Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study

Marije Michel[a], Andrea Révész[b], Xiaojun Lu[b], Nektaria Kourtali[c], MinJin Lee[d], and Lais Borges[b]

[a] Groningen University, NL, and Lancaster University, UK

[b] University College London, UK

[c] University of Liverpool, UK

[d] Yonsei University, South Korea

Corresponding author

Address: Dr Marije Michel, Groningen University, Dept. Applied Linguistics,
Harmoniecomplex, Oude Kijk in 't Jatstraat 26, 9712 EK Groningen,
The Netherlands

Email: m.c.michel@rug.nl

Acknowledgements

We would like to thank the Educational Testing Service (ETS) for providing financial support to carry out the project on which this study is based. We are also grateful to ETS, and John Norris in particular, for constructive feedback and practical support at various stages of the project. In addition, we would like to thank Khaled Barkaoui for sharing his typing test and Ana Pellicer-Sanchez for advice on eye-tracking issues.

Funding

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).

Abstract

Most research into second language (L2) writing has focused on the products of writing tasks; much less empirical work has examined the behaviours in which L2 writers engage and the cognitive processes that underlie writing behaviours. We aimed to fill this gap by investigating the extent to which writing speed fluency, pausing, eye-gaze behaviours and the cognitive processes associated with pausing may vary across independent and integrated tasks throughout the whole, and at five different stages, of the writing process. Sixty L2 writers performed two independent and two integrated TOEFL iBT writing tasks counterbalanced across participants. While writing, we logged participants' keystrokes and captured their eye-movements. Participants took part in a stimulated recall interview based on the last task they had completed. Mixed effects regressions and qualitative analyses revealed that, apart from source use on the integrated task, L2 writers engaged in similar writing behaviours and cognitive processes during the independent and integrated tasks. The integrated task, however, elicited more dynamic and varied behaviours and cognitive processes across writing stages. Adopting a mixed-methods approach enabled us to gain more complete and specific insights than using a single method.

Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study

Most existing research into second language (L2) writing performance, development and assessment has focused on the product of writing (see Cumming, 2016, Polio and Lee, 2017), with researchers relying on increasingly more sophisticated analytical tools to study text quality including corpus-based, natural language processing techniques (e.g., Alexopoulou, Michel, Murakami and Meurers, 2017). Less is known about the L2 writing processes in which L2 writers engage (e.g., Roca de Larios, Murphy and Manchón, 1999; Spelman Miller, 2000; Stevenson, Schoonen and De Glopper, 2006), although there is a growing body of research on L2 writing behaviours (e.g., fluency, pausing) and associated cognitive processes using more and more advanced research methodology (Révész and Michel, 2019). We conducted this study to contribute to and expand this work. Through the triangulation of keystroke logging, eye-tracking, and stimulated recall, we investigated the extent to which writing behaviours (speed fluency, pausing, and eye-movements) and cognitive writing processes associated with pausing (inferred from stimulated recall protocols) may vary across independent and integrated tasks throughout the whole, and at different stages, of the writing process. Few studies have looked into the effects of task type on L2 writing behaviours (Barkaoui, 2015) and cognitive processes (Plakans, 2008), and none have employed a mixed-methods approach combining keystroke logging and verbal protocols with eye-tracking.

Background

Theoretical background

We have adopted Kellogg's (1996) model of writing as a theoretical framework. Like most existing theoretical accounts of writing (e.g., Flower and Hayes, 1980; Galbraith, 2009; Hayes, 2012; Scardamalia and Bereiter, 1987), this model was developed to explain first language writing. Compared to other writing models, however, Kellogg’s framework puts greater emphasis on linguistic encoding processes. This makes it particularly suitable for studying L2 writing, given that encoding ideas into written form tends to require more conscious attention and effort in one’s second than first language **(**Kormos, 2012; Roca de Larios et al., 1999).

Kellogg's (1996) model distinguishes three main processes: formulation, execution, and monitoring. Formulation involves higher-order processes such as planning content, retrieving ideas from the task input and/or from long-term memory, and organizing ideas into a coherent plan. In addition, formulation entails the lower-order translation processes of lexical retrieval, syntactic encoding, and creating cohesion, which translate the writer's plan into language form. At the execution stage, the writer uses motor movements to hand-write or type their text. During monitoring, writers ensure that the written text maps onto their intended plan and, if needed, revision is triggered. Importantly, the processes of formulation, execution, and monitoring are assumed to take place concurrently and cyclically until the text expresses the writer's plan.

Researching L2 writing behaviours and cognitive processes

To test this and other models of writing, researchers of L2 writing processes have employed verbal protocols such as the think-aloud and stimulated-recall procedures (e.g., Roca de Larios, Manchón, Murphy and Marín, 2008). Although these techniques have generated useful information, the validity of the think-aloud procedure has been questioned on the grounds of causing reactivity (i.e., altering the writing process), and both verbalization techniques have been argued to carry the risk of veridicality (i.e., not capturing writers' thoughts in full). To address these limitations, scholars increasingly rely on keystroke logging, alone or together with verbal protocols, to study L2 writing processes (Barkaoui, 2019; Révész, Kourtali, and Mazgutova, 2017; Spelman Miller, 2000; Stevenson, Schoonen, and De Glopper, 2006; van Waes and Leijten, 2015). However, even when combined with verbal reports, keystroke logging does not give insight into what writers look at when they compose. This shortcoming may be addressed through triangulating keystroke logging and verbal report data with eye-gaze recordings. As argued below, adding information about, for example, reading during writing, will provide a more complete picture of the processes underlying writing (Révész and Michel, 2019).

Only a few L2 studies have employed eye-gaze measurements to tap into looking behaviours during writing, with all of them triangulating eye-gaze recordings with other techniques (Chukharev-Hudilainen, Feng, Saricaoğlu and Torrance, 2019; Gánem-Gutiérrez and Gilmore, 2018; Révész, Michel and Lee, 2017a, 2019). The quantitative measurement of real-time written production and/or eye-gaze recordings, especially when combined with the qualitative examination of thought processes, should provide a fuller and more specific description of the behaviours and cognitive processes of L2 writers.

To date, only a few L2 writing studies have adopted a mixed-method approach. For example, Révész et al. (2017b) utilised keystroke logging and stimulated recall to examine the speed fluency, pausing, and revision behaviours of L2 writers and associated cognitive processes during an argumentative task with or without content support. Triangulating data from the two methodologies, the researchers concluded that content support likely decreased the pressure on planning processes (e.g., inducing fewer between-sentence pauses and more below-clause level revisions), thereby freeing up attentional resources for linguistic encoding (i.e., more translation-related pauses and revisions). Khuder and Harwood (2015) combined keystroke logging, stimulated recall, and screen recordings to compare L2 writers performing tasks under a test versus no test condition. Findings revealed more translation and surface revision processes under the test condition, but higher proportions of meaning-focused revisions and evaluation in the non-test situation, with differences being more pronounced at the last writing stage.

Recently, researchers have also begun to include eye-tracking methodology in mixed-methods studies when investigating L2 writing processes. For instance, Gánem-Gutiérrez and Gilmore (2018) complemented digital screen capture data with eye tracking, video recording, and stimulated recall when studying Japanese L2 English writers. Qualitative analyses, which considered the number and frequency of writing activities (e.g., rereading, use of external sources), revealed that most of the writing time was dedicated to text construction, while other activities took up comparatively little time. Analyses across writing stages additionally found that, as participants progressed with the task, they spent gradually less time on text construction and increasingly more time rereading their work. Révész, Michel and Lee (2017a, 2019) combined eye-tracking with stimulated recall and keystroke logging to examine writing processes of Chinese L2 users of English completing an argumentative essay. Like the present study, one aim was to investigate pausing behaviours and associated cognitive processes. The researchers obtained measures of pause frequency and length, classified according to location—whether they occurred within words, between words, or between sentences. Of interest was whether eye-gazes remained at inscription point or moved back within the word/phrase, clause, sentence, or paragraph preceding the inscription point. As hypothesised, when participants paused between sentences, pauses were longer, they looked back at longer stretches of text, and they engaged in higher-order writing activities. Pauses within and between words were shorter, induced shorter lookbacks, and involved lower-order writing processes.

Overall, findings of these mixed-methods studies mirror trends in previous writing research, where longer pauses were associated with higher textual units (between clause and sentences), and shorter pauses were linked to lower textual units (within and between words) (e.g., Deane and Zhang, 2015; Spelman Miller, 2006; Spelman Miller, Lindgren and Sullivan, 2008; Xu and Qi, 2017). However, the combination of data sources provided a fuller picture of writing behaviours and more complete understanding of the underlying processes. Inspired by this earlier work and methodological advances in L2 writing-process research, the present study also adopted a mixed-methods approach employing keystroke logging, eye-tracking, and stimulated recall. Specifically, we investigated L2 writing processes across independent versus integrated tasks, hoping that the triangulation of methods would afford deeper insights into writing processes across these task types.

The role of task type: Independent versus integrated tasks

It is well documented in L2 writing research that the type of task in which learners engage has an impact on the writing product (e.g., Alexopoulou et al., 2017; Lu, 2011). Given this understanding, researchers have shown a keen interest in exploring how writing performances might be affected by the distinction between independent and integrated tasks, the latter task type being employed increasingly in language assessment contexts in an attempt to increase authenticity. Independent tasks typically ask writers to address a prompt or respond to a question relying on their own resources. Integrated tasks “require learners or test takers to incorporate substantive content from source materials" (Cumming 2013, p.1), thus writers need to synthesize information from, for example, a listening and/or reading input and summarize it into a coherent text. While in the assessment literature there is a growing body of research comparing L2 performances on independent versus integrated writing tasks, most existing work has examined the extent to which the product of writing or text quality may vary across the task types (e.g., Biber and Gray, 2013; David, 2015). To the best of our knowledge, only three studies have looked into writing processes and behaviours on integrated vs. independent tasks, all examining the TOEFL iBT test.

Based on Kellogg’s model (1996), one might expect integrated tasks to engage writers in planning and translation processes to a lesser degree, because writers can rely on the sources for help with content and language. This, in turn, might elicit greater fluency and more time for monitoring. Previous studies largely reflect these predictions. Using a think-aloud procedure, Plakans (2008) found that in the TOEFL iBT integrated task, students were more likely to reread the prompt, engage in thinking to interpret the task, and do during-writing planning. In contrast, the independent task elicited more initial but less during-writing planning and more frequent rereading. Unlike Plakans, Barkaoui (2015) employed stimulated recalls, and revealed that, during the independent task, participants planned more, experienced greater difficulty with generating content and language, and revised language more frequently. In a keystroke-logging study, Barkaoui (2019), like the present research, examined the impact of task type on pausing behaviours. When working on the integrated task, participants paused longer on average, probably because they went back to the reading while composing. They also paused longer between paragraphs on the independent task, suggesting more time was required for planning content. Finally, writers produced more revision pauses on the independent task, maybe because they could not extract language from a provided text. Through the use of a mixed-methods approach including eye-tracking, we aimed to substantiate and expand this research.

Stage of writing

Largely motivated by the work of Rijlaarsdam and Van den Bergh (1996), a growing number of studies examined the temporal distribution of writing activities, demonstrating that the behaviours and the cognitive processes in which writers engage differ during the composing process. Previous work, mostly focusing on independent, argumentative writing tasks, suggests that L2 writers tend to plan more initially, while formulation activities are more frequent in the middle phases (e.g., Barkaoui, 2015; Roca de Larios et al., 2008; Tillema, 2012; Van Weijen, 2009). Less uniform patterns were observed for revision and rereading. Some studies found increased revision over time (Barkaoui, 2015; Roca de Larios et al., 2008), whereas others reported stable amounts of revision across stages (Gánem-Gutiérrez and Gillmore, 2018; Tillema, 2012). For rereading, Tillema (2012) observed similar amounts throughout the writing process, but in Gánem-Gutiérrez and Gillmore (2018) there was a decrease during the task. In addition, Roca de Larios et al. (2008) found that more proficient writers showed greater variety of activities and were more versatile in responding to the different demands of the evolving text, whereas lower level writers exhibited similar processes and behaviour across stages.

Only a few studies have considered the distribution of writing activities in integrated writing tasks. In comparing independent and integrated tasks, Barkaoui's previously discussed (2015) study found that the independent task initially elicited reflection, followed by planning and text generation, while the final stage was characterised by evaluation and revision. On the integrated task, participants interacted with the sources in the first stage, but the other activities were largely parallel to those observed for the independent task. Leijten et al. (2019) revealed similar trends for source-based writing: writers spent most of the first interval consulting the various source texts, followed by an intensive writing period involving only short switches to the sources. In the final stages, hardly any sources were consulted; high achievers engaged in revising their texts. Finally, Barkaoui (2019) observed longer pauses in the first third of the writing process compared to the second and last stage irrespective of task type. However, the average frequency and length of pausing differed across stages and task type. For the first interval, there were more but shorter pauses in the independent than the integrated task. Also, pause frequency was almost equal across the three stages on the independent task, but pauses were three times as frequent in the second and third stages as in the first on the integrated task. Barkaoui attributed this difference to the participants' rereading of the source text during the initial stage of the integrated task and look-backs to the source at later stages. One aim of this study was to gain direct evidence about eye-gaze behaviours during writing to reach firmer conclusions about processes such as rereading.

Study

Through a mixed-methods approach employing keystroke logging, eye-tracking, and verbal protocols, we pursued the following research questions and sub-questions:

1. To what extent does task type influence the behaviours of L2 writers
    a. during the whole writing process?
    b. at various stages of the writing process?

2. To what extent does task type influence the cognitive processes underlying writing behaviours
    c. during the whole writing process?
    d. at various stages of the writing process?

Task type was operationalised as differences between independent and integrated writing tasks. Writing behaviours were measured in terms of measures of speed fluency and pausing obtained through keystroke logging and indices derived from eye-gaze recordings. Underlying cognitive processes were investigated by eliciting stimulated recall comments on participants' composing processes.

Methodology

Participants

Participants were 60 L2 users, 20 students each at levels B1, B2, and C1 of the Common European Framework of Reference (CEFR). They were all Chinese L2 users of English, studying at the University of London. We recruited an initial pool of 103 participants. Of these, 84 students were invited to participate based on their performance on the research form of the TOEFL iBT listening and reading tests and a typing test. From among these students, 24 were excluded due to technical issues or because they failed to complete all tasks. The final cohort were mostly females (*n* = 55), aged 18 to 36 (*M*=23.76, *SD*=3.22). The majority were studying for an MA (*n* = 55), and two were working towards a BA and three towards a doctorate.

Instruments and procedures

Typing test

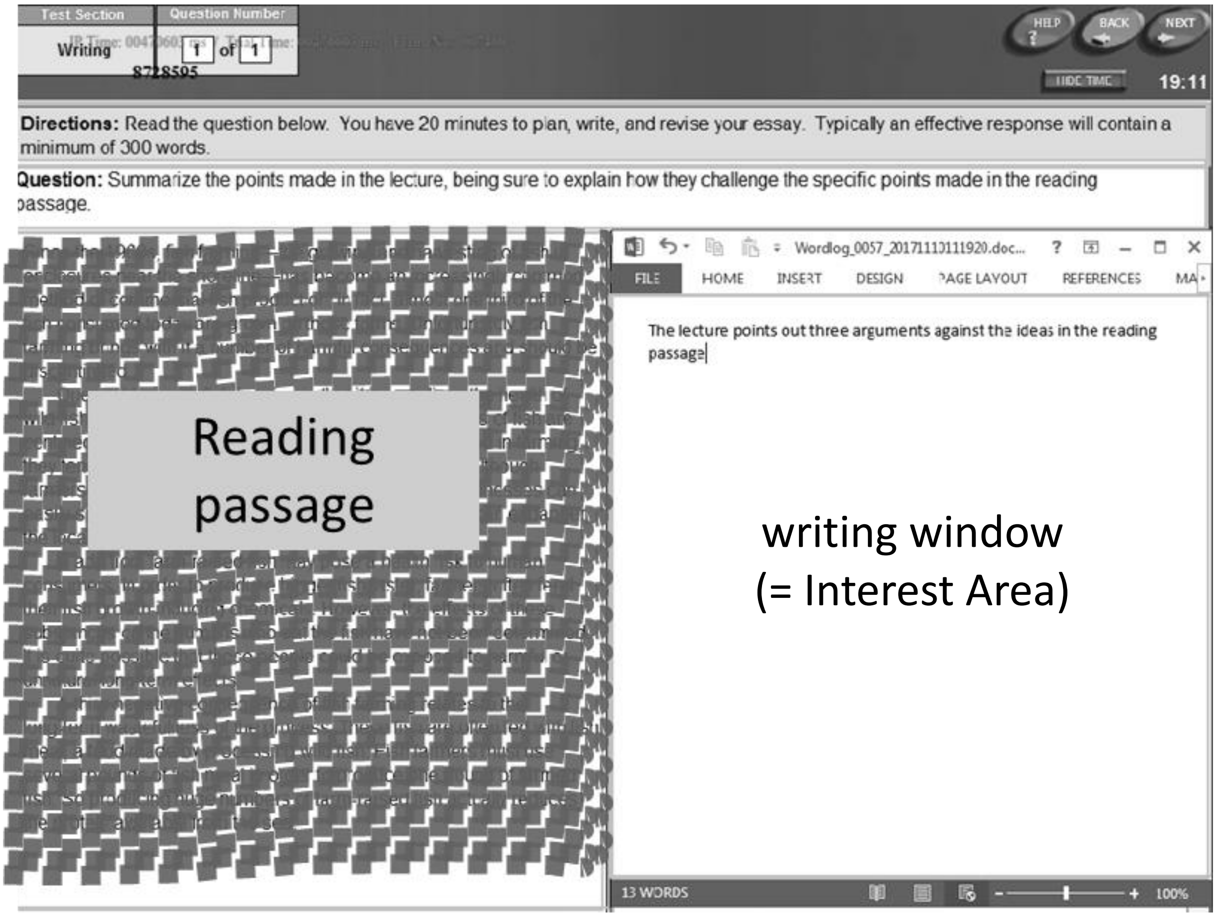
We controlled for keyboarding skills using the software, Typing Test Pro (Barkaoui 2014). Per proficiency level, all individual scores of net typing speed (i.e., words per minute adjusted for accuracy) were within 2 SDs from the mean per group (B1: M=25.32, SD=7.52; B2: M=26.60, SD=11.71; C1: M=36.55, SD=10.90).

Writing tasks

Participants completed two research versions of the TOEFL iBT independent and integrated writing tasks to control for prompt effects, resulting in 240 performances altogether. The order of the four tasks was counterbalanced across participants. The independent tasks asked participants to write an argumentative essay in 30 minutes. For the integrated task, participants first read a passage, then listened to a lecture on the same topic. While reading and listening, they could take notes on paper. Next, their task was to summarise the points made in the lecture, explaining how the named aspects in the lecture cast doubt on the points that were put forward in the reading passage. The reading text and any notes taken were available in the 20 minutes participants had for writing their summaries.

The tasks were administered in the TOEFL iBT research platform, without additional planning time. The actual writing, however, was completed in a Microsoft (MS) Word document, as the Inputlog software logged data in MS Word. The MS Word document was opened on top of the TOEFL iBT environment and set up in such a way that the font type, font size, spacing and editing tools mimicked the original TOEFL iBT writing window (see Figure

1 for an example of the set-up of the integrated task; in the independent task, the reading passage area was blank). While writing, an Eyelink1000 with a temporal resolution of 1000 Hz recorded participants' eye movements.



Directions: Read the question below. You have 20 minutes to plan, write, and revise your essay. Typically an effective response will contain a minimum of 300 words.

Question: Summarize the points made in the lecture, being sure to explain how they challenge the specific points made in the reading passage.

The lecture points out three arguments against the ideas in the reading passage

Figure 1. Screen set-up for integrated task

Stimulated recall

Immediately after completing the writing tasks, all participants engaged in a stimulated recall session to elicit the participants’ thoughts about the writing. Recall was based on the last writing task they had performed, that is, we collected these data for 30 independent and 30 integrated performances. To prompt recall, we used the recordings of participants’ keystrokes and eye-movements while writing. We explained to participants how to interpret eye-gaze recording data in everyday language, for example, that the red circles (eye-fixations) and lines (saccades) in the eye-gaze recordings indicated their eye movements, and that larger circles represented longer eye-fixations. Participants also listened to a sample stimulated-recall performance based on a different writing task. Participants could stop the recording at any point when they wanted to share their during-writing thoughts. Additionally, the researcher elicited their thoughts when they paused, revised, or produced unexpected or interesting eye-movements (e.g., longer fixations, regressions) but did not produce a comment. We stressed to only report on what they were thinking during task completion. Participants could use their native language, given that the third author, a Mandarin speaker, conducted all stimulated recall sessions. We video-recorded all sessions to capture participants' verbal comments as well as gestures (e.g., pointing to the screen). The sessions lasted approximately 60-90 minutes.

Data collection

First, participants took part in a group session in a computer lab. After providing informed consent, they completed a background questionnaire (10min), the listening (60-90min) and reading (60-80min) components of the TOEFL iBT test, and a typing test (10min). Those who obtained appropriate proficiency and typing scores attended two individual sessions. In the first session, they completed the first two tasks (60-70min), in the second session, participants performed the remaining writing tasks (60-70min), followed by the stimulated recall session (30-45min).

Before participants began a writing task, we launched the Inputlog software and the eye-tracker. Participants sat about 60cms away from the centre of the screen. Once calibration on a 9-point grid was successful, we started the *SR Research Screen Recorder* software and opened the appropriate version of the research version of the TOEFL iBT writing task. To ensure ecological validity, we used the remote set-up of the eye-tracker, which allowed participants to move their head freely during writing. We recalibrated participant's eyes between each writing task, but no re-calibration was done during writing tasks. We monitored participants’ eye-movements throughout the writing session on the researcher's screen and adjusted the seating of the participant if tracking was lost. To account for track loss, we measured blink duration and number for each participant and calculated the mean percentage of blink duration (see Table S1, online supportive material). There were fewer but on average longer blinks on the integrated than the independent task. Percentage of track loss was around 30% for both tasks, a lower rate than reported in earlier work (e.g., Ganém Gutierrez and Gillmore, 2018). Participants took short breaks between the writing tasks and prior to the stimulated recall interview.

Data analysis

Writing behaviours

We obtained speed fluency and pausing measures from Inputlog 7 (Leijten and van Waes 2013). We used a pause threshold of 200ms when calculating the fluency and pausing indices, as this low threshold allowed us to capture lower-level writing processes (van Waes and Leijten, 2015). Speed fluency was expressed with two measures: characters per P-burst (i.e., number of characters produced between pauses) and mean duration of character production (i.e., total writing time excluding pauses divided by number of characters produced).

We classified the pause frequency and length measures by location - whether pauses occurred within words, between words, or between sentences - while counting between-word pauses as one pause (adding up the pause before and after pressing the spacebar).

Eye-gaze data were analysed with the SR Research Data Viewer software. Given that the TOEFL iBT research environment leaves little white space around words and lines, we used relatively coarse eye-gaze measures to gauge viewing behaviours within the writing window as a whole. That is, for this study, the box allocated for writing was defined as the interest area (AOI). For this AOI, we calculated the following indices (Brunfaut and McCray 2015): fixation count; total fixation duration; mean fixation duration; number of forward and of backward saccades; median length of forward and of backward saccades (in degrees of visual angle); and proportion of regressive movements (i.e., number of backward saccades divided by the total number of saccades). Forward and backward saccades were defined as eye-movements that had a positive (forward) and negative (backward) angle between the horizontal plane and the direction of the current saccade, respectively. We corrected for time on task by dividing the measures by the time needed for task completion for the measures total fixation duration and number of fixations, forward saccades, and backward saccades.

Cognitive processes

The analysis of the stimulated recall comments involved five steps. First, the data were transcribed. Second, the third author reviewed the comments related to pausing and identified emergent categories. Third, the resulting micro-categories were merged into more general categories following Kellogg (1996) (see Table 1). Fourth, the third author coded all the comments. Another Mandarin speaker of L2 English with an L2 research background coded 20 percent of the data, yielding a high inter-coder reliability (Cohen's kappa = .91). Finally, comments were added up resulting in a frequency count per participant by category.

Table 1. Examples for Stimulated Recall Comments by Coding Category

Process/Subprocess	*Example (translated from Mandarin)*
Planning	
Content	I was thinking what to write for the second point. I was trying to find out any point that I could put there.
Organisation	I was thinking I would write two paragraphs to support my view and write another paragraph starting with 'admittedly'.
Translation	
Lexical retrieval	I was thinking what verb to use after 'to'. The words I used were very similar to those in the question.
Syntactic encoding	I was thinking of adding a clause after 'study the subjects' to express 'you are interested in'. [...] I didn't want to phrase it in the same way as the question did.
Cohesion	I was wondering what linking word to use here.
Unspecified	*I wanted to express this point clearly and in detail, but I just could not.*
Monitoring	
Reading for monitoring	I was reading the sentence again to see whether [... it] conveyed what I wanted to express.
Resource use (integrated task only)	
Content	I was thinking of the third point. What was that point about? So I went back to the reading passage.
Organization	I looked at the reading passage again. It was in three paragraphs, so I thought I would write three paragraphs as well.
Lexis	I spelt 'infection' in my mind, but I was afraid that I could not get it right. So I went back to the reading passage to look for the word.
Syntax	There was no plural 's' in my notes, but then I thought the observations could be multiple. So I might need the 's', and then I was hesitant about whether to add it or not.
Monitoring	I just wanted to check if the second paragraph was complete and if there was anything else to add.
Other	I was looking at my notes.

Stages of writing

All analyses were conducted for the overall writing process. Following earlier research (e.g., Tillema et al. 2011), we also divided the total time participants spent writing each task into five equal intervals and calculated all indices (keystroke logging, eye-gaze, and stimulated recall) for these intervals. This allowed us to capture potential changes in processes as a function of writing stage within participants and to compare writing processes by stage across participants and tasks.

Hypotheses

Based on Kellogg's (1996) framework, Rijlaarsdam and Van den Bergh's (1996) temporal model of writing and earlier empirical research, we expected that the independent and integrated writing tasks would yield different writing behaviours, both when considered as a whole and when looking at writing stages.

For task type, we hypothesised that the availability of oral and written sources would ease planning and translation processes during the integrated task, resulting in greater writing fluency (i.e., characters per P-burst and mean duration of character production) and fewer and shorter pauses (particularly between higher textual units). We also anticipated fewer and shorter fixations on the integrated task, accompanied by fewer but longer forward and backward saccades, given that participants were expected to return to the source text and listening notes while writing. A higher proportion of backward saccades was expected during the integrated task, indicating more rereading (a signal of monitoring). Stimulated recall comments were hypothesised to be aligned with these prognoses.

Concerning different writing stages, we hypothesized that differences due to source use would be most pronounced at the initial stages, while later stages were anticipated to yield more similar behaviours across the task types, demonstrating focused writing in the middle stages (e.g., fewer/shorter pauses, fixations and saccades) and monitoring (e.g., longer pauses, longer saccades, higher proportion of backward saccades) in the last stage.

Statistical analyses

According to G*Power (Faul, Erdfelder, Lang and Buchner 2007), a sample size of 60 allowed us to identify medium-size relationships, given the within-subject design and number of observations. To address the research questions, we constructed linear mixed effects models using the *lmer* function of the *lme4* package in the R statistical environment. The *r.squared GLMM* function in the *MuMln* package was used to compute effect sizes (R^2) for fixed effects, and Cohen's d was employed to obtain effect sizes for Tukey post-hoc tests. Following Plonsky and Oswald (2014), *d* values of .60, 1.00 and 1.40 were considered as small, medium, and large. The alpha level was set at .05 for initial analyses and at .01 for any post-hoc tests. Residual plots were used to check the linearity, homoscedasticity, and normality assumptions for the models; the data met the assumptions.

Results

Task type and L2 writing behaviours

Research question 1a investigated the extent to which task type influenced writing behaviours during the whole writing process. In all statistical models, a writing behaviour index served as the dependent variable, the fixed effect was task type, and participant and prompt were the random effects. We also added by-participant random slopes for task type to account for the potentially differential effects of task type on the participants.

The descriptive statistics for the measures of speed fluency, pausing, eye-fixations, and (forward and backward) saccades are summarized in Tables 2 and 3, respectively (see online material Table S2 to S5 for complete figures), while Table 4 provides relevant inferential statistics.

Table 2. Fluency and Pausing Measures by Task Type and Pause Location (n=60). Full descriptives (including data by stages) are available as online supportive material S2 and S3.

	Fluency				Pausing				
	Characters per P-burst		Active writing time per character (min)			Pause number per minute		Median pause length	
Task type	M	SD	M	SD	Location	M	SD	M	SD
Independent	1.53	.44	.15	.11	Ww	25.04	11.65	.30	.04
					Bw	14.46	6.18	.78	.31
					Bs	.63	.43	4.46	12.37
					Total	58.98	22.18	.39	.06
Integrated	1.52	.48	.20	.66	Ww	23.78	9.87	.30	.05
					Bw	13.99	5.73	.80	.36
					Bs	.68	.50	4.59	20.75
					Total	56.18	16.20	.40	.08

Note. P-burst=between two pauses; Ww=within word; Bw=between words; Bs=between sentences

Table 3. Eye-fixation and Saccade Measures by Task Type (n=60). Full descriptives (including data by stages) are available as online supportive material S4 and S5.

	Fixation data						Saccade data									
	Total fixation duration (ms)		Fixation Count		Mean fixation length (ms)		Backward Saccades Number		Backward Saccades Median Length		Forward Saccades Number		Forward Saccades Median Length		Proportion of Backward Saccades	
Task type	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Independent	31894.38	8050.13	84.55	24.53	374.43	8.87	51.68	16.35	2.46	0.58	56.35	17.37	2.26	0.41	0.48	0.00
Integrated	25680.93	751.62	71.62	6.58	363.00	1.20	35.31	4.50	2.14	0.30	39.63	3.42	2.02	0.30	0.47	0.01

Table 4. Significant Effects Identified by the Models Examining the Effects of Task Type on Writing Behaviours

Dependent variable	Pred	Est	SE	t	p	R^2m	R^2c
Speed fluency							
Active writing time per chars	Task	.34	.03	10.96	<.01	.39	.75
Pausing							
Median pause length	Task	.01	<.01	3.40	<.01	<.01	.49
Eye-gaze behaviours							
Total fixation duration	Task	-6305.87	1660.84	-3.80	.01	.05	.46
Fixation count	Task	-13.19	4.63	-2.85	.04	.04	.37
Backward sac median length	Task	-0.33	0.10	-3.26	<.01	.03	.41
Number of forward saccades	Task	-16.84	2.80	-6.01	<.01	.14	.41
Forward sac median length	Task	-0.24	0.09	-2.87	.01	.02	.47

** Task = Task type, R^2m = R^2 marginal, R^2c = R^2 conditional, chars = characters, sac = saccade*

Task type was found to have a significant effect on seven indices. Participants showed greater speed fluency, as measured by active writing time per characters, on the independent as compared to the integrated task, with task type accounting for 39% of the variation. The independent task also yielded significantly shorter pauses, but task type explained less than 1% of the variance. Of the eye-tracking indices, participants fixated significantly longer and more often on the writing window during the independent than the integrated task and made more forward saccades and longer forward and backward saccades when completing the independent task. However, the eye-tracking measures only explained 2-14% of the variance.

Research question 1b examined the extent to which task type influenced L2 writing behaviours at various writing stages. In the series of mixed effects analyses the dependent variable was a writing behaviour measure; the fixed effects were task type, stage of writing, and their interaction; and the random effects were participant and prompt. By-participant random slopes for task type and writing stage were also added to take into account participant-by-stage and participant-by-task type variation. For some dependent variables, the participant-by-stage random slope (characters per P-burst, median length of backward saccades) or the participant-by-task slope (total pause number) were removed to ensure model convergence. The predictors of interest were the interactions between task type and writing stage, with significant effects meaning that participants behaved differently in the independent and integrated tasks during a particular writing stage. The analyses yielded a significant interaction effect for 15 measures: characters per P-burst; active writing time per characters; pause length total and between sentences; pause frequency total, within words, between words and between sentences; total fixation duration; fixation count; number of backward and forward saccades;

median length of forward and backward saccades; and proportion of backward saccades (see Tables S9-S13, online supportve material).

To investigate the interaction effects, we ran another series of mixed effects analyses for the independent and integrated tasks separately. This time, writing stage was the single fixed effect in the models, and the random effects remained participant and prompt. By-participant random slopes for writing stage were also added, but these were removed for some dependent variables to achieve convergence (fixation count and number of forward saccades for both independent and integrated tasks; characters per P-burst, median pause length total, and number of backward saccades for independent task only; active writing time per characters, and pause frequency total and between words for integrated task only). As shown in Table 5, Bonferroni post-hoc tests revealed that, overall, stage of writing had a greater influence on writing behaviours during the integrated task, with the analyses yielding considerably more significant differences among writing stages for this task. Notably, although a significant overall interaction effect was identified for median pause length between sentences, median length of forward and backward saccades, and proportion of backward saccades, no significant stage effects emerged in the post-hoc analyses. The significant stage effects are visually represented in Figures 2-6.

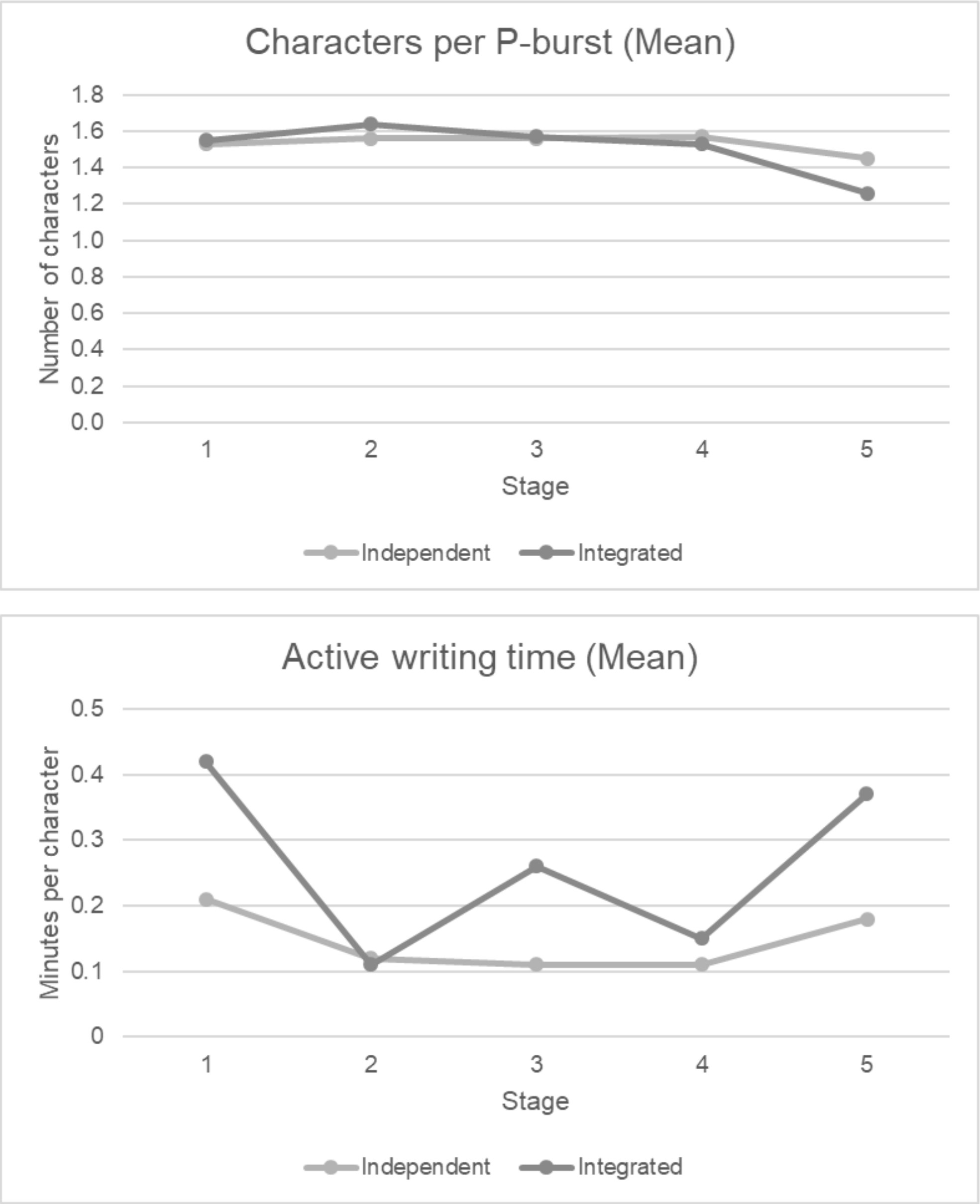


Figure 2. Significant Interaction Effects Identified for Fluency Measures Across Stages

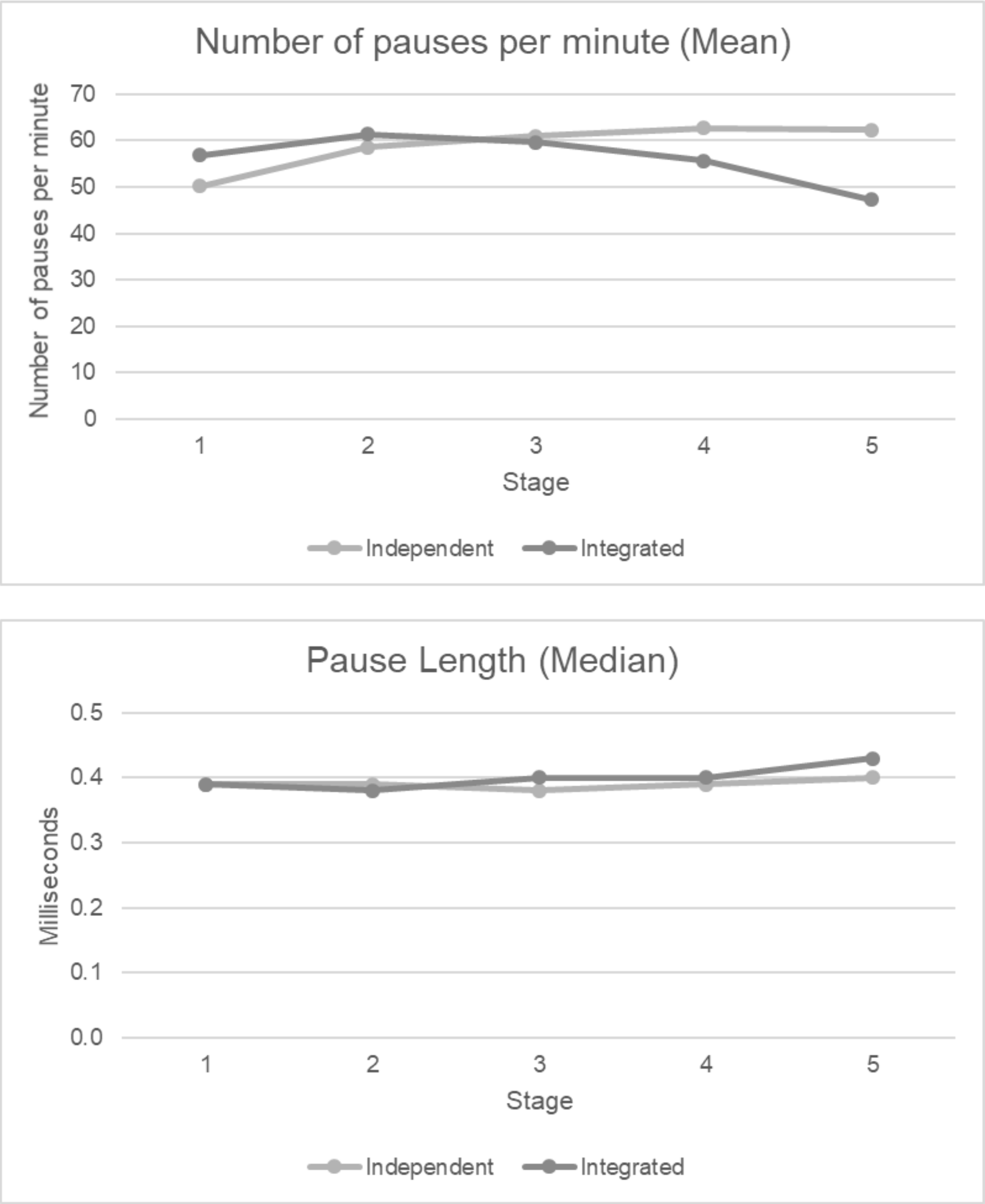


Figure 3. Significant Interaction Effects Identified for Pausing Totals Across Stages

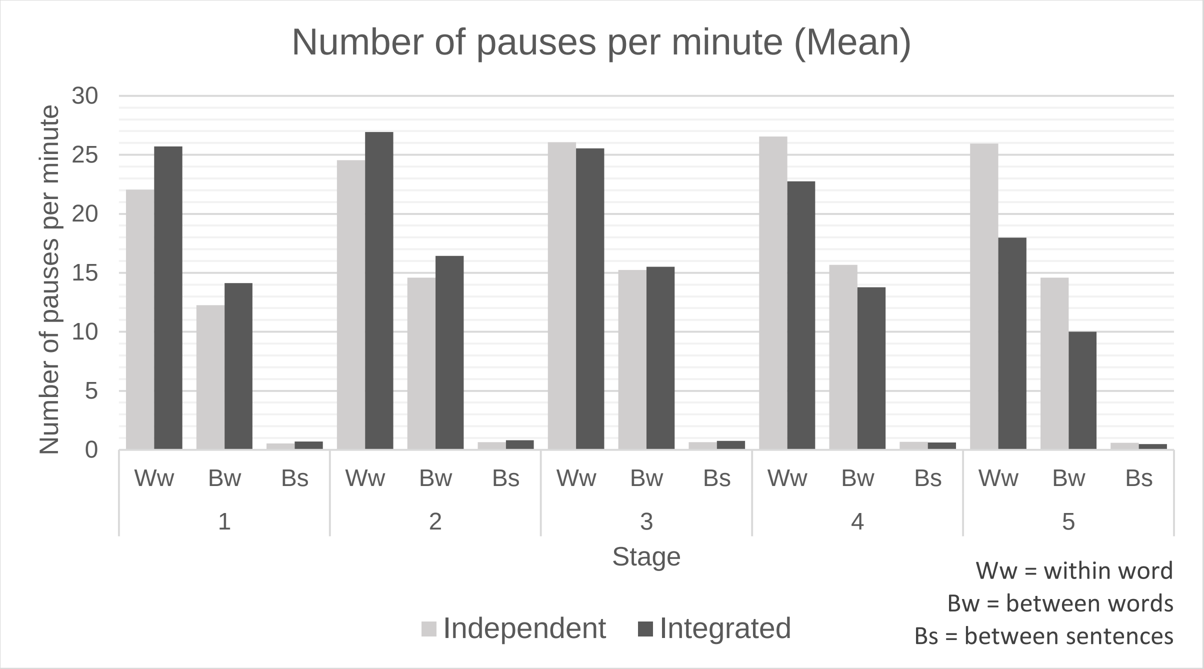


Figure 4. Significant Interaction Effects Identified for Pause Frequency per Location Across Stages.

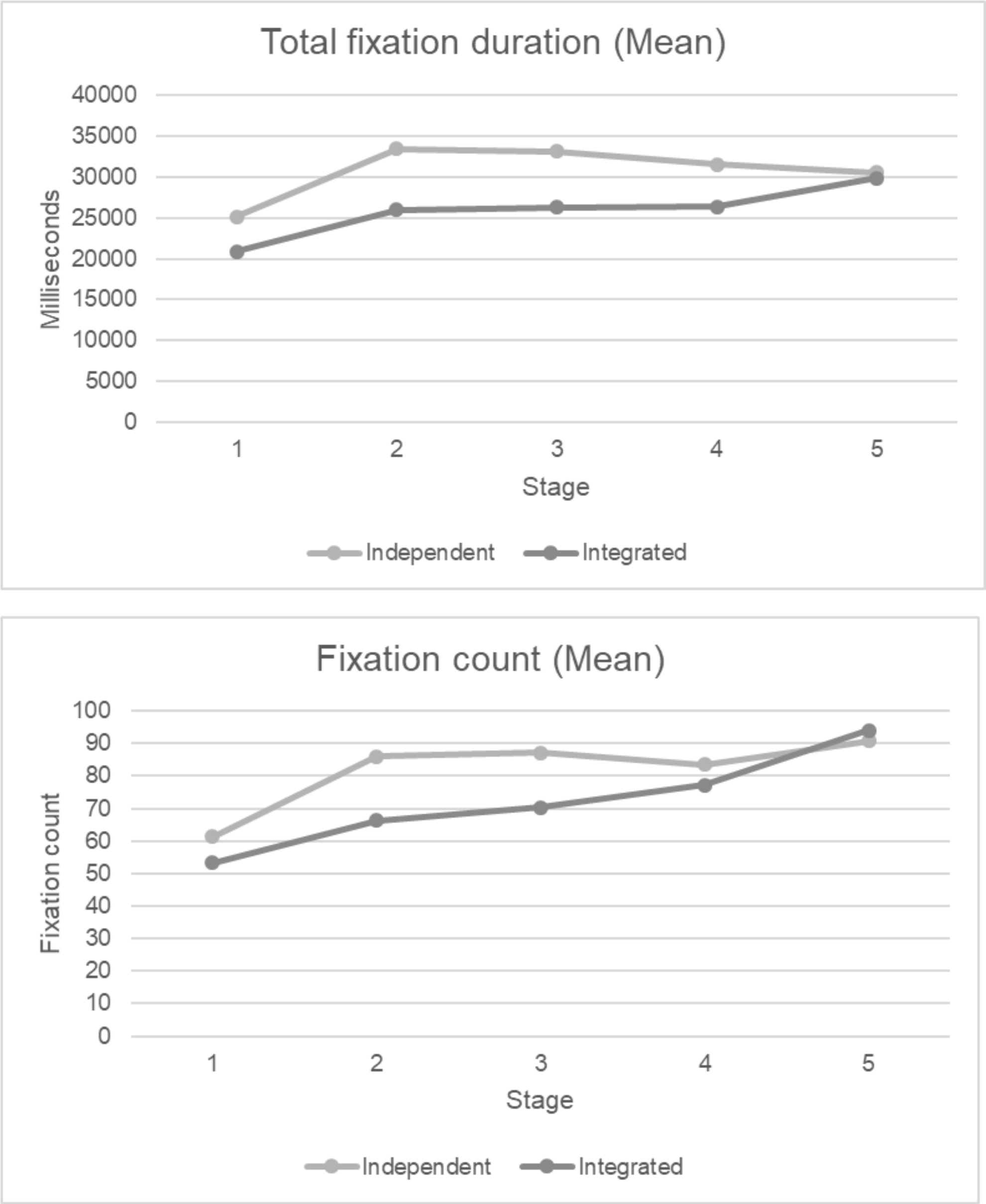


Figure 5. Significant Interaction Effects Identified for Eye-Tracking Indices based on Fixations Across Stages

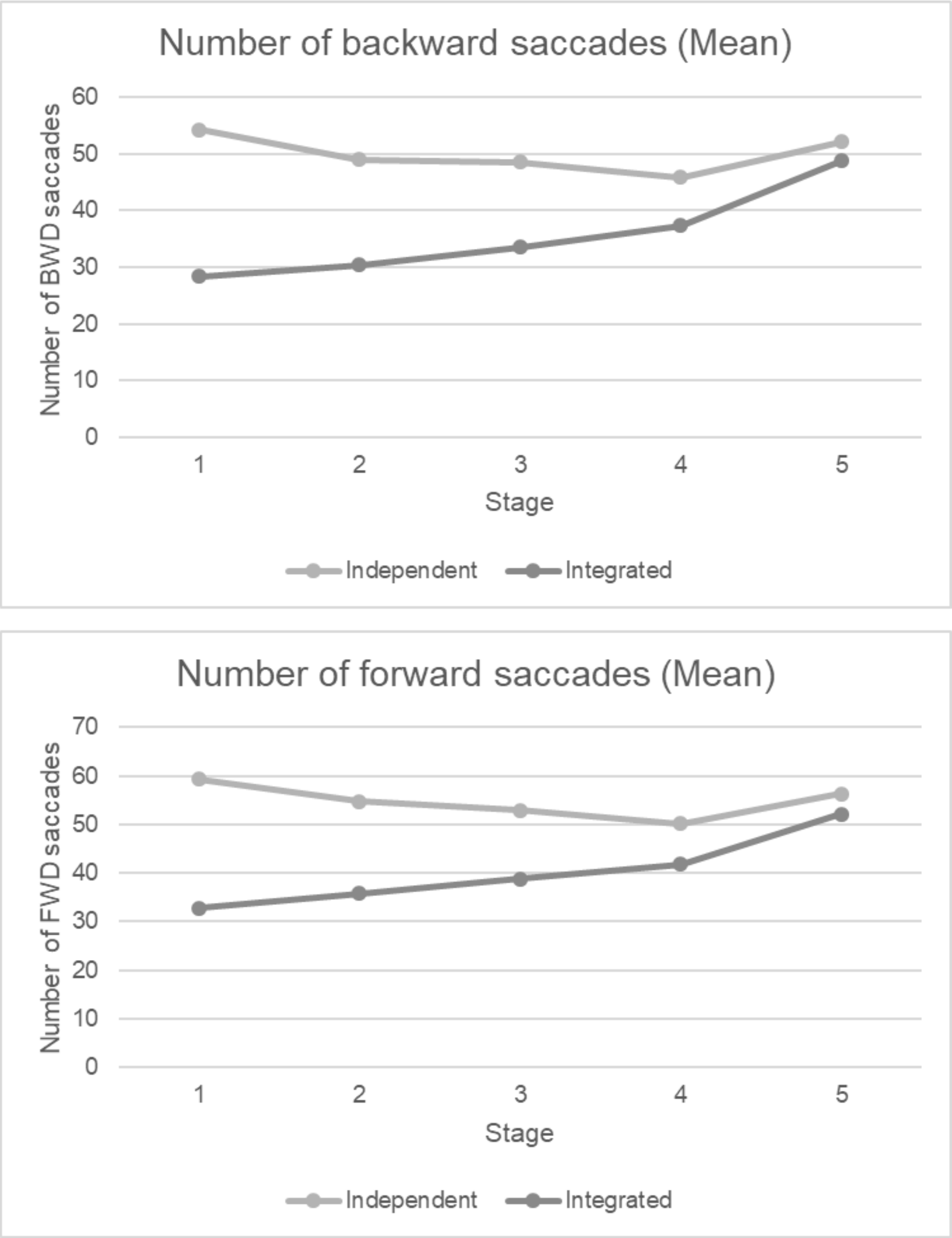


Figure 6. Significant Interaction Effects Identified for Eye-Tracking Indices based on Saccades Across Stages

Most differences on the integrated task were found between stage 1, stage 5, and the rest of the stages. During stage 5, participants produced significantly fewer characters between pauses and paused less frequently, as compared to earlier stages. Exceptions to this trend were the lower number of pauses observed during stage 1 (overall) and stage 4 (overall,) than stage 5. According to the eye-gaze data, participants looked at the writing window more often and longer during stage 5 than earlier stages, as evidenced in higher total fixation time and number of fixations and saccades.

During stage 1, there was more pausing than at stage 5 (overall, within and between words, between sentences), but overall pause length was shorter. In addition, the eye-gaze data revealed that, at stage 1, participants spent less time viewing the writing window than at later stages, reflected in shorter total fixation durations and fewer fixations and saccades. Most effect sizes were in the small range, but a few effect sizes for the eye-gaze indices were large.

On the independent task, speed fluency, as measured by active writing time per characters, was higher during stages 2-4 than during stages 1 and 5, with a small effect size. For pause frequency, most differences were observed between stage 1 and stages 2-4. Participants paused less during stage 1 than subsequent stages, the only exception being pause frequency within words, where more pauses were observed during stage 1 as compared to stage 4. The effect sizes were in the small range. Turning to eye-gaze behaviours, participants fixated shorter and less often on the writing window during stage 1 as compared to later stages, and more saccades were observed during stage 1 than stage 4. The effect sizes were in the small to medium range.

Table 5. Significant Results for Post-hoc Tukey Tests Examining the Effects of Task Type and Stage of Writing on Writing Behaviours (p < 0.01). Full Table S6 including p- and SE-values is available as online supportive material.

	Independent					*Integrated*							
Characters per P-burst													
Post-hoc						S1>S5	S2>S5	S3>S5	S4>S5				
d						.60	.71	.61	.62				
Active writing time per characters													
Post-hoc	S1>S2	S1>S3	S1>S4	S3<S5	S4<S5								
d	.59	.21	.65	.39	.41								
Pause frequency total													
Post-hoc	S1<S2	S1<S3	S1<S4			S1>S5	S2>S5	S3>S5	S4<S5				
d	.70	.81	.79			.42	.65	.61	.45				
Pause length total													
Post-hoc						S1<S5							
d						.39							
Pause frequency within words													
Post-hoc	S1<S3	S1>S4				S1>S5	S2>S4	S2>S5	S3>S5	S4>S5			
d	.61	.57				.62	.47	.76	.69	.46			
Pause frequency between words													
Post-hoc	S1<S2	S1<S3	S1<S4			S1<S2	S1>S5	S2>S4	S2>S5	S3>S5	S4>S5		
d	.62	.71	.66			.63	.06	.47	.80	.69	.60		
Pause frequency between sentences													
Post-hoc						S1>S5	S2>S5	S3>S5					
d						.43	.48	.43					
Total fixation duration													
Post-hoc	S1<S2	S1<S3	S1<S4	S1<S5		S1<S2	S1<S3	S1<S4	S1<S5	S2<S5	S3<S5	S4<S5	
d	1.00	.73	.55	.47		.66	.48	.51	.94	.40	.32	.38	
Fixation count													
Post-hoc	S1<S2	S1<S3	S1<S4	S1<S5		S1<S2	S1<S3	S1<S4	S1<S5	S2<S4	S2<S5	S3<S5	S4<S5
d	1.14	1.00	.78	.89		.76	.57	.72	1.33	.33	.85	.67	.54
Number of backward saccades													

Post-hoc	S1>S4	S1<S4	S1<S5	S2<S4	S2<S5	S3<S5	S4<S5	
d	.46	.45	.95	.42	1.01	.93	.81	
Number of forward saccades								
Post-hoc	S1>S4	S1<S3	S1<S4	S1<S5	S2<S4	S2<S5	S3<S5	S4<S5
d	.52	.32	.42	.95	.41	1.03	.78	.71

Task type and cognitive processes underlying L2 writing behaviours

Research question 2a examined the extent to which task type influenced cognitive writing processes during the whole writing period, as evidenced in the stimulated recall comments prompted by pauses. The comments are summarised in Figure 7 for the independent and integrated tasks, respectively (see online supportive material Tables S7 and S8 for exact statistics). For the independent task, almost half the comments concerned translation processes, about one-third referred to planning, and approximately a sixth described monitoring.

Similarly, on the integrated task, participants referred to translation processes in about half the comments (including resource use), and reported monitoring a sixth of the time. However, they described spending only a fifth of the time planning (including resource use). Overall, about 30 percent of the comments described resource use. Interestingly, resource use was associated with almost 40 percent of the translation-related comments but only a fifth of the planning-related comments.

Research question 2b was concerned with the effects of task type on cognitive processes as a function of writing stage. The distribution of stimulated recall comments (cf. Figure 7) showed some changes across the five stages. On the independent task, planning- and translation-related comments demonstrated a small decrease across the stages, whereas the comments describing monitoring displayed an increase. Likewise, on the integrated task, there was a decrease in comments referring to translation (excluding resource use) and an increasing number of monitoring-related comments. Unlike on the independent task, however, participants reported slightly more planning (excluding resource use) towards later stages. Resource use was mentioned with gradually lower frequency.

The two task types yielded similar trends regarding pause locations. Within-word pauses and between-word pauses were found to be primarily related to translation processes, mostly lexical retrieval. Only when between-word pauses were associated with resource use, participants mentioned planning more than translation. On both task types, between-sentence pauses were mostly linked to planning (mainly content). Patterns for pause location were similar across the five stages.

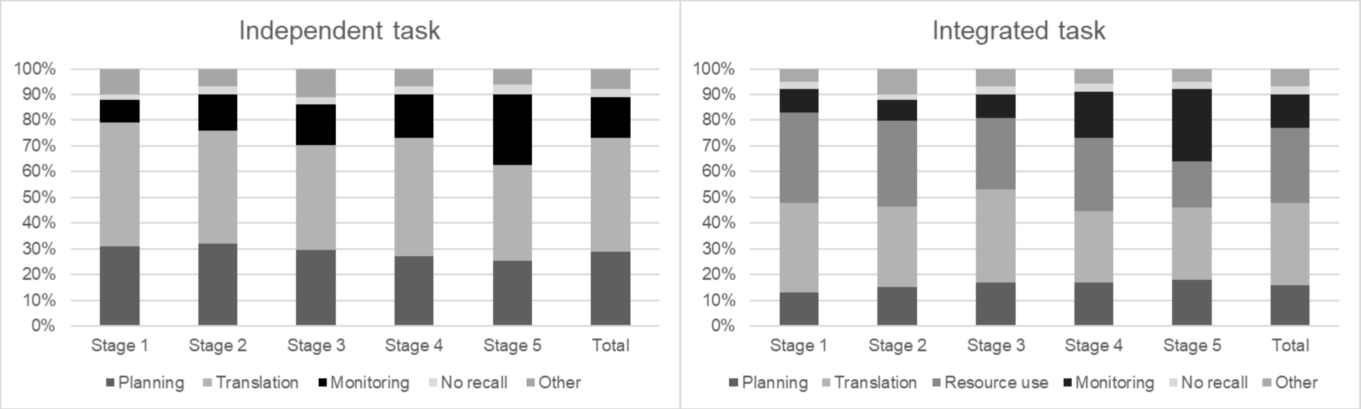


Figure 7. Stimulated Recall Comments Across Stages for the Integrated and Independent Task

Discussion

This study aimed to contribute to and expand on existing work on L2 writing processes (e.g., Révész and Michel, 2019) by investigating how writing behaviours and the cognitive processes underlying them may differ across independent and integrated tasks during the whole, and at different stages, of the writing process. To this end, we used keystroke logging to measure speed fluency and pausing behaviours, eye-tracking methodology to gauge viewing behaviours during writing, and stimulated recall to tap into the cognitive processes of L2 writers.

Writing behaviours and underlying cognitive processes across independent versus integrated tasks

We found that task type had a significant impact on six behavioural indices. Participants took a relatively shorter time to produce characters on average and had shorter pauses in the independent than the integrated task. During the independent task, they made more looks to and spent more time viewing the writing window, and made more and longer forward saccades as well as longer backward saccades. The stimulated recall comments revealed that participants used relatively more pauses for planning (about one-third) in the independent than the integrated task (about one-fifth), and about 30 percent of the pauses on the integrated task were linked to resource use. The three data sources converged on the interpretation that, when working on the integrated task, participants spent proportionately more time viewing the reading text and/or notes that they had taken while listening. This is consistent with the fewer eye-fixations and saccades observed in the writing window and the lower active writing time for the integrated task. In other words, increased source use resulted in decreased time spent on writing. The availability of sources accounts for the reduced time spent planning on the integrated task, given that content could be mined from the reading text and listening notes. No significant task type differences emerged for the other speed fluency, pausing, and eye-gaze measures. Neither did the stimulated recall comments find differences between time spent on translation (about a half) and monitoring (about a sixth) processes across the task types. These results suggest that, apart from using the source text and/or notes in the integrated task, L2 writers engaged in similar writing behaviours during the two task types.

Our results partially replicate previous research findings. Similar to Barkaoui (2015) and Plakans (2008), we observed that the largest proportion of the verbal protocol comments referred to source use on the integrated task, and like Barkaoui (2015), we found more planning-related comments during the independent tasks. Yet, our keystroke logging results run counter to Barkaoui's (2019) findings, where participants displayed differential pausing behaviours during TOEFL iBT independent and integrated tasks. Barkaoui's independent task elicited longer pauses between higher textual units, suggesting that more planning time was required (Schilperoord, 1996).

Although we found no task type effects for the pause measures, it is worth highlighting that, like existing findings (e.g., Révész et al., 2017a, 2019; Spellman Miller, 2000), pause durations were longer between larger (e.g., between sentences) than smaller (e.g., within words) textual units. Also, the stimulated recall comments confirmed the assumption that pauses between smaller and larger textual units tend to be associated with lower- and higher-order writing processes, respectively (Schilperoord, 1996).

Importantly, when interpreting our findings for pause behaviours, we have to consider that we used a 200ms pause threshold to better capture lower-level writing processes (van Waes and Leijten, 2015). This inevitably resulted in smaller average figures for speed fluency (e.g., characters per P-burst: *M*=1.54) compared to earlier work using a 2-second threshold. Indeed, when we apply the 2-second threshold to our data, the mean values (independent: *M*=24.14; integrated: *M*=23.08) are comparable to those in Révész et al. (2017a; *M*=20). The higher speed observed in van Waes and Leijten (2015, *M*=55) may be explained by differences in proficiency and/or L1 background across the studies. A further result of applying a lower pause threshold is that most pauses in our data occurred within words. This contrasts with previous work using a 2-second threshold (e.g., Barkaoui, 2019; Révész et al., 2017a), where pause frequency was highest between words. Again, applying a 2-second threshold would bring our data in line with earlier work.

It is also interesting to consider what insights the eye tracking data offered in addition to the information we gained from the keystroke-logging and stimulated recall data. We learnt from the keystroke-logging data that the integrated task led to slower writing and longer pauses, and the stimulated recall data revealed that, on the integrated task, about a third of the pauses were associated with resource use and fewer pauses were linked to planning. On the one hand, the eye-tracking data, an objective measure, substantiated some of the patterns emerging from the subjective, stimulated recall comments. The fewer visits and less time spent viewing the writing window during the integrated task is compatible with the stimulated recall finding that participants often consulted sources (outside the writing window) when they stopped writing. On the other hand, the eye-tracking data yielded additional information that could not have been derived from keystroke-logging or stimulated recall alone. The fact that the independent task generated more planning-related comments as well as more and longer saccades supports the interpretation that participants might have reread previously produced text for the purpose of generating new content.

The little eye-tracking research that exists on L2 writing processes has used different approaches to analysing eye-gaze behaviours during writing (e.g., Chukharev-Hudilainen et al., 2019; Gánem-Gutiérrez and Gilmore, 2018; Révész et al., 2017a, 2019), thus our findings cannot be directly related to them. It is worthwhile, however, to compare our results to those of L2 reading research. Interestingly, we observed considerably higher proportion of backward saccades (independent: *M*=.48; integrated: *M*=.47) than Brunfaut and McCray (*M*=.28), which supports the view that, compared to regular reading, rereading during writing serves different functions, such as generating new content, managing cohesion, and applying metacognitive revision strategies (Wengelin et al. 2009).

The role of writing stage across independent and integrated tasks

The analyses tapping into the role of writing stage revealed that behaviours were considerably more varied during the integrated than the independent tasks (51 vs. 23 significant differences). Most differences set apart stage 1 and/or stage 5 from the middle stages.

For the independent task, the initial stage was characterised by slower writing, fewer pauses, and shorter and fewer fixations in the writing window than the middle stages. Stages 3 and 4 demonstrated faster writing speed as compared to stage 5, and fewer saccades were observed in stage 4 than stage 1. According to the stimulated recall data (elicited in relation to pausing), both planning- and translation-related comments demonstrated a small decrease over time, whereas comments on monitoring increased. Taking the behavioural and verbal protocol data together, participants at stage 1 were probably more in a 'planning mode' than during the middle stages, which resulted in less continuous writing than in later periods. Greater speed at stage 4 might reflect that participants were trying to finish the task and thus focus on text production. This interpretation is also compatible with the fewer saccades observed at Stage 4. As participants were concerned with text production, they were less likely to reread previous texts, resulting in their eye-gazes remaining more often at the inscription point. Then, at stage 5 they likely slowed down to focus on rereading and monitoring. These results are consistent with Barkaoui (2019), who observed fewer pauses initially on a TOEFL iBT independent task, and others, who found that planning decreased from the initial to later writing stages (Barkaoui 2015; Roca de Larios et al. 2008; Tillema 2012; Van Weijen 2009).

The integrated task reveals a somewhat more varied picture. According to the eye-gaze data, participants spent less time viewing the writing window at stage 1 than at later stages. Pause frequency, at all locations, was highest at stages 2 and 3. At stage 5, participants viewed the writing window more often than previously, reflected in higher total fixation and counts; and moved more within the text, evidenced in greater number of forward and backward saccades. The stimulated recall comments revealed a decrease in translation and resource use but an increase in planning and monitoring at later stages.

Triangulating these findings we may infer that, as expected, during stage 1 participants focused on reading the source text and/or notes. In stages 2 to 4 they primarily engaged in text construction involving both higher- and lower-order writing processes. At stage 5, participants allocated most of their attention to their text, probably to monitor and revise their summaries to ensure that it reflected their intended content. These patterns are well-aligned with previous research on integrated tasks, which reports greater source use at initial stages of writing, increasing attention to own text construction in middle stages, and a primary focus on revision in the last stage (Barkaoui, 2015, 2019; Leijten et al., 2019). The addition of eye-gaze data helped us confirm an initial visual focus on the source text and greater visual engagement with participants' own texts towards the end of writing.

Limitations and directions for future work

This study has a number of limitations. Our participants were London-based Chinese university students, which affects generalisability to other populations. Next, participants did not complete the TOEFL iBT test under high-stakes circumstances. Accordingly, our results may not transfer to real testing conditions.

A number of limitations pertain to the eye-tracking methodology. We utilised relatively coarse eye-gaze measures focusing on the whole writing window. Word-level analyses were not possible, given that the design of the TOEFL iBT platform did not provide sufficient white space between words and lines, nor was the font size large enough (cf. Chukharev-Hudilainen et al. 2019; Révész et al. 2017a, 2019). Another artefact of the test setup was that, for the independent task, the space where the reading text was positioned in the integrated task remained empty. Thus, an increased attention to the writing window on the independent task was not unexpected. Furthermore, due to space limitations, we restricted the eye-gaze analyses to the writing window only. In our future work, we aim to report on gaze information for the reading window (integrated task only), directions and question, as well as switching behaviours between these interest areas.

A further issue concerns individual variation in typing style. For writers who are not touch typists, there is considerable track loss when participants look at the keyboard (around 30% on average in our study, including both touch and non-touch typists). A possible solution is to recruit touch typists only, but this would limit generalisability. Alternatively, one may consider the use of eye-tracking glasses. However, this would introduce different limitations (e.g., accuracy, cf. Conklin, Pellicer-Sanchez and Carrol, 2018). For now, we need to report on and consider track loss when interpreting eye-gaze data during writing.

Another methodological issue concerns the 200ms pause threshold we adopted. This lower threshold enabled us to gain information about lower-level processes. As a consequence, however, our results are not directly comparable to much of the existing research on L2 pausing, where a 2 s threshold was employed.

Finally, following seemingly arbitrary criteria (Tillema 2011), we compared writing behaviours and processes across five stages. It would be valuable to consider writing processes according to the stages of planning, translation and monitoring, as observed in the replay of writing or as reported by participants. For our study, this would have resulted in highly individualised data, making a comparison across participants and tasks challenging. We would, however, encourage future work to explore this approach.

Conclusion

We investigated the behaviours and cognitive processes of L2 writers when completing independent and integrated writing tasks. In the integrated task, source use was most prevalent initially, with participants dedicating gradually more attention to the writing pane. Apart from source use, however, L2 writers engaged in similar types of writing behaviours and cognitive processes during the two tasks. However, the distribution of writing activities varied across the different stages in the two task types. The integrated task elicited more dynamic and varied behaviours and cognitive processes. From a practical perspective, these findings help provide evidence of response process validity of the TOEFL iBT writing test. The two task types generated, as intended (Cumming, Kantor, Powers, Santos, and Taylor 2000; Enright and Tyson 2011), partially different writing behaviors and underlying processes, assisting the measurement of different aspects of the construct. Moreover, it is worth noting that the study yielded no evidence for construct irrelevant behaviours. Finally, it is important to highlight that adopting a mixed-methods approach, through the use of keystroke logging, eye-tracking and stimulated recall, enabled us to gain more complete and specific insights than the use of a single method would have made possible.

References

Alexopoulou T, Michel M, Murakami A, Meurers D. (2017) Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1):180-208.

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

Barkaoui, K (2015) *Test takers' writing activities during the TOEFL iBT writing tasks: A stimulated recall study* (Research Report No. RR-15-04). Princeton, NJ: Educational Testing Service.

Barkaoui, K (2019) What can L2 Writers' pausing behaviour tell us about their L2 writing processes? *Studies in Second Language Acquisition*. Epub ahead of print XXX

Biber D and Gray B (2013) Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis (Research Report TOEFL iBT-19). Princeton, NJ: Educational Testing Service.

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

Chukharev-Hudilainen E, Feng HH, Saricaoglu A and Torrance M (2019). Combined deployable keystroke logging and eyetracking for investigating cognitive processes that underlie L2 writing. *Studies in Second Language Acquisition.* Epub ahead of print XXX

Conklin, K., Pellicer-Sánchez, A., & Carrol, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge, UK: Cambridge University Press

Cumming A (2013) Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly 10(1)*:1-8.

Cumming A (2016) Theoretical orientations to L2 writing. In Manchon R and Matsuda PK (eds) *Handbook of second and foreign language writing*. Berlin: Walter de Gruyter, pp. 65–88.

Cumming A, Kantor R, Powers D, Santos T and Taylor C (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series No. 18). Princeton, NJ: Educational Testing Service.

David V (2015) *Impromptu timed-writing and process-based timed-writing exams: Comparing students' performance and investigating students' and raters' perceptions*. Michigan State University, ProQuest Dissertations Publishing.

Deane P and Zhang M (2015) *Exploring the feasibility of using writing process features to assess text production skills* (Research Report No. RR-15-02). Princeton, NJ: Educational Testing Service.

Enright M and Tyson E (2011) Validity evidence supporting the interpretation and use of TOEFL iBT scores (TOEFL iBT Research Insight, Series I, Volume 4). Princeton, NJ: Educational Testing Service.

Faul F, Erdfelder E, Lang AG and Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioural, and biomedical sciences. *Behaviour Research Methods 39:* 175–191.

Flower L and Hayes JR (1980) The cognition of discovery: Defining a rhetorical problem. *College Composition and Communication 31*: 21-32.

Galbraith D (2009) Writing as discovery. *British Journal of Educational Psychology Monograph Series II 6 - Teaching and Learning Writing*: 5-26. doi: 10.1348/978185409X421129

Gánem-Gutiérrez GA and Gilmore A (2018) Tracking the Real-Time Evolution of a Writing Event: Second Language Writers at Different Proficiency Levels. *Language Learning*. doi:10.1111/lang.12280

Hayes JR (1996) A new framework for understanding cognition and affect in writing. In Levy CM and Ransdell S (eds) *The science of writing*. Mahwah, NJ: Erlbaum, pp.1–27.

Kellogg RT (1996) A model of working memory in writing. In Levy, CM and Ransdell, S (eds), *The science of writing: Theories, methods, individual differences and applications.* Hillsdale, NJ: Erlbaum, pp.57–71.

Khuder B and 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:* 390–403

Leijten M and van Waes L (2013) Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. Written Communication *30*: 358–392.

Leijten M, van Waes L, Schrijver I, Bernolet S and Vangehuchten L (2019) Mapping MA-level students' use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition.* Epub ahead of print XXX

Lu X (2011) A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45,* 36–62. doi:10.5054/tq.2011.240859

Plakans L (2008) Comparing composing in writing-only and reading-to-write test tasks. *Assessing Writing 13:* 111–129.

Plonsky L and Oswald FL (2014) How big is "big"? Interpreting effect sizes in L2 research. *Language Learning 64*: 878-912.

Polio C and Lee J (2017) Written language learning. In Loewen S and Sato M (eds) *The Routledge handbook of instructed second language acquisition*. New York: Routledge, pp.299-318.

Révész A, Kourtali NE and Mazgutova D, (2017b) Effects of task complexity on L2 writing behaviours and linguistic complexity. *Language Learning 67(1)*: 208-241.

Révész A and Michel M (2019) Introduction to the special issue. *Studies in Second Language Acquisition.* Epub ahead of print XXX

Révész A, Michel M and Lee MJ (2019) Exploring second language writers' pausing and revision behaviours: A mixed methods study. *Studies in Second Language Acquisition.* Epub ahead of print XXX

Révész A, Michel M and Lee MJ (2017a) Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory. *IELTS Research Report*. London: The British Council.

Rijlaarsdam, G and van den Bergh H (1996) The dynamics of composing—An agenda for research into an interactive compensatory model of writing: Many questions, some answers. In Levy CM and Ransdell S (eds) The science of writing: Theories, methods, individual differences, and applications. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, pp. 107-125.

Roca De Larios J, Manchón R, Murphy L, Marín J (2008) The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing 17(1)*: 30-47.

Roca De Larios J, Murphy L, Manchon 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(1): 13-44.

Scardamalia M and Bereiter C (1987) Knowledge telling and knowledge transforming in written composition. *Advances in Applied Psycholinguistics 2*: 142-175.

Schilperoord J (1996) *It's about time: Temporal aspects of cognitive processes in text production*. Amsterdam: Rodopi.

Spelman Miller K (2000) Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research 4*: 123–148.

Spelman Miller K (2006) Pausing, productivity and the processing of topic in online writing. In Sullivan KPH and Lindgren E (eds) *Computer key-stroke logging: Methods and applications*. Oxford: Elsevier, pp. 131–155.

Spelman Miller K, Lindgren E and Sullivan KPH (2008) The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly 42:* 433–453.

Stevenson M, Schoonen R and 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.

Tillema M (2012) *Writing in first and second language: Empirical studies on text quality and writing processes.* Utrecht (NL): LOT.

Tillema M, Van den Bergh H, Rijlaarsdam G and Sanders T (2011) Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning 6:* 229-253.

van Waes L and 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.

Van Weijen D (2009) *Writing processes, text quality, and tasks effects. Empirical studies in first and second language writing.* Utrecht (NL): LOT.

Wengelin A, Torrance M, Holmqvist K, Simpson S, Galbraith D, Johansson V and Johansson R (2009) Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behaviour Research Methods* 41: 337-351.

Xu C and Qi Y (2017) Analyzing pauses in computer-assisted EFL writing: A computer-keystroke-log perspective. *Journal of Educational Technology and Society 20*(4): 24-34.