Journal of Second Language Writing 58 (2022) 100927

Contents lists available at ScienceDirect

Journal of Second Language Writing

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

The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing☆

Andrea Révész [a,*], Marije Michel [b], Xiaojun Lu [c], Nektaria Kourtali [d], Minjin Lee [e], Laís Borges [f]

[a] University College London, UK
[b] University of Groningen, the Netherlands
[c] University of Nottingham Ningbo, China
[d] University of Liverpool, UK
[e] Yonsei University, South Korea
[f] Universidade Católica de Brasília, Brazil

ARTICLE INFO

Keywords:
Fluency
Pausing
Keystroke logging
Eye-tracking
Stimulated recall

ABSTRACT

In this study we investigated the extent to which writing proficiency predicts L2 writers' speed fluency, pausing, and eye-gaze behaviours and the cognitive processes that underlie pausing. Additionally, we explored whether these relationships were influenced by stage of writing (beginning, middle stages, end). The participants were 60 Chinese second language users of English, with proficiency levels ranging from CEFR B1 to C1 levels. They all completed two independent TOEFL iBT writing tasks over two sessions, with the tasks being counterbalanced across participants. While composing, we recorded participants' keystrokes and eye-gaze behaviours. Participants also took part in a stimulated recall session based on the last writing task they had completed. A series of mixed-effects regression models found that proficiency was the strongest predictor of speed fluency. The stimulated recall analysis revealed considerably less variation in cognitive activities among lower- than higher-proficiency writers across writing stages.

1. Introduction

The past two decades have seen an increased interest in investigating the processes in which second language (L2) writers engage, understood here as the behaviours of L2 writers (i.e., directly observable features such as fluency and pausing) and the cognitive operations which are associated with L2 text generation (e.g., planning, linguistic encoding; for reviews, see Cumming, 2016; Polio, 2012; Roca de Larios et al., 2016). Part of this research has been concerned with exploring how L2 proficiency might be linked to writing behaviours and underlying cognitive processes. Researchers have generated considerable insights into the relationship of proficiency to speed fluency (i.e., the rate at which writing is produced; e.g., Spelmann Miller et al., 2008) and cognitive processes (e.g., Roca de Larios et al., 2008). Little is known, however, about how proficiency relates to pausing and viewing behaviours and

☆ This research was funded by Educational Testing Service (ETS) under a Committee of Examiners and the Test of English as a Foreign Language research grant. ETS does not discount or endorse the methodology, results, implications, or opinions presented by the researcher(s).

* Corresponding author.
E-mail address: a.revesz@ucl.ac.uk (A. Révész).

https://doi.org/10.1016/j.jslw.2022.100927
Received 3 May 2021; Received in revised form 5 September 2022; Accepted 6 September 2022
Available online 27 September 2022

associated cognitive processes (cf., Barkaoui, 2019; Gánem-Gutiérrez & Gilmore, 2018). It also remains unexplored how stage of writing (beginning, middle, end) may moderate these links.

In this study, our aim was to obtain a fuller picture of the role of proficiency in the L2 writing process. To this end, we intended to contribute to previous research by further examining the relationship of proficiency to L2 writers' speed fluency and pausing behaviours, and to extend existing studies by exploring how proficiency links to eye-gaze behaviours and pause-related cognitive processes during writing. Additionally, our goals were to investigate whether these relationships are influenced by stage of writing and whether the behaviours and cognitive activities occur in the early, mid, or later stages of the writing process. Studying how proficiency relates to writing processes is of both theoretical and practical significance. Theoretical models of writing, by nature, need to account for the role of proficiency when applied to L2 writing. Therefore, to inform and test models of writing in L2 contexts, it is necessary to investigate how proficiency relates to writing behaviours and associated cognitive processes. On the practical front, information about the link between proficiency and writing processes may inform language teaching and assessment. Lower-proficiency writers, for example, might benefit from explicit instruction highlighting how more proficient students allocate their attention during various stages of writing, whether they write faster, pause more, and/or re-read their text more frequently in the early, mid, or later stages. Also, links between proficiency and writing process measures (e.g., Barkaoui, 2019; Révész, Michel et al., 2017) could provide the impetus for validation work in the field of assessment with a view to incorporating process indices to automatic scoring practices.

From a methodological perspective, a novel aspect of our research was the joint use of keystroke-logging, eye-tracking, and stimulated recall to investigate writing processes. To date, most cognitively-oriented empirical research into L2 writing processes has utilised verbal reports such as think-aloud and stimulated recall procedures. However, due to limitations associated with these techniques (Polio & Freedman, 2017; Révész et al., 2019) and increasing prevalence of computer-based writing (McKee & de Voss, 2007), L2 researchers have begun to triangulate verbal protocol data with keystroke-logging and/or eye-tracking recordings to tap digital writing behaviours and corresponding cognitive activities (Gánem-Gutiérrez & Gilmore, 2018; Michel et al., 2020; Révész, Kourtali et al., 2017; Révész, Michel et al., 2017, 2019; Stevenson et al., 2006). Nevertheless, multiple-method studies are still rare in L2 writing research. Combining these three methods enabled us to triangulate information about real-time text production behaviours (keystroke logging), viewing behaviours such as the re-reading of previously produced text (eye-tracking), and writers' thought processes during pausing (stimulated recall). As a result, we were able to establish a deeper understanding of the writing process than a single method would have made possible. Using keystroke logging alone would have prevented us from obtaining direct information about viewing behaviours during writing and about the conscious cognitive processes associated with writing behaviours. The stimulated recall protocol data helped resolve this potential limitation by providing insights into writers' conscious cognitive activities, whereas eye-tracking, through recording the writer's eye-gaze behaviours moment by moment, enabled us to capture viewing processes such as the re-reading of the writing prompt and the writer's evolving text.

2. Background

2.1. Theoretical background

For the past three decades, several cognitive models of writing (e.g., Flower & Hayes, 1980; Galbraith, 2009; Hayes, 2012; Kellogg, 1996) have been adopted as a theoretical basis for investigating L2 writing processes. Among these, we have chosen Kellogg's (1996) model as our theoretical framework. This model was originally proposed to describe first language writing processes. Compared to other accounts, however, it provides greater detail about linguistic encoding processes, making this model more suited for studying L2 writing. Linguistic encoding processes are less automatized for L2 than L1 writers. Thus, they tend to pose greater cognitive demands, requiring more conscious attention (Kormos, 2012; Roca de Larios et al., 1999).

There are three main processes posited by Kellogg (1996). The first is formulation, which entails higher-order activities such as planning content and organising ideas to achieve a coherent plan. Formulation also involves lower-order, translation processes, which convert the writer's ideas into linguistic form through lexical retrieval, syntactic encoding, and use of cohesive devices. The second phase is execution, where motor skills are deployed to generate a hand-written or typed text. The last step is monitoring, during which writers check that their evolving piece reflects their plan. These processes are assumed to work in cycles and in parallel.

There is growing recognition that writing is also dynamic in a more macro sense—that is, writers engage in different composing processes to a differential extent at various stages of writing (e.g., beginning, middle, end). The temporal distribution of cognitive activities is assumed to reflect how the writer's perception of the task changes during the composing process (Khuder & Harwood, 2015; Nicolás-Conesa et al., 2014; Rijlaarsdam & van den Bergh, 1996). That is, the cognitive processes in which writers engage are expected to vary according to the changing task environment.

2.2. Proficiency and writing processes

Based on Kellogg's (1996) model, we expected proficiency to have a strong relationship with L2 writers' speed fluency, pausing, and underlying cognitive processes. According to Hulstijn (2011), the construct of L2 proficiency is made up of linguistic cognition (i.e., phonetic-phonological, morphonological, morphosyntactic, and lexical knowledge and the speed at which these components can be applied by learners) and metacognitive competences. Given that lower proficiency is associated with less automatized linguistic skills (Lindgren et al., 2008; Schoonen et al., 2009) and less developed strategic orientation of problem-solving behaviours (López-Serrano et al., 2019; Manchón et al., 2009; Roca de Larios et al., 2008), less proficient L2 writers will struggle more with handling the competing demands imposed on planning, translation, and monitoring by working memory constraints. Therefore, assuming that

equal pressure is exerted on planning activities due to conceptualisation demands, lower proficiency writers will probably display decreased speed fluency and greater frequency and length of pausing. Pausing is likely to become especially frequent and lengthy at lower textual units (within and between words), as pauses at such locations have been shown to be associated more often with linguistic encoding (e.g., Révész, Kourtali et al., 2017; Révész et al., 2019).

Given that less previous research has investigated eye-gaze behaviours in the context of writing, our hypotheses are more tentative for the link between proficiency and eye-movements. Nevertheless, previous research has demonstrated that L2 writers are more likely to look back on shorter stretches of text when they pause at lower textual units (Chukharev-Hudilainen et al., 2019; Révész et al., 2019) or revise shorter units (Révész et al., 2019). Thus, we expected lower-proficiency writers to display more local viewing behaviour. Specifically, we anticipated that they would show shorter lookbacks during pauses and revisions, given that they will probably pause more at lower textual units and focus more on local revisions, as they struggle with linguistic encoding to a greater extent. In general, as compared to higher-proficiency writers, they will likely have fewer attentional resources to go further back in previously written text to facilitate monitoring and the generation of new content.

The few studies that have examined the link between proficiency and speed fluency have confirmed that, with growing proficiency, writers do indeed increase their writing speed. For example, Palviainen et al. (2012) examined the performance of Finnish university students of L2 English or Swedish on a narrative and an argumentative writing task. The students' proficiency ranged from B1 to C2 according to the Common European Framework of Reference (CEFR). As expected, the researchers found that higher-proficiency writers displayed higher speed fluency. Barkaoui (2019) yielded similar results comparing pre-admission (below CEFR B2) and post-admission (CEFR B2 or above) university students' performance on TOEFL writing tasks. Further evidence for the positive relationship between proficiency and speed fluency comes from two three-year longitudinal studies. Spelman Miller et al. (2008) investigated the writing development of Swedish high-school students in L2 English. The participants were 14 years old and had had 3.5 years of previous English instruction. Kowal (2014) assessed changes in the speed fluency of Polish university students studying L2 Swedish, who had no prior knowledge of Swedish. Both studies established greater fluency as a function of rising proficiency. Importantly, all these studies employed keystroke-logging to obtain process-based measures of fluency (Abdel Latif, 2012), such as words and/or characters per minute (Barkaoui, 2019; Palviainen et al., 2012; Spelman Miller et al., 2008), length of bursts (Kowal, 2014; Palviainen et al., 2012; Spelman Miller et al., 2008), and fluency in bursts (Palviainen et al., 2012; Spelman Miller et al., 2008).

The few studies that have explored the relationship of proficiency to pause frequency and length have yielded mixed findings. Xu and Ding (2014) compared Chinese writers of L2 English with writing scores of 8 or below to 12 or higher on the College English Test 4, a national English proficiency test. Contrary to the predictions above, the researchers found that L2 writing skill did not influence the total frequency and duration of pauses on an argumentative task. On the other hand, Barkaoui (2019), as hypothesised above, observed that writers at lower proficiency paused more often overall than higher-proficiency L2 users. Interestingly, any differences detected in pause frequency, duration, and proportion according to pause location (within words; between words, sentences, and paragraphs) did not vary by proficiency. A limitation of these studies was that they exclusively used keystroke-logging to investigate pausing, providing no information about writers' thought processes during pauses. As a result, it remains to be tested whether the cognitive activities associated with pauses at various locations differ depending on proficiency. Although pauses have been shown to be more often associated with planning and linguistic encoding at higher and lower textual units respectively (e.g., Révész, Kourtali et al., 2017; Révész et al., 2019), previous studies have focused only on high proficiency learners, leaving it to be examined whether these trends apply across proficiency levels.

While no research has looked into how proficiency might relate to cognitive processes during pausing, there is considerable evidence suggesting that, indeed, high-proficiency writers are better able to deal with the competing demands posed by writing subprocesses, as predicted by Kellogg's (1996) model (e.g., Cumming, 1989; Gánem-Gutiérrez & Gilmore, 2018; Plakans, 2009; Raimes, 1987; Roca de Larios et al., 2008; Sasaki, 2002; Tillema, 2012). For example, in one of the first studies investigating the link between proficiency and cognitive writing processes, Raimes (1987) found that ESL students in a non-remedial composition course planned, rehearsed, scanned, and revised their work more frequently than their counterparts in an ESL remedial composition programme. Similar, Sasaki (2002) observed that expert Japanese EFL writers engaged in more global planning, re-reading, and revising as compared to novices. Replicating these patterns, in Roca de Larios et al. (2008), more proficient writers demonstrated a more balanced division of time to various cognitive activities overall, devoting more attention to planning, evaluation, and revision and somewhat less time to formulation processes. In all these studies, verbal protocols, alone or together with other techniques, were employed to tap internal composing activities.

To date, only a few studies have investigated the viewing behaviours of L2 writers like re-reading previously produced text (Chukharev-Hudilainen et al., 2019; Gánem-Gutiérrez & Gilmore, 2018; Révész, Michel et al., 2017, 2019). Among them, a single study, Gánem-Gutiérrez and Gilmore (2018), has addressed how proficiency may relate to viewing during writing. The study included Japanese L2 writers of English whose proficiency ranged from elementary to advanced. The researchers employed digital screen capture data as the main elicitation tool, whereas eye tracking, video recording, and stimulated recall served as complimentary instruments. The eye-movement data were primarily used to capture participants' re-reading patterns. In contrast to the hypotheses we posed above, the study yielded no meaningful relationships between proficiency and eye-gaze behaviours.

To summarise, although researchers have begun to explore the relationship of proficiency to L2 writers' speed fluency, pausing, and eye-gaze behaviours, previous work on some of these links is still limited. Little is known about how proficiency may be linked to pause-related cognitive processes and eye-gaze behaviours during writing (see, however, Gánem-Gutiérrez & Gilmore, 2008). Also, most previous research has utilised a single elicitation tool to tap the L2 writing process, mainly employing verbal protocols or keystroke logging. Especially few studies have utilised eye-tracking to examine writing processes (cf., Chukharev-Hudilainen et al., 2019; Gánem-Gutiérrez & Gilmore, 2018; Michel et al., 2020; Révész, Michel et al., 2017, 2019). Although multiple-method studies are

on the rise, further research would benefit from combining data sources (e.g., verbal protocols, keystroke-logging, and eye-tracking) to capture how proficiency may relate to different writing processes such as composing, typing, and viewing.

2.3. Writing stage, proficiency, and writing processes

In this research, our other goal was to investigate whether and to what extent the stage of writing may moderate the relationship of L2 proficiency to speed fluency, pausing, and eye-gaze behaviours and cognitive processes underlying pausing. Some researchers (Roca de Larios et al., 2008) have suggested that proficiency may be a key determinant of how L2 writing activities are distributed during the writing process, with high-proficiency learners being better able to decide which writing activities to focus on at what point during writing against the changing task environment. If so, for speed fluency, pausing, and eye-gaze behaviours, we might expect less variation among lower proficiency writers across various writing stages.

Several empirical studies have provided evidence in support of the assumption that higher-proficiency writers are more apt at focusing their attention on different activities during various writing stages, including the previously cited studies by Sasaki (2002), Roca de Larios et al. (2008), and Gánem-Gutiérrez and Gilmore (2018). In Sasaki's (2002) work, higher-proficiency writers spent more time on initial planning and engaged in less pausing subsequently. In contrast, lower-proficiency writers tended to prepare less detailed pre-writing plans and do more local, online planning as their text evolved. More proficient participants in Roca de Larios et al.'s (2008) study also did most of their planning at the beginning of the writing process, followed by a gradual decrease in planning activities. Formulation predominantly occurred in the middle stage, and revision activities were the most frequent in the final stage. Less proficient writers, on the other hand, displayed less variation in writing processes, maintaining a similar temporal distribution of cognitive activities. Although Gánem-Gutiérrez and Gilmore (2018) detected no significant difference in the relationship between proficiency and writing processes depending on time period, qualitative analyses revealed that the activities of lower- and higher-proficiency writers considerably varied across stages. Interestingly, Tillema (2012) did not observe notable links between proficiency, operationalised in terms of L2 English vocabulary size, and writing processes as a function of writing stage (and overall).

Little is known about whether stage of writing, as predicted, would moderate the relationship of proficiency to speed fluency, pausing, and viewing behaviours during writing.

3. Research questions

1. To what extent does proficiency relate to the speed fluency, pausing, and eye-gaze behaviours of L2 writers and pause-related cognitive processes?
2. To what extent does stage of writing influence the relationship of proficiency to the speed fluency, pausing, and eye-gaze behaviours of L2 writers and pause-related cognitive processes?

We operationalised proficiency as participants' combined scores on the TOEFL iBT listening and reading components. We operationalised speed fluency and pausing in terms of key-stroke-logging measures and used eye-tracking equipment to measure gaze behaviours. We employed stimulated recall methodology to tap pause-related cognitive processes—that is, the writing processes (planning, translation, or monitoring) in which participants engaged when they paused. We divided the writing process into five equal time intervals to gauge the effects of writing stage.

4. Methodology

4.1. Design

The dataset for this study comes from a larger project. Unlike in other reports on the project (Michel et al., 2020), our focus here is how proficiency alone and together with writing stage relate to writing processes. Sixty L2 users of English participated in the project. First, they took a research version of the TOEFL iBT listening and reading tests, followed by a typing test. Next, they were administered two independent and two integrated research versions of TOEFL iBT writing tasks. Task type and prompt were counterbalanced across participants following a Latin-square design. In this study, our focus is on the independent tasks, 120 performances altogether. We used the keystroke-logging software InputLog 7.00 (Leijten & Van Waes, 2013) to capture fluency and pausing behaviours and an Eyelink1000 eye-tracker to record eye movements. Immediately after finishing the writing tasks, participants took part in a stimulated recall interview using their last writing performance as a prompt. Thus, the recall sessions yielded retrospective comments for 30 independent task performances given the Latin-square design we used for counterbalancing.

4.2. Participants

All 60 participants were Chinese L2 users of English studying at the University of London. The initial participant pool included 103 students, who were recruited over a period of six months. Of these, 84 students were invited to complete the writing tests. We only invited participants to continue if they achieved scores within the B1 to C1 CEFR thresholds and if and their typing speed was within 2 SDs from the mean for their respective proficiency levels. We excluded a further 24 participants due to technical issues or failure to

complete all writing tasks. Thus, the final participant pool included 60 students. According to their listening and reading scores on the TOEFL iBT test, 20 students were at level B1, 20 at level B2, and 20 at level C1 in terms of the CEFR. We recruited participants from different proficiency levels to have sufficient variability in the dataset to address our research questions. Most participants were female (n=55), and the mean age was 23.76 (SD=3.22). The majority were enroled in masters' programmes (n=55); three students were working towards a bachelor's degree, and two participants were completing a doctorate. As a token of our appreciation, participants received a £30 gift card for their involvement.

4.3. Instruments and procedures

4.3.1. Typing test

Typing Test Pro, an online software, was used to assess participants' keyboarding skills, following Barkaoui (2014). The individual scores for net typing speed were within 2 SDs from the mean for each proficiency level (B1: M=25.32, SD=7.52; B2: M=26.60, SD=11.71; C1: M=36.55, SD=10.90).

4.3.2. Writing tasks

Each participant performed two parallel research versions of the TOEFL iBT independent writing tasks to avoid prompt effects. The independent writing tasks asked participants to write argumentative essays on a given topic. Participants were likely to be familiar with this genre, as the majority had indicated taking the IELTS test, which includes an argumentative essay task. Following TOEFL iBT procedures, 30 min were allowed to complete each independent task. We used the TOEFL iBT research platform to administer the tasks. The actual writing, however, was performed in a Microsoft Word document, given that data is logged in Microsoft Word by the InputLog software. We opened the Microsoft Word document on top of the TOEFL iBT environment, and adjusted it to have the same the size as the TOEFL iBT writing window. In the Microsoft Word file, we set the font size, font type, spacing and editing tools to mimic the TOEFL iBT writing window.

4.3.3. Stimulated recall

The stimulated recall interviews aimed to elicit participants' thoughts during writing. We only asked participants to recall their thoughts in the last writing task to avoid reactivity—that is, the stimulated recall procedure affecting participants' performance on subsequent tasks. To decrease memory decay, the sessions immediately followed the last writing task. We used the keystroke and eye-gaze recordings to prompt participants' recall. To help participants interpret their eye-tracking data, we explained that the circles and lines in the recordings indicated their eye gazes (fixations) and movements between eye-gazes (saccades) respectively, with larger circles denoting longer gazes. We also exemplified the procedure by a brief stimulated-recall interview, which we elicited by a different writing task. In this example, the writer demonstrated recalling their thoughts during pausing and revision prompted by the keystroke and eye-gaze recording of their performance. We encouraged participants to pause the recording whenever they wanted to describe what they were thinking while writing. Additionally, we invited them to share their thoughts if they made pauses, revisions, or interesting or unexpected eye-movements (e.g., regressions, longer fixations) but did not comment on these behaviours. We told participants to describe only what their thoughts were at the time they were engaged in the task. They were also encouraged to comment only if they remembered what they were thinking. The third author, a first language speaker of Mandarin, conducted the stimulated recall interviews in Mandarin. However, participants occasionally switched to English. We video-recorded the sessions to ensure that participants' verbal comments as well as spatial movements (e.g., pointing to the screen) were captured. The stimulated recall sessions took approximately 60–90 min. This procedure was adopted from Révész et al. (2019) and successfully piloted for the purposes of the present study.

4.4. Data collection

Fig. 1 illustrates the data collection procedures. We administered the first session in groups in a computer lab. The session started with obtaining informed consent, followed by the background questionnaire (10 min), a listening (60–90 min) and reading (60–80 min) component of the research form of the TOEFL iBT test, and a typing test (10 min). Then, we invited participants with appropriate proficiency and typing scores to take part in two individual sessions. In both individual sessions, participants completed two out of the four writing tasks (60–70 min). In the second session, participants additionally took part in the stimulated recall interview (60–90 min). We encouraged participants to take a short break between the writing tasks and before the stimulated recall interview.

Before participants began the writing tasks, we started the InputLog software. Next, we calibrated the eye-tracking system, an Eyelink1000 with a temporal resolution of 1000 Hz. We positioned participants 60 cms from the centre of the computer screen, and employed a 9-point calibration grid. Once we had calibrated the participants' eyes, we launched the *SR Research Screen Recorder* software. Finally, we launched the appropriate research version of the TOEFL iBT writing task. To increase ecological validity, we employed the head-free-to-move remote set-up of the eye-tracker, allowing participants to move relatively freely while writing. We re-calibrated participant's eyes before each writing task. We monitored eye-movements on the researcher's screen, and adjusted the

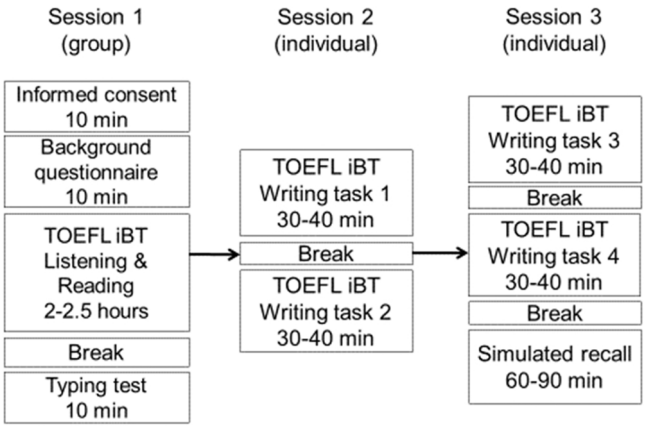

Fig. 1. Data collection procedures.

seating of the participant if we lost tracking. To take account of track loss, we assessed blink number and duration for each participant and computed mean percentages for blink duration (see Table S1 in Supplementary Information Online). We found around 30% of track loss, a lower percentage than reported in previous writing research (e.g., Gánem-Gutiérrez & Gilmore, 2018).

4.5. Data analysis

4.5.1. Proficiency scores

The research forms of the TOEFL iBT listening and reading tests were scored by the Educational Testing Service (ETS) using standard scoring procedures. Scores for both components ranged from 0 to 30 points. The TOEFL iBT writing performances were scored by two independent ETS raters employing TOEFL rubrics.

4.5.2. Speed fluency, pausing, and eye-gaze behaviours

We obtained the speed fluency and pausing measures employing InputLog 7.00 (Leijten & Van Waes, 2013). We adopted a pause threshold of 200 ms, given that this low threshold enabled us to capture lower-level writing processes (Van Waes & Leijten, 2015). We assessed speed fluency with two measures: characters per P-burst (i.e., the number of characters typed between pauses) and mean duration of character production, calculated by dividing the total time spent writing (excluding pauses) by the number of characters typed. We expressed pausing behaviours in terms of pause frequency and length. Using InputLog 7.00, we examined whether pause length and frequency differed according to location, and whether pauses occurred within words, between words, or between sentences. Our rationale for considering pause location was that, as discussed earlier, pauses at higher and lower textual units are expected to be more often associated with planning and linguistic coding processes, respectively (e.g., Révész, Kourtali et al., 2017; Révész et al., 2019). As between-word pauses frequently include one pause before the spacebar is pressed and one pause before the next word begins, Inputlog treats these as a single pause. Similarly, Inputlog counts a pause before a spacebar and a pause before the next sentence as one between-sentence pause.

We used the SR Research Data Viewer software to analyse the eye-gaze data and defined our area of interest as the TOEFL iBT writing window—that is, the box allocated for writing. We employed rather coarse-grained eye-movement indices, as in the TOEFL iBT set-up there is relatively little white space around words and between lines, excluding the possibility of obtaining reliable word-level measures. Given the fact that we focused on the whole writing window, scrolling up and down by participants did not affect data analyses. We computed the following measures (Brunfaut & McCray, 2015): total fixation count; total fixation duration; mean fixation length; number of forward and of backward saccades; median length of forward and backward saccades (in degrees of visual angle); and proportion of backward saccades (i.e., number of backward saccades divided by the number of all saccades). Forward and backward saccades were defined, respectively, as eye-movements with a positive (forward) and negative (backward) angle between the direction of the current saccade and the horizontal plane. To correct for time on task, we divided the raw values by the amount of time spent on task for the following indices: total fixation count, total fixation duration, and number of forward and backward saccades.

We obtained these measures for the entire writing process and for five consecutive stages of writing. For each participant and task, we defined the five stages by dividing the total amount of time participants spent completing the task (M=29.31 min) into five equal intervals (M=5.86 min). In this way, we could detect changes in writing behaviours across stages within and between participants. Dividing the writing process into five (e.g., Gánem-Gutiérrez & Gilmore, 2018) rather than three stages (e.g., Roca de Larios et al., 2008) allowed us to obtain a more sophisticated picture of possible stage variation.

4.5.3. Stimulated recall comments

We analysed the stimulated recall comments in five steps. First, the third author transcribed the data. Second, they identified the

pause-related comments. Each cognitive activity recalled was treated as a segment. A new segment was identified as a switch to a new cognitive activity (Gánem-Gutiérrez & Gilmore, 2018; Tillema, 2012). The segments identified were reviewed by the third author to identify micro-categories emerging from the data. Third, the third author combined the resulting micro-categories into more general categories informed by Kellogg's (1996) model (e.g., "thinking of what to write for second point" coded as planning content point and "I was thinking of an example" coded as planning example were combined into planning content). This process generated the coding scheme shown in Table S2 in the Supplementary Information Online (also presented in Michel et al., 2020). Fourth, the third author coded all the comments using this scheme. To establish inter-coder agreement, another L1 Mandarin speaker with a background in L2 research coded 20% of the data. Cohen's Kappa was high (0.91). Finally, for each participant, we added up the comments per category to form frequency counts, and then calculated the percentage of stimulated recall comments falling into various categories by pause location and in total, with the total number of stimulated recall comment produced by the participant serving as the denominator. We used the resulting proportions in further statistical analyses.

4.6. Statistical analyses

First, we computed Spearman correlations among participants' performance on the research forms of the listening, reading, and writing TOEFL iBT test components. To address the research questions, we constructed linear mixed effects models and linear regressions using the functions *lmer* (for writing behaviours) and *lm* (for stimulated recall comments), respectively. We used mixed-effects models for writing behaviours given that this procedure allowed us to control for the potential random effects of the writing prompts. This was not necessary when analysing the stimulated recall data, as participants were asked to recall their thoughts for only one of the four writing tasks they had performed. We used the *r.squared GLMM* function in the *MuMIn* package to compute effect sizes for the *lmer* models. Specifically, we obtained marginal and conditional R^2 values (R^2m, R^2c) to assess the variance explained by the fixed and random effects in the models, respectively. We set the alpha level at. 05. We used residual plots to check the linearity, homoscedasticity, and normality assumptions for the models. In a few cases, we needed to use squareroot transformations to make sure that the data meet these assumptions.

5. Results

5.1. Proficiency scores

The mean listening and reading proficiency scores (out of 30) were 18.70 (*SD*=6.56) and 21.90 (*SD*=6.03). That is, the average listening and reading scores were equivalent to CEFR B2 level (https://www.ets.org/toefl/score-users/scores-admissions/compare). The mean writing proficiency score, calculated based on the independent and integrated TOEFL iBT performances, was 3.16 (SD=1.01) out of 5.

To establish the relationships between the proficiency measures, we ran a series of Spearman correlations, using the average of the four writing scores in the calculations. The analyses yielded strong correlations (Plonsky & Oswald, 2014) between each pair of tests (listening-reading: rho=0.69, $p < .01$; writing-reading: rho=0.63, $p < .01$; writing-listening: rho=0.59, $p < .01$). We used the combined listening and reading scores as our proficiency measure in further analyses to ensure that our measure of proficiency was independent of the writing performances (our dependent variable).

5.2. RQ1: The relationship of proficiency to writing behaviours and pause-related cognitive processes

We answered the first research question by examining the extent to which proficiency related to speed fluency, pausing, and eye-gaze behaviours as well as pause-related cognitive processes. Tables 1–5 give the descriptive statistics for these measures. To address the first research question, first we conducted a series of linear mixed effects analyses for writing behaviours. We began the modelling by constructing null models, which only contained random intercepts and slopes for Participant and Prompt. In each model, the dependent variable was a measure of writing behaviour. Then, we added Proficiency to the null model and conducted a likelihood ratio test to assess whether Proficiency improved model fit. If we identified a significant effect for Proficiency, we constructed a maximal

Table 1
Fluency measures by stage (n = 60).

Stage	Characters per P-burst				Active writing time per character (min)			
	M	SD	95%CI Low	95% CI Up	M	SD	95% CI Low	95% CI Up
1	1.53	0.44	1.46	1.61	0.21	0.15	0.18	0.23
2	1.56	0.47	1.48	1.65	0.12	0.04	0.11	0.13
3	1.56	0.44	1.48	1.64	0.11	0.04	0.10	0.12
4	1.57	0.45	1.49	1.65	0.11	0.03	0.10	0.12
5	1.45	0.42	1.37	1.52	0.18	0.18	0.15	0.21
Tot	1.53	0.44	1.50	1.57	0.15	0.11	0.14	0.15

Table 2
Pausing measures by stage and pause location (n = 60).

Stage	Pause location	Pause number per minute				Median pause length (ms)			
		M	SD	95% CI Low	95% CI Up	M	SD	95% CI Low	95% CI Up
1	Ww	22.05	8.13	20.61	23.51	0.30	0.04	0.29	0.31
1	Bw	12.25	4.56	11.46	13.08	0.80	0.30	0.75	0.87
1	Bs	0.55	0.38	0.48	0.62	11.03	25.59	6.40	15.65
1	Tot	50.25	15.60	47.52	53.03	0.39	0.06	0.37	0.40
2	Ww	24.55	7.72	23.21	26.01	0.30	0.04	0.29	0.31
2	Bw	14.58	4.58	13.75	15.43	0.79	0.32	0.74	0.83
2	Bs	0.65	0.37	0.58	0.71	3.00	3.66	2.33	3.66
2	Tot	58.60	12.79	56.29	60.91	0.39	0.06	0.38	0.39
3	Ww	26.07	7.96	24.68	27.49	0.30	0.04	0.29	0.30
3	Bw	15.23	4.64	14.38	16.07	0.78	0.31	0.72	0.84
3	Bs	0.65	0.39	0.58	0.72	2.57	2.55	2.11	3.03
3	Tot	61.03	12.23	58.82	63.24	0.38	0.06	0.37	0.40
4	Ww	26.55	8.70	24.98	28.09	0.29	0.04	0.29	0.30
4	Bw	15.67	4.60	14.86	16.49	0.79	0.33	0.73	0.85
4	Bs	0.68	0.47	0.60	0.77	3.19	5.24	2.25	4.14
4	Tot	62.71	13.33	60.30	65.03	0.39	0.06	0.37	0.39
5	Ww	25.96	20.15	22.59	29.85	0.30	0.04	0.29	0.30
5	Bw	14.60	10.06	12.97	16.47	0.76	0.27	0.72	0.81
5	Bs	0.60	0.50	0.52	0.69	2.43	3.12	1.82	3.04
5	Tot	62.32	40.44	55.73	69.85	0.40	0.07	0.39	0.41
Tot	Ww	25.04	11.65	24.16	26.00	0.30	0.04	0.29	0.30
Tot	Bw	14.46	6.18	13.97	14.98	0.78	0.31	0.76	0.81
Tot	Bs	0.63	0.43	0.59	0.66	4.46	12.37	3.43	5.50
Tot	Tot	58.98	22.18	57.31	60.96	0.39	0.06	0.38	0.39

Note. Ww=within words, Bw=between words, Bs=between sentences.

Table 3
Eye-fixation Measures by Stage (n = 60).

Stage	Total fixation duration (ms)				Fixation count				Mean fixation length (ms)			
	Mean	SD	95% CI Low	95% CI Up	Mean	SD	95% CI Low	95% CI Up	Mean	SD	95% CI Low	95% CI Up
1	25,121.91	1040.87	24,935.68	25,308.15	61.39	6.83	60.17	62.61	413.13	2.14	412.75	413.52
2	33,417.59	459.41	33,335.39	33,499.79	85.96	8.66	84.41	87.51	392.42	18.36	389.14	395.71
3	33,089.21	1455.67	32,828.75	33,349.66	87.09	9.78	85.34	88.84	380.77	3.58	380.13	381.41
4	31,565.69	1461.53	31,304.19	31,827.19	83.46	10.04	81.66	85.25	373.27	5.89	372.22	374.33
5	30,570.49	962.71	30,398.24	30,742.74	90.87	7.21	89.58	92.16	331.44	29.84	326.10	336.78
Tot	31,894.38	8050.13	30,454.02	33,334.73	84.55	24.53	80.16	88.94	374.43	8.87	372.84	376.02

model by adding a by-prompt random slope for proficiency to examine prompt-by-proficiency variation.

Out of 18 analyses, the likelihood ratio tests revealed that proficiency had a significant relationship to three indices: characters per P-burst, $\chi^2(1)= 8.53$, $p < .01$; median pause length between words, $\chi^2(1)= 8.32$, $p < .01$; and number of pauses between sentences, $\chi^2(1)= 4.82$, $p = .03$ (see Table S3 in Supplementary Information Online for all test results). As shown in Table 6, follow-up mixed effects models found that, as proficiency increased, participants produced more characters per P-bursts, paused for shorter periods between words, and paused more frequently between sentences. These relationships are illustrated in Fig. 2. The effect sizes for characters per P-burst and between-word pause length were considerable, with proficiency explaining 13% of the variation. For pause number between sentences, we found the effect size to be small; proficiency accounted for 6% of the variance.

Next, we conducted a series of simple regression analyses to assess whether proficiency was related to the proportion of stimulated recall comments focusing on planning, translation, or monitoring. In each regression, proficiency was the independent variable, and the proportion of stimulated recall comments on planning, translation or monitoring at a particular pause location served as the dependent variable. As shown in Table 7, the analyses generated no proficiency effects.

5.3. RQ2: writing stage, proficiency, and writing behaviours and pause-related cognitive processes

The second research question investigated whether writing stage influenced the extent to which proficiency related to L2 writers' speed fluency, pausing, and eye-gaze behaviours and pause-related cognitive processes. First, we conducted another series of linear mixed effects analyses including the writing behaviour indices. In our initial models, we included Proficiency and Stage as fixed effects and Participant and Prompt as random effects. The dependent variable was a writing index. Next, we added the Interaction between

Table 4
Saccade measures by stage (n = 60).

	Backward Saccade Number				Backward Saccade Median Length*				Forward Saccade Number				Forward Saccade Median Length*				Proportion of Backward Saccades			
	M	SD	95% CI Lower	95% CI Upper	M	SD	95% CI Lower	95% CI Upper	M	SD	95% CI Lower	95% CI Upper	M	SD	95% CI Lower	95% CI Upper	M	SD	95% CI Lower	95% CI Upper
1	54.20	2.07	53.83	54.57	2.60	0.17	2.57	2.63	59.29	2.16	58.91	59.68	2.41	0.15	2.38	2.43	0.48	0.00	0.48	0.48
2	48.98	3.00	48.45	49.52	2.57	1.43	2.31	2.83	54.74	1.50	54.47	55.01	2.26	0.41	2.19	2.34	0.47	0.01	0.47	0.47
3	48.55	4.07	47.83	49.28	2.54	1.24	2.32	2.76	52.95	4.35	52.17	53.73	2.42	1.06	2.23	2.61	0.48	0.01	0.48	0.48
4	45.83	4.59	45.01	46.65	2.32	0.32	2.26	2.38	50.21	4.47	49.41	51.01	2.16	0.14	2.14	2.18	0.48	0.02	0.47	0.48
5	52.13	7.47	50.80	53.47	2.36	0.87	2.21	2.52	56.34	5.83	55.29	57.38	2.29	1.08	2.10	2.49	0.48	0.03	0.47	0.48
Tot	51.68	16.35	48.76	54.61	2.46	0.58	2.36	2.56	56.35	17.37	53.24	59.46	2.26	0.41	2.19	2.33	0.48	0.00	0.48	0.48

Note. The unit of measurement is degree of visual angle.

Table 5
Stimulated recall comments for pausing by stage and pause location (n = 30).

Stage	Pause location	Total number of comments		Planning		Translation		Monitoring		No recall*		Other	
		M	M%	M	M%	M	M%	M	M%	M	M%	M	M%
1	Ww	2.87	15%	0.30	2%	2.00	11%	0.03	0%	0.07	0%	0.47	3%
1	Bw	6.80	36%	2.13	12%	3.40	17%	0.48	3%	0.15	1%	0.63	3%
1	Bs	2.03	11%	1.00	5%	0.37	2%	0.53	3%	0.03	0%	0.10	1%
1	Tot	18.83	100%	5.73	31%	9.23	48%	1.63	9%	0.40	2%	1.83	10%
2	Ww	2.33	11%	0.17	1%	1.87	9%	0.10	0%	0.03	0%	0.17	1%
2	Bw	6.65	31%	1.92	9%	3.48	15%	0.60	3%	0.15	1%	0.50	2%
2	Bs	4.43	21%	2.00	10%	0.77	4%	1.23	6%	0.23	1%	0.20	1%
2	Tot	21.20	100%	6.40	32%	9.70	44%	3.03	14%	0.60	3%	1.47	7%
3	Ww	1.90	10%	0.30	1%	1.10	6%	0.03	0%	0.10	1%	0.37	2%
3	Bw	6.52	32%	1.78	9%	3.38	16%	0.62	3%	0.13	1%	0.60	3%
3	Bs	4.23	21%	1.80	9%	0.60	3%	1.33	7%	0.17	1%	0.33	2%
3	Tot	20.23	100%	6.07	30%	8.50	41%	3.07	16%	0.57	3%	2.03	11%
4	Ww	1.87	10%	0.13	1%	1.43	8%	0.07	0%	0.03	0%	0.20	1%
4	Bw	6.45	32%	1.80	9%	3.42	17%	0.45	2%	0.22	1%	0.57	3%
4	Bs	3.17	16%	1.33	7%	0.43	2%	1.20	6%	0.10	0%	0.10	0%
4	Tot	19.43	100%	5.30	27%	8.83	46%	3.27	17%	0.60	3%	1.43	7%
5	Ww	1.53	7%	0.17	1%	1.20	6%	0.00	0%	0.07	0%	0.10	0%
5	Bw	5.18	26%	1.65	8%	2.70	14%	0.35	2%	0.17	1%	0.33	2%
5	Bs	2.80	14%	1.30	6%	0.33	2%	0.83	4%	0.20	1%	0.14	1%
5	Tot	19.30	100%	4.90	25%	7.27	37%	5.27	27%	0.77	4%	1.10	6%
Tot	Ww	1.50	11%	1.07	1%	7.60	8%	0.23	0%	0.30	0%	1.30	1%
Tot	Bw	31.60	32%	9.28	9%	16.38	16%	2.50	2%	0.82	1%	2.62	3%
Tot	Bs	16.67	17%	7.43	8%	2.50	2%	5.13	5%	0.73	1%	0.87	1%
Tot	Tot	99	100%	28.40	29%	43.53	44%	16.27	16%	2.93	3%	7.87	8%

Note. Percentages do not necessarily add up to 100% as the values indicate averages across participants.
*No recall refers to instances when participants did not recall their thoughts when prompted.

Table 6
Results for maximum models examining the effects of proficiency on writing behaviours.

Dependent variable	Fixed effects							Random effects	
	Pred	*Est*	*SE*	*T*	*p*	R^2m	R^2c	Factor	*SD*
Speed fluency									
Characters per P-burst	Prof	0.01	< 0.01	2.99	< 0.01	0.13	0.92	Part (Int)	0.36
								Prompt(Int)	0.03
								Prompt(Prof)	< 0.01
Pausing									
Median pause length between words	Prof	-0.01	< 0.01	-2.98	< 0.01	0.13	0.93	Part (Int)	0.26
								Prompt(Int)	0.02
								Prompt(Prof)	< 0.01
Pause number between sentences	Prof	0.01	< 0.01	2.12	0.05	0.06	0.59	Part (Int)	0.19
								Prompt(Int)	< 0.01
								Prompt(Prof)	< 0.01

Note. Int=Intercept, Prof=Proficiency.

Proficiency and Stage to the models, given that the Interaction was our predictor of interest. A significant interaction would mean that, depending on stage of writing, proficiency had a differential relationship to the measure of writing behaviour in the model. In other words, when a significant interaction effect was detected, this indicated that the relationship between proficiency and the writing behaviour varied across stages of writing. When a significant Interaction effect emerged, we constructed a maximal model by adding by-prompt and by-participant random slopes for Stage and by-prompt random slopes for Proficiency. The resulting maximal models failed to converge for all variables; thus, for each model, we removed the random effect parameters explaining the least variance one by one until convergence was reached. We ran the best-fitting models for the various stages as reference points to identify all possible interactions between proficiency and writing stage.

Out of 18 analyses, the likelihood ratio tests yielded a significant interaction effect for five indices: characters per P-burst, $\chi^2(1)= 27.45$, $p < 0.01$; active writing time, $\chi^2(1)= 25.93$, $p < 0.01$; median pause length total, $\chi^2(1)= 9.48$, $p = 0.05$; median pause length between words, $\chi^2(1)= 13.44$, $p < 0.01$; and mean fixation length, $\chi^2(1)= 19.58$, $p < 0.01$ (see Table S4 in the Supplementary Information Online for all test results).

The results for the follow-up maximal models are available in Table S5. Table 8 presents the significant interactions identified, and Fig. 3 illustrates these relationships. The speed fluency measures, characters per P-burst, and active writing time indicate that higher-

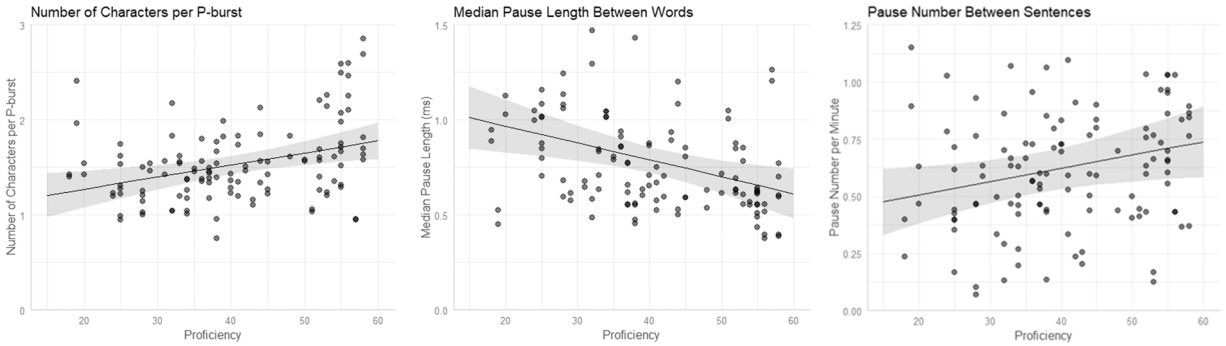

Fig. 2. Proficiency predicting writing behaviours.

Table 7
Results for regression models examining the effects of proficiency on proportion of stimulated recall comments by focus and pause location.

Dependent variable	Pred	*Est*	*SE*	*t*	*p*	R^2
Planning-related comments						
Within words	Prof	< 0.01	< 0.01	0.38	0.71	< 0.01
Between words	Prof	< 0.01	0.01	0.68	0.50	< 0.02
Between sentences	Prof	< 0.01	< 0.01	1.57	0.13	0.08
Total	Prof	< 0.01	< 0.01	1.51	0.14	0.08
Translation-related comments						
Within words	Prof	< −0.01	< 0.01	-1.06	0.30	0.04
Between words	Prof	< −0.01	< 0.01	-0.29	0.77	< 0.01
Between sentences	Prof	< −0.01	< 0.01	-0.41	0.69	< 0.01
Total	Prof	< −0.01	< 0.01	-0.98	0.33	0.03
Monitoring-related comments						
Within words	Prof	< 0.01	< 0.01	0.36	0.72	< 0.01
Between words	Prof	< −0.01	< 0.01	-0.88	0.39	0.03
Between sentences	Prof	< −0.01	< 0.01	-1.65	0.11	0.09
Total	Prof	< 0.01	< 0.01	0.07	0.94	< 0.01

Note. Prof=Proficiency.

Table 8
Significant proficiency-stage interaction effects identified by maximum models for writing behaviours.

Dependent variable/Ref level	Pred	*Est*	*SE*	*t*	*p*
Characters per P-burst					
Stage 1	Prof:Stage5	-0.01	< 0.01	-3.75	< 0.01
Stage 2	Prof:Stage5	-0.01	< 0.01	-4.98	< 0.01
Stage 3	Prof:Stage5	-0.01	< 0.01	-3.96	< 0.01
Stage 4	Prof:Stage2	< 0.01	< 0.01	2.07	0.04
Stage 4	Prof:Stage5	-0.01	< 0.01	-2.91	< 0.01
Active writing time					
Stage 1	Prof:Stage5	0.01	< 0.01	2.51	0.01
Stage 2	Prof:Stage5	< 0.01	< 0.01	3.15	< 0.01
Stage 3	Prof:Stage5	< 0.01	< 0.01	2.66	0.01
Stage 4	Prof:Stage5	< 0.01	< 0.01	2.63	0.01
Median pause length between words					
Stage 1	Prof:Stage5	< 0.01	< 0.01	2.51	0.01
Stage 2	Prof:Stage5	< 0.01	< 0.01	3.00	< 0.01
Stage 4	Prof:Stage2	< 0.01	< 0.01	-2.18	0.03
Median pause length total					
Stage 2	Prof:Stage5	< 0.01	< 0.01	2.70	< 0.01
Stage 3	Prof:Stage5	< 0.01	< 0.01	2.14	0.03
Stage 4	Prof:Stage5	< 0.01	< 0.01	2.62	0.01
Mean fixation length					
Stage 1	Prof:Stage4	-1.82	0.86	-2.12	0.04
Stage 1	Prof:Stage5	-3.24	1.01	-3.21	< 0.01
Stage 2	Prof:Stage5	-2.77	0.92	-3.02	< 0.01
Stage 3	Prof:Stage5	-1.79	0.88	-2.03	0.04

Note. Prof=Proficiency.

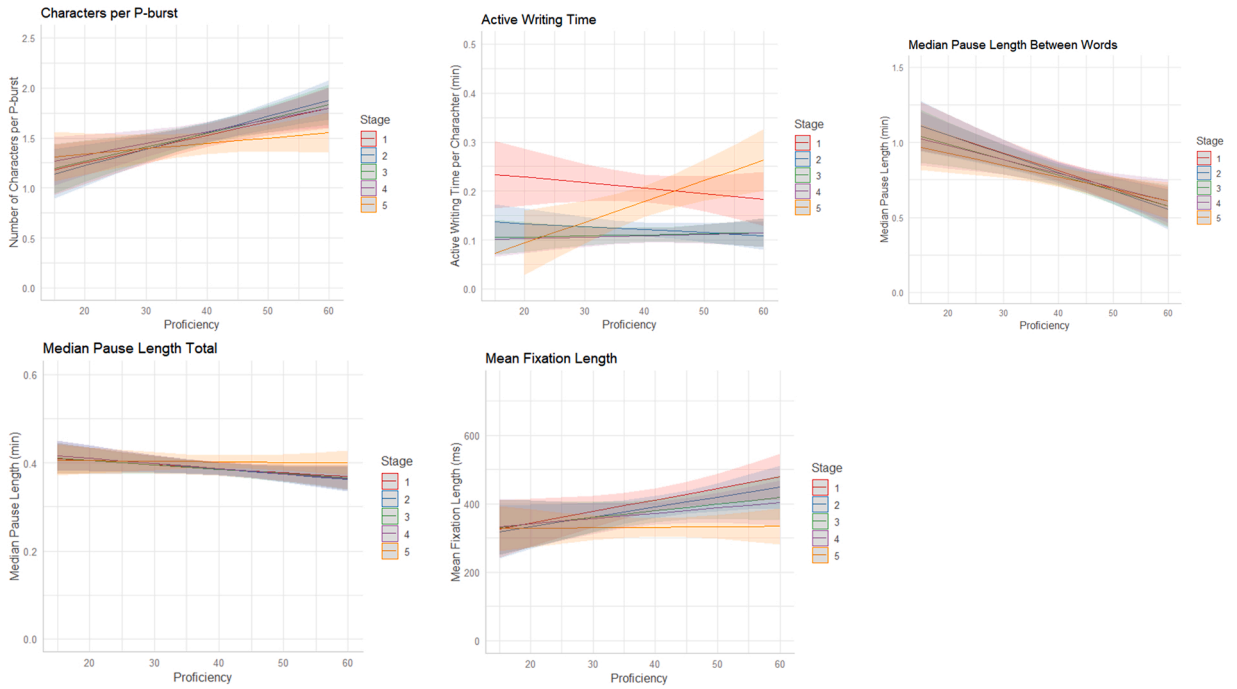

Fig. 3. Effects of stage on the relationship of proficiency to writing behaviours.

proficiency writers engaged in less speedy writing in stage 5 as compared to earlier stages (1−4). For characters per P-burst, we also found that, with growing proficiency, participants produced fewer characters per P-burst during stage 4 as compared to stage 2. Similar trends were observed for pause length: higher-proficiency writers paused longer in later than earlier stages of writing. More proficient writers paused longer overall during stage 5 as compared stages 2–4. With increasing proficiency, participants also produced longer between-word pauses in stage 5 than stages 2 and 4 and in stage 4 in comparison to stage 2. Finally, as proficiency increased, participants produced shorter fixations in stage 5 than stages 1–3 and in stage 4 as compared to stage 1.

Finally, we carried out a series of multiple regressions to examine whether writing stage influenced the link between proficiency and the proportion of planning-, translation-, and monitoring-related stimulated recall comments at different pause locations. Proficiency, writing stage, and their interaction were the independent variables, and a stimulated recall proportion functioned as the dependent variable. Again, the interaction was the predictor of interest (see Tables S6 in the Supplementary Information Online for all regression results).

The regressions yielded a significant interaction effect for four stimulated recall proportions: within-word and total planning-related comments, and within-word and total monitoring-related comments. These relationships are summarised in Table 9 and in Fig. 4. As proficiency grew, participants made more planning-related comments at earlier stages. Specifically, more proficient writers reported more planning during within-word pauses in stage 1 as compared to stage 4. For pauses overall, higher-proficiency writers referred to more planning in stage 2 than stage 5. Turning to monitoring-related comments, with increasing proficiency, participants produced more monitoring-related comments in stage 4 than stages 1, 3 and 5 when describing within-word pauses. More proficient writers also referred to more monitoring in stage 4 than stage 2 when reporting their thoughts for pauses overall. However, we need to interpret the results for within-word monitoring-related comments with caution, as the number of comments falling into this category was low.

Table 9
Significant proficiency-stage interaction effects identified by regression for proportion of stimulated comments.

Reference level	Pred	*Est*	*SE*	*t*	*p*
Planning-related comments within words					
Stage 1	Prof:Stage4	< 0.01	< 0.01	-2.13	0.03
Planning-related comments in total					
Stage 2	Prof:Stage5	-0.01	< 0.01	-2.18	0.03
Monitoring-related comments within words					
Stage 1	Prof:Stage4	< 0.01	< 0.01	3.03	< 0.01
Stage 3	Prof:Stage4	< 0.01	< 0.01	3.03	< 0.01
Stage 4	Prof:Stage5	< 0.01	< 0.01	-2.37	0.02
Monitoring-related comments in total					
Stage 2	Prof:Stage4	< 0.01	< 0.01	2.14	0.03

Note. Prof=Proficiency

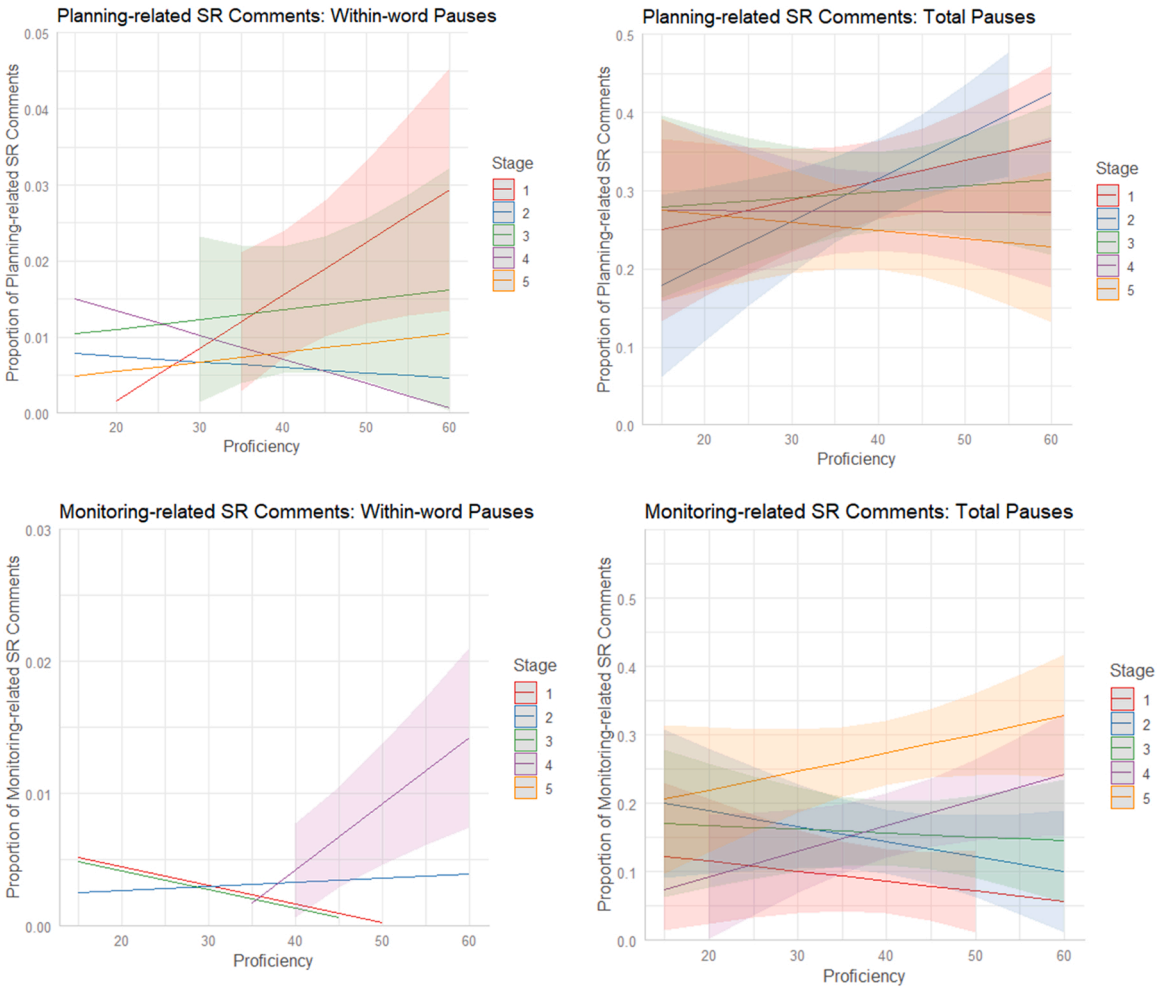

Fig. 4. Effects of stage on the relationship of proficiency to pause-related cognitive processes.

6. Discussion

6.1. *Proficiency and L2 writing behaviours and processes*

We addressed the first research questions by examining to what extent proficiency was related to L2 writer's speed fluency, pausing, and eye-gaze behaviours and pause-related cognitive processes. Based on Kellogg's (1996) model of writing, we predicted that, given the competing demands on writing subprocesses, lower proficiency would be associated with decreased speed fluency and more frequent and longer pauses, especially at lower textual units (e.g., Révész, Kourtali et al., 2017; Révész et al., 2019).

In line with our predictions, we found that proficiency had a considerable relationship to writing fluency, as measured by characters per P-burst, explaining 13% of the variation in speed fluency. Our results confirm the findings of previous research that there is a positive link between proficiency and the speed fluency of L2 writers (Barkaoui, 2019; Kowal, 2014; Palviainen et al., 2012; Spelman Miller et al., 2008).

Our predictions for pausing received partial confirmation. As expected, higher-proficiency participants paused for shorter periods at lower textual units (between words), but contrary to our predictions, they paused more frequently between sentences. These findings also run counter to those of Barkaoui (2019). While in our research higher-proficiency writers paused more often but for shorter periods, Barkaoui (2019) found that more proficient participants paused less often overall. A possible explanation might be that, following Van Waes and Leijten (2015), we used a considerably lower pause threshold (200 ms) than Barkaoui's (2019) 2 s. Probably, a lower pause threshold enabled us to detect more of the pauses produced by higher-proficiency writers. It is likely that more proficient writers produced pauses shorter than 2 s more frequently, due to their more automatized linguistic and writing skills. It is not surprising either that this difference was observed for pauses between sentences in our study. Pauses between sentences tend to be longer, reflecting syntactic encoding and/or planning processes, which are less likely to take shorter than 2 s for lower-proficiency writers with less automatized skills.

Given the results of Révész et al. (2019) and Chukharev-Hudilainen et al. (2019), we expected more local viewing behaviours (shorter saccades) by lower-proficiency students and more global viewing behaviours (longer saccades) by higher-proficiency writers. These predictions, however, were not confirmed. None of the eye-tracking indices varied according to proficiency. One reason for this might be that our relatively coarse eye-gaze measures did not allow for word- or sentence-level analyses of gaze patterns.

Although we observed some differences in terms of pause length and frequency, we found no significant relationship of proficiency to how planning-, translation-, and monitoring-related stimulated recall comments were distributed at various pause locations. This finding suggests that, despite differences in pause length and frequency at some locations, more and less proficient writers allocated their conscious attention in similar proportions to various types of cognitive activities at different pause locations. As an anonymous reviewer suggested, a possible explanation for this might be that lower-proficiency writers produced more verbalisations describing compensatory strategies (i.e., making up for lack of linguistic knowledge), whereas higher-proficiency writers focused more on upgrading (i.e., fine-tuning linguistic expression; López Serrano et al., 2019).

6.2. Stage of writing, proficiency, and L2 writing behaviours and processes

Our second research question was concerned with whether the relationship of proficiency to L2 writers' speed fluency, pausing, and eye-gaze behaviours and pause-related cognitive processes would differ across writing stages. In light of the results of previous empirical research (e.g., Roca de Larios et al., 2008), we expected less variation in writing behaviours and associated processes among lower-proficiency participants, given that they might be less able to determine what writing activities to prioritise in response to the changing task environment.

Our analyses revealed that, indeed, writing stage significantly impacted the relationship between proficiency and five writing behaviour indices (characters per P-burst, active writing time, median pause length between words and in total, and mean fixation length). As proficiency grew, writers engaged in slower writing, paused longer, and produced shorter mean fixations in later stages (4, 5) than earlier stages of writing. The stimulated recall comments also indicated that, depending on proficiency, there was some variation in writing processes across stages, as reflected in two planning- and two monitoring-related measures. With increasing proficiency, participants referred to more planning in earlier than later stages and to more monitoring in later than earlier stages. To summarise, the three data sources converge on the finding that higher proficiency writers engaged in more varied activities across writing stages. Specifically, monitoring activities were more probable towards the end of the writing process, as manifest in less speedy writing, longer pausing, shorter fixations (probably associated with re-reading familiar, previously written text), and more monitoring-related stimulated recall comments. In addition, the stimulated recall comments indicate that higher-proficiency writers were more likely to engage in planning in the earlier rather than final stages. Interestingly, greater engagement in planning was not reflected in the keystroke-logging and eye-fixation indices, as no differences were found between the initial and mid-stages in terms of these measures. Possibly, initial planning also involved producing text (e.g., brainstorming through writing), masking any differences in terms of the keystroke-logging and eye-gaze indices.

These findings, overall, are in line with the results of Sasaki (2002), Roca de Larios et al. (2008), and Gánem-Gutiérrez and Gilmore (2018) in that, depending on proficiency, participants engaged in different writing activities to a differential degree across various stages of writing. Like in the study by Roca de Larios et al. (2008), participants seemed to have engaged in more monitoring in later stages of writing. Our results are also aligned with the findings of Sasaki (2002) and Roca de Larios et al. (2008), who reported more planning time for higher-proficiency writers at the initial stages of writing. We found, however, no evidence for translation processes dominating the middle stages. This is different from the pattern reported by Roca de Larios et al. (2008), but similar to the results of Tillema (2012) and Gánem-Gutiérrez and Gilmore (2008), who found no evidence either that higher proficiency writers engaged in more translation in the middle stages. A possible explanation for lower alignment of our and others' results with those of Roca de Larios et al. (2008) may lie in that the high-proficiency group in Roca de Larios et al.'s research were, unlike in the rest of the studies, graduates from a five-year English degree programme. Their background might have made them more inclined to focus on quality of expression throughout the whole writing process.

To summarise, the three data sources converge on the finding that higher proficiency writers carried out more varied activities across writing stages. We found the strongest evidence for more proficient writers engaging in greater amount of monitoring in later stages, as all three data sources yielded results consistent with this. The stimulated recall data also indicated that higher-proficiency writers were more likely to be involved in planning in earlier stages of writing, but this was not reflected in the keystroke-logging and eye-tracking indices.

6.3. Theoretical, practical, and methodological implications

Now we turn to a discussion of the broader significance and implications of our results. At the theoretical level, our study yielded some useful information about the applicability of Kellogg's (1996) model to L2 writing. As we predicted based on this model, less proficient writers appeared less able to deal with the parallel demands posed by various writing subprocesses, probably given the need to focus more on linguistic encoding. This was reflected in lower speed fluency and longer pauses at lower textual units, which are more likely to reflect linguistic encoding issues. The results we obtained for writing stage are also consistent with our prediction that greater pressure on linguistic encoding will make lower proficiency writers less capable of switching attention among writing subprocesses. Lower-proficiency writers proved less able to distribute their attentional resources effectively during different writing stages. For example, while more proficient writers focused more on monitoring in later stages of writing, less proficient writers displayed less variation in cognitive activities throughout writing. Nevertheless, some of our theoretical predictions deduced from Kellogg's model

were not borne out by the data. This suggests that, to enable a full theoretical account of L2 writing processes, the construction of a model specific to L2 writing is warranted.

Our results also have some possible implications for language assessment. We found that proficiency had a relationship with three writing behaviour indices: characters per P-burst, pause length between words, and pause number between sentences. Based on our results and those of previous research, speed fluency appears to have a reliable link to proficiency. This suggests that it might be worth conducting follow-up validation studies exploring whether indices of speed fluency could be usefully integrated into automatic scoring tools. That is, it would be interesting to investigate whether speed fluency measures could help increase the reliability of automatic writing evaluation outcomes, which are currently based exclusively on analyses of writing products using statistical modelling, natural language processing, and computational linguistics tools. If our results for pausing are confirmed in future research, pausing measures could also be further investigated with the same purpose in mind. We found, however, little evidence for a relationship between proficiency and the eye-gaze measures, suggesting that the coarse eye-gaze indices obtained in this study may be less amenable to informing scoring practices.

Besides assessment, the study yielded some tentative implications for pedagogy. The fact that more proficient writers showed more variance in activities across writing stages implies that lower-proficiency writers might benefit from explicit instruction on how to allocate their attentional resources during writing (e.g., students could be encouraged to pay more attention to monitoring in later writing stages). Clearly, it is for further research to assess the validity of this suggestion.

Finally, it is worthwhile to consider some methodological implications. Employing three data sources (keystroke-logging, eye-tracking, and stimulated recall) proved to be valuable in achieving a more valid and fuller picture of the writing process. In some cases, the various elicitation instruments yielded converging findings, which allowed for reaching less tentative conclusions. For example, we found that the patterns emerging from all three data sources were consistent with the assumption that higher proficiency writers engage in more monitoring towards the end of the writing process. In other cases, not all three methods revealed significant differences across stages. For example, if we had only used stimulated recall, we would have failed to see some differences between the middle and later stages of writing. In light of these observations, we would encourage researchers to combine data sources when investigating writing processes.

6.4. Limitations and future research directions

Our study suffers from several limitations. First, all participants were Chinese university students, which limits the generalisability of the results. Second, our eye-gaze measures were relatively coarse, focusing on the entire writing window. Given the constraints of the TOEFL iBT research platform, the font size was not sufficiently large to carry out analyses at the word level (cf., Chukharev-Hudilainen et al., 2019; Révész, Michel et al., 2017; Révész et al., 2019). Further research would profit from investigating eye-gaze behaviours at different proficiency levels employing more fine-grained indices. Third, although using a threshold of 200 ms allowed us to capture lower-level writing processes, this threshold made it impossible to compare our results directly to most existing studies, which traditionally used longer pause thresholds. Future studies would benefit from adopting different pause thresholds to address this shortcoming (Van Waes & Leijten, 2015). Another weakness of our pause analysis was that we did not run our analyses for before-space and word-initial pauses separately, but treated a combination of these pauses as between-word and between-sentence pauses using Inputlog. It would be interesting to explore whether separating these latencies would yield differential patterns. A further limitation of our research was that we triangulated our data sources only at the group level. Future studies could triangulate the data at the individual level to yield more fine-tuned information about writing processes. Finally, our proficiency measure was exclusively based on receptive skills. Using a full TOEFL test assessing all four skills as an index for proficiency, however, would have had the disadvantage of the TOEFL writing component being part of the assessment, leading to some overlap between our independent factor (proficiency) and dependent variables (writing process indices).

7. Conclusion

In this study, our aim was to examine the relationship of proficiency to L2 speed fluency, pausing, and eye-gaze behaviours and pausing-related cognitive processes. In addition, we explored whether these links varied across different stages of writing. Two main findings emerged. First, we found proficiency to have the strongest relationship to speed fluency and pause length. Second, there was considerably less variation in cognitive activities among lower- than higher-proficiency writers across stages. From a practical perspective, these results suggest that it might be worthwhile to consider in further validation studies whether speed fluency would be useful to integrate into automatic scoring procedures. On the methodological front, the study demonstrates that triangulating keystroke logging, eye-tracking, and stimulated recall is helpful in obtaining an in-depth account of how proficiency relates to L2 writing behaviours and processes.

Data availability

The authors do not have permission to share data.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jslw.2022.100927.

References

Abdel Latif, M. M. M. (2012). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics, 34*(*1*), 99–105. https://doi.org/10.1093/applin/ams073

Barkaoui, K. (2014). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL iBT writing tasks. *Language Testing, 31*(2), 241–259. https://doi.org/10.1177/0265532213509810

Barkaoui, K. (2019). What can L2 writers' pausing behaviour tell us about their L2 writing processes? *Studies in Second Language Acquisition, 41*(3), 529–554. https://doi.org/10.1017/S027226311900010X

Brunfaut, T., & McCray, G. (2015). *Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. ARAGs Research. Reports Online; Vol. AR/2015/001. The British Council.

Chukharev-Hudilainen, E., Feng, H. H., Saricaoglu, A., & Torrance, M. (2019). Combined deployable keystroke logging and eyetracking for investigating cognitive processes that underlie L2 writing. *Studies in Second Language Acquisition, 41*, 583–604.

Cumming, A. (1989). Writing expertise and second-language proficiency. *Language Learning: A Journal of Applied Linguistics, 39*, 81–141. https://doi.org/10.1111/j.1467-1770.1989.tb00592.x

Cumming, A. (2016). Theoretical orientations to L2 writing. In R. M. Manchón, & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 65–88). Mouton: De Gruyter.

Flower, L., & Hayes, J. R. (1980). The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication, 31*, 21–32. https://doi.org/10.2307/356630

Galbraith, D. (2009). Cognitive models of writing. *German as a Foreign Language, 2–3*, 7–22.

Gánem-Gutiérrez, G. A., & Gilmore, A. (2018). Tracking the real-time evolution of a writing event: Second language writers at different proficiency levels. *Language Learning, 68*(2), 469–506. https://doi.org/10.1111/lang.12280

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication, 29*(3), 369–388. https://doi.org/10.1177/0741088312451260

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly, 8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844

Kellogg, R. (1996). A model of working memory in writing. In M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–72). Lawrence Erlbaum Associates.

Khuder, B., & Harwood, N. (2015). L2 writing in test and non-test situations: Process and product. *Journal of Writing Research, 6*, 233–278.

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing, 21*(4), 390–403. https://doi.org/10.1016/j.jslw.2012.09.003

Kowal, I. (2014). Fluency in second language writing: A developmental perspective. *Studia Linguistica Universitatis Iagellonicae Cracoviensis, 131*(*3*), 229–246.

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

Lindgren, E., Spelman Miller, K. S., & Sullivan, K. P. (2008). Development of fluency and revision in L1 and L2 writing in Swedish high school years eight and nine. *International Journal of Applied Linguistics, 156*, 133–151.

López-Serrano, S., Roca de Larios, J., & Manchón, R. (2019). Language reflection fostered by individual L2 writing tasks: Developing a theoretically motivated and empirically based coding system. *Studies in Second Language Acquisition, 41*(3), 503–527. https://doi.org/10.1017/S0272263119000275

Manchón, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal dimension and problem-solving nature of foreign language composing. Implications for theory. In R. M. Manchón (Ed.), *Foreign language writing: Learning, teaching and research* (pp. 102–124). Multilingual Matters.

McKee, H. A., & DeVoss, D. (Eds.). (2007). *Digital writing research: Technologies, methodologies, and ethical issues*. Hampton Press.

Michel, M., Révész, A., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed methods study. *Second Language Research, 36*(3), 277–304. https://doi.org/10.1177/0267658320915501

Nicolás-Conesa, F., Roca de Larios, J., & Coyle, Y. (2014). Development of EFL students' mental models of writing and their effects on performance. *Journal of Second Language Writing, 24*, 1–19. https://doi.org/10.1016/j.jslw.2014.02.004

Palviainen, Å., Kalaja, P., & Mäntylä, K. (2012). Development of L2 writing: fluency and proficiency. *AFinLA-E: Soveltavan Kielitieteen Tutkimuksia, 4*, 47–59.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8*(4), 252–266. https://doi.org/10.1016/j.jeap.2009.05.001

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

Polio, C. (2012). Second language writing. In S. Gass, & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 319–334). Routledge.

Polio, C., & Freedman, D. (2017). *Understanding, evaluating and conducting second language writing research*. Routledge/Taylor & Francis,.

Raimes, A. (1987). Language proficiency, writing ability, and composing strategies: A study of ESL college student writers. *Language Learning, 37*(3), 439–468. https://doi.org/10.1111/j.1467-1770.1987.tb00579.x

Révész, A., Kourtali, N., & Mazgutova, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning, 67*, 208–241.

Révész, A., Michel, M., & Lee, M., 2017, Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory. IELTS Research Reports Online Series, 2017/3.

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*, 605–631.

Rijlaarsdam, G., & Van Den Bergh, G. (1996). The dynamic of composing-An agenda for research into an interactive compensatory model of writing: Many questions, some answers. In C. M. Levy, & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 107–126). Lawrence Erlbaum Associates.

Roca de Larios, J., Nicolás-Conesa, F., & Coyle, Y. (2016). Focus on writers: Processes and strategies. In R. Manchón, & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 267–286). DeGruyter.

Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behavior in the allocation of time to writing processes. *Journal of Second Language Writing, 17*, 30–47.

Roca de Larios, J., Murphy, L., & Manchón, R. (1999). The use of restructuring strategies in EFL writing: A study of Spanish learners of English as a Foreign Language. *Journal of Second Language Writing, 8*, 13–44. https://doi.org/10.1016/S1060-3743(99)80111-8

Sasaki, M. (2002). Building an empirically-based model of EFL learners' writing processes. In G. Rijlaarsdam (series) (Volume eds.). In S. Ransdell, & M. Barbier (Eds.), *New directions for research in L2 writing* (pp. 49–80). Kluwer (Volume eds.).

Schoonen, R., Snellings, P., Stevenson, M., & Van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 77–101). Multilingual Matters.

Spelman Miller, K., 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*, 433–453. https://doi.org/10.1002/j.1545-7249.2008.tb00140.x

Stevenson, M., Schoonen, R., & De Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing, 15*, 201–233. https://doi.org/10.1016/j.jslw.2006.06.002

Tillema, M. (2012). Writing in first and second language. Empirical studies on text quality and writing processes [Unpublished doctoral thesis]. Netherlands Graduate School of Linguistics (LOT).

Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition: An International Journal, 38*, 79–95. https://doi.org/10.1016/j.compcom.2015.09.012

Xu, C., & Ding, Y. (2014). An exploratory study of pauses in computer-assisted EFL writing. *Language Learning and Technology, 18*, 80–96.