relying on standard, but coarse, proficiency levels. Using a newly developed written version of the EI task, which incorporates the keystroke logging technique, we examined whether incorporating fluency measures alongside conventional accuracy measures could enhance the task's capacity to capture proficiency on a finer-grained scale. The fluency measure (i.e., speed and pausing) emerged as a stronger predictor of self-assessed proficiency than the conventional accuracy measure, particularly among advanced-level L2 users. Furthermore, our findings indicate that individuals prioritize accuracy and fluency differently in EI production, underscoring the value of evaluating both dimensions of EI performance. The current study illustrates how keystroke data can be collected within a written EI task and analyzed to assess EI fluency. The findings provide new insights into the effects of individual differences and task-specific factors, offering a solid foundation for further research into the measurement of L2 language knowledge and performance.

Funding sources

This work was supported by The Leverhulme Trust [grant number RL-2016–001] awarded to Dagmar Divjak, which funded all authors as well as the research assistant who collected the data. The study was designed and set up while Hui Sun was a member of staff at the University of Birmingham.

CRediT authorship contribution statement

Hui Sun: Conceptualization, Methodology, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Dagmar Divjak:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Petar Milin:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing.

Declaration of competing interest

All authors declare no conflict of interest.

Acknowledgements

We would like to thank Dr Laurence Romain for creating the text stimuli and assisting in calibrating the annotation scheme; Oghenetekevwe Kwakpovwe for recording the audio stimuli; Yuhan Luo and Rupali Limachya for collecting the data; and Rupali Limachya for first-coding the keystroke logging data for accuracy. We extend our gratitude to Dr Petar Čolović for general discussion of psychometric measurements, which influenced the analytic approach adopted in this study. This study forms part of a larger study looking at ultimate attainment in different populations of learners. We are thankful to the members of the Out of Our Minds team at Birmingham for their feedback over the lifetime of this project.

Appendix A. : Summary of McDonald's omega analysis on proficiency measures

Table 8
McDonald's omega analysis on measures of English language proficiency, featuring general component loadings (g) and loadings on the first two principal components (F1*, F2*), communality (h2), uniqueness (u2), and general factor variance (p2). Component loadings below an absolute value of 0.2 are not presented.

Proficiency variable	g	F1*	F2*	h2	u2	p2
Listening	0.36	0.68		0.59	0.41	0.22
Writing	0.39	0.79		0.77	0.23	0.19
Speaking	0.33	0.67		0.56	0.44	0.20
Reading	0.46	0.74		0.77	0.23	0.28
Grammar	0.31		−0.58	0.43	0.57	0.22
Vocabulary (scaled)	0.38		−0.70	0.64	0.36	0.23
Sum of squares	0.85	2.07	0.86			

Appendix B. Distributions of variables pre- and post-transformations

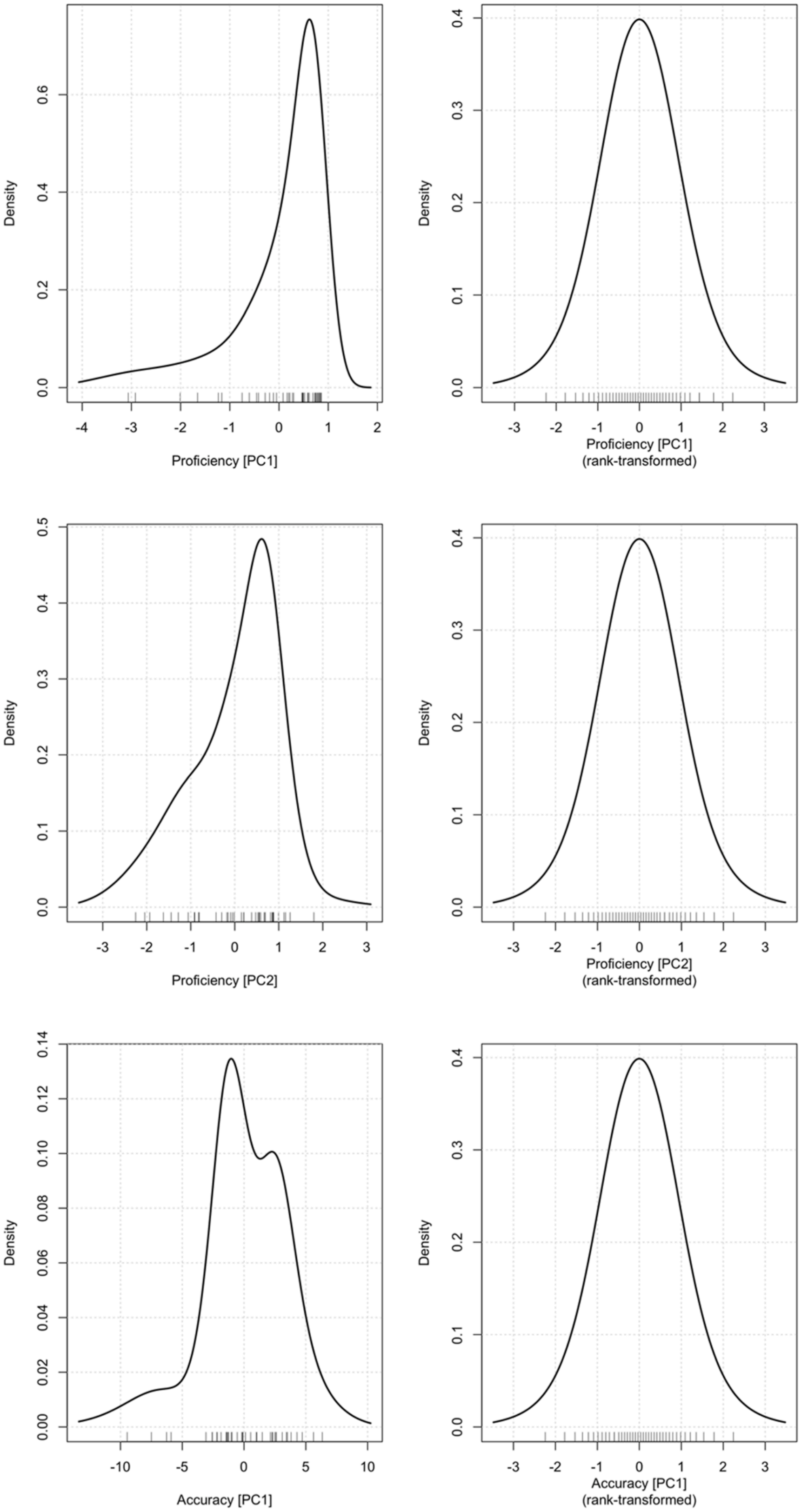

Fig. 3. Density distribution comparison plots of variables before (left panel) and after (right panel) the transformations.

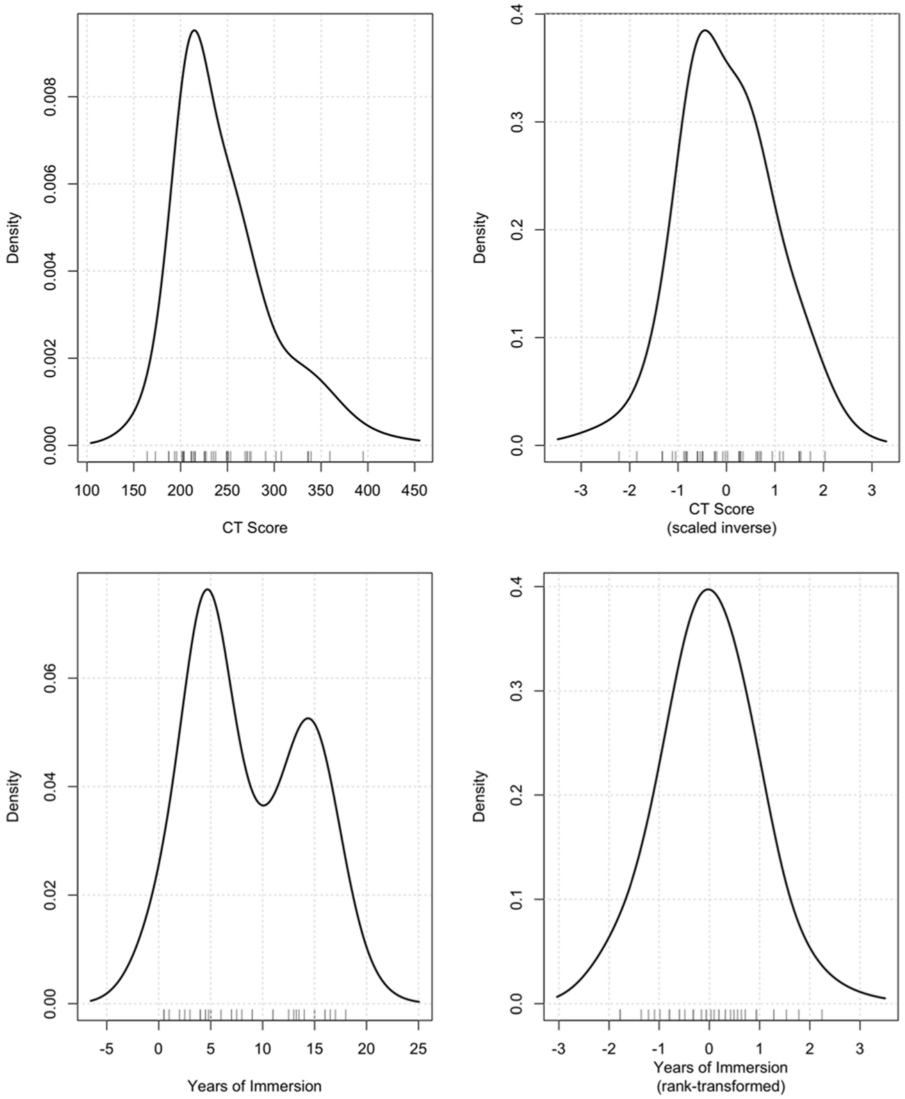

Fig. 3. (*continued*).

Appendix C. Summary of Random Forest analysis on proficiency measures

In line with results of the GAM approach, the RF analysis identified MeanFluency and Accuracy.PC1 as the most influential predictors for self-evaluated proficiency (Proficiency.PC1). The influence of other predictors and control variables was considerably smaller. These findings are detailed in Table 9, which presents the increase in Mean Square Error (MSE) resulting from the permutation of each variable (i.e., the random shuffling of variable values). Note that a higher (or steeper) increase in MSE due to permutation indicates a more significant impact of random alterations on the model's accuracy, thus highlighting that MeanFluency showed greater predictive importance than Accuracy.PC1.

Table 9
Importance of predictors and controls in contributing to the model of self-evaluated proficiency (Proficiency.PC1).

Variable	Increase in MSE
Mean Fluency	20.632
Accuracy.PC1	12.082
pSTM	3.643

(*continued on next page*)

Table 9 (*continued*)

Variable	Increase in MSE
Years of Immersion	3.046
CV Fluency	0.317
CT Score	−2.317

The RF analysis for the test-evaluated proficiency scores (Proficiency.PC2) indicated generally low importance for all examined variables, which corroborated the GAM models where no significant predictor was found. Among them, the CV in fluency (CVFluency) emerged as the most significant predictor, with a modest impact, confirmed by an increase of <7 units in MSE after permutations. These findings are summarized in Table 10, which details the relative influence of each variable on Proficiency.PC2 scores.

Table 10
Importance of predictors and controls in contributing to the model of test-evaluated proficiency (Proficiency.PC2).

Variable	Increase in MSE
CV Fluency	6.837
pSTM	2.605
CT Score	2.248
Accuracy.PC1	−0.712
Mean Fluency	−2.641
Years of Immersion	−4.736

Data availability

The data and materials that support the findings of this study are openly available on the University of Birmingham Institutional Research Archive, UBIRA at https://doi.org/10.25500/edata.bham.00001200. The R-scripts for all analyses presented here are available on the Lab's GitHub repository at https://github.com/ooominds/FluencyMeasuresInElicitedImitationTask.

References

Agresti, A. (2013). *Categorical data analysis* (3rdEdition). John Wiley & Sons.

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

Al-Ghezi, R., Voskoboinik, K., Getman, Y., Von Zansen, A., Kallio, H., Kurimo, M., et al. (2023). Automatic speaking assessment of spontaneous L2 Finnish and Swedish. *Language Assessment Quarterly, 20*(4–5), 421–444. https://doi.org/10.1080/15434303.2023.2292265

Baaijen, V. M., Galbraith, D., & de Glopper, K. (2012). Keystroke analysis: Reflections on procedures and measures. *Written Communication, 29*(3), 246–277. https://doi.org/10.1177/0741088312451108

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research, 3*(2), 12–28. https://doi.org/10.21500/20112084.807

Bardovi-Harlig, K. (2000). *Tense and aspect in second language acquisition: Form, meaning, and use*. Oxford: Blackwell.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Breiman, L. (2001). Random forests. *Machine learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*(3), 249–253. https://doi.org/10.1177/014662168300700301

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages: A learner corpus study. *International Journal of Learner Corpus Research, 2*(1), 68–100. https://doi.org/10.1075/ijlcr.2.1.03cro

de Jong, N. H., Pacilly, J., & Heeren, W. (2021). PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education Principles Policy and Practice, 28*(4), 456–476. https://doi.org/10.1080/0969594x.2021.1951162

Dewaele, J.-M., & Furnham, A. (2000). Personality and speech production: A pilot study of second language learners. *Personality and Individual Differences, 28*(2), 355–365. https://doi.org/10.1016/s0191-8869(99)00106-3

Divjak, D., Romain, L., & Milin, P. (2023). From their point of view: The article category as a hierarchically structured referent tracking system. *Linguistics, 61*(4), 1027–1068. https://doi.org/10.1515/ling-2022-0186

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141–172. https://doi.org/10.1017/s0272263105050096

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464–491. https://doi.org/10.1093/applin/aml001

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression*. Sage publications.

Gaffney, C. (2021). *The Effects of Intelligence, Personality Traits, L1 Fluency, and L2 proficiency on L2 Spoken Fluency*. Available from ProQuest Dissertations & Theses Global; ProQuest One Literature; Social Science Premium Collection. (2609739064). https://www.proquest.com/dissertations-theses/effects-intelligence-personality-traits-l1/docview/2609739064/se-2.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task: Elicited imitation. *Language Learning, 66*(2), 419–447. https://doi.org/10.1111/lang.12157

García-amaya, L., & Cintrón-valentín, M. C. (2021). The effects of textually enhanced captions on written elicited imitation in L2 grammar. *Modern Language Journal, 105*(4), 919–935. https://doi.org/10.1111/modl.12740

Gathercole, S. E., Pickering, S. J., Hall, M., & Peaker, S. M. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *The Quarterly Journal of Experimental Psychology Section A, 54*(1), 1–30. https://doi.org/10.1080/02724980042000002

Genuer, R., & Poggi, J.-M. (2020). *Random forests with R*. Springer.

Granena, G. (2016). Elicited imitation as a measure of implicit L2 knowledge. In G. Granena, D. O. Jackson, & Y. Yilmaz (Eds.), *Cognitive individual differences in second language processing and acquisition* (pp. 185–204). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.3.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than cronbach's alpha for estimating reliability. *But…. Communication Methods and Measures, 14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hulstijn, J. H., Van Gelderen, A., & Schoonen, R. (2009). Automatization in second language acquisition: What does the coefficient of variation tell us? *Applied Psycholinguistics, 30*(4), 555–582. https://doi.org/10.1017/s0142716409990014

Isbell, D. R., Kim, K. M., & Chen, X. (2023). Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test. *Research Methods in Applied Linguistics, 2*(3), Article 100076. https://doi.org/10.1016/j.rmal.2023.100076

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika, 36*(1/2), 149–176. https://doi.org/10.2307/2332539

Karssen, L. C., van Duijn, C. M., & Aulchenko, Y. S. (2016). The GenABEL Project for statistical genomics. *F1000Research, 5*, 914. https://doi.org/10.12688/f1000research.8733.1

Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an elicited imitation test for SLA research. *Modern Language Journal, 100*(3), 655–673. https://doi.org/10.1111/modl.12346

Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of l2 proficiency: A meta-analysis. *Studies in Second Language Acquisition, 44*(3), 886–911. https://doi.org/10.1017/s0272263121000395

Leech, G. (1992). 100 million words of English: The British National Corpus (BNC). *Language Research, 28*, 1–13.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392. https://doi.org/10.1177/0741088313491692

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news, 2*(3), 18–22.

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

McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *The Journal of Speech and Hearing Disorders, 47*(1), 19–24. https://doi.org/10.1044/jshd.4701.19

Mcmanus, K., & Marsden, E. (2019). Signatures of automaticity during practice: Explicit instruction about L1 processing routines can improve L2 grammatical processing. *Applied Psycholinguistics, 40*(1), 205–234. https://doi.org/10.1017/s0142716418000553

Milin, P., Feldman, L. B., Ramscar, M., Hendrix, P., & Baayen, R. H. (2017). Discrimination in lexical decision. *PloS One, 12*(2), Article e0171935. https://doi.org/10.1371/journal.pone.0171935

Miller, K. S., Lindgren, E., & Sullivan, K. P. H. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly, 42*(3), 433–454. https://doi.org/10.1002/j.1545-7249.2008.tb00140.x

Oya, T., Manalo, E., & Greenwood, J. (2004). The influence of personality and anxiety on the oral performance of Japanese speakers of English. *Applied Cognitive Psychology, 18*(7), 841–855. https://doi.org/10.1002/acp.1063

Park, H. I. N., Solon, M., Henderson, C., & Dehghan-chaleshtori, M. (2020). The roles of working memory and oral language abilities in elicited imitation performance. *Modern Language Journal, 104*(1), 133–151. https://doi.org/10.1111/modl.12618

Perkins, K., Brutten, S. R., & Angelis, P. J. (1986). Derivational complexity and item difficulty in a sentence repetition task. *Language Learning, 36*(2), 125–141. https://doi.org/10.1111/j.1467-1770.1986.tb00375.x

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research*. In (Version R package version 2.3.6) Northwestern University. https://CRAN.R-project.org/package=psych.

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviours: A mixed-methods study. *Studies in Second Language Acquisition, 41*(3), 605–631. https://doi.org/10.1017/S027226311900024X

Révész, Andrea, Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing, 58*(100927), Article 100927. https://doi.org/10.1016/j.jslw.2022.100927

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). Cambridge University Press. https://doi.org/10.1017/cbo9781139524780.012.

Segalowitz, N. S., & Segalowitz, S. J. (1993). Skilled performance, practice, and the differentiation of speed-up from automatization effects: Evidence from second language word recognition. *Applied Psycholinguistics, 14*(3), 369–385. https://doi.org/10.1017/s0142716400010845

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies: Validating an elicited imitation task. *Language Learning, 65*(3), 723–751. https://doi.org/10.1111/lang.12129

Spanos, A. (2019). *Probability theory and statistical inference: Econometric modeling with observational data*. Cambridge University Press. https://doi.org/10.1017/CBO9780511754081

Suzuki, S., & Révész, A. (2023). Measuring speaking and writing fluency: A methodological synthesis focusing on automaticity. In Y. Suzuki (Ed.), *Practice and automatization in second language research* (pp. 235–264). Routledge. https://doi.org/10.4324/9781003414643.

Suzuki, Yuichi, & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge: Elicited imitation and word monitoring. *Language Learning, 65*(4), 860–895. https://doi.org/10.1111/lang.12138

Suzuki, Yuichi, Jeong, H., Cui, H., Okamoto, K., Kawashima, R., & Sugiura, M. (2023). fMRI reveals the dynamic interface between explicit and implicit knowledge recruited during elicited imitation task. *Research Methods in Applied Linguistics, 2*(2), Article 100051. https://doi.org/10.1016/j.rmal.2023.100051

Suzuki, Yuichi, & 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

Tracy-Ventura, N., Mcmanus, K., Norris, J. M., & Ortega, L. (2014). Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Multilingual Matters. https://doi.org/10.21832/9781783092291-011.

Van Rij, J., Wieling, M., Baayen, R.H., & van Rijn, D. (2022). *itsadug: Interpreting time series and autocorrelated data using GAMMs*. In (Version R package version 2.4.1) https://CRAN.R-project.org/package=itsadug.

Van Waes, L., Leijten, M., Pauwaert, T., & Van Horenbeeck, E. (2019). A multilingual copy task: Measuring typing and motor skills in writing with inputlog. *Journal of Open Research Software, 7*(1), 30. https://doi.org/10.5334/jors.234

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models: Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society. Series B, Statistical Methodology, 73*(1), 3–36. https://doi.org/10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd Edition). Chapman and Hall/CRC. https://doi.org/10.1201/9781315370279

Xu, C., & Xia, J. (2021). Scaffolding process knowledge in L2 writing development: Insights from computer keystroke log and process graph. *Computer Assisted Language Learning, 34*(4), 583–608. https://doi.org/10.1080/09588221.2019.1632901

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