Tracking the Real-Time Evolution of a Writing Event: Second Language Writers at Different Proficiency Levels

Gabriela Adela Gánem-Gutiérrez and Alexander Gilmore

University of Essex and University of Tokyo

We would like to thank the editor of *Language Learning*, the anonymous reviewers, and Florence Myles for their constructive comments on earlier versions of this paper. We are deeply grateful to Monika Schmid and Phil Scholfield for their invaluable help with the statistical analysis. A special thank-you to our participants as well as to Michiyo Yamazaki for her help at earlier stages of the project.

Correspondence concerning this article should be addressed to Gabriela Adela Gánem-Gutiérrez, University of Essex, Department of Language and Linguistics, Wivenhoe Park, Colchester, CO4 3SQ, UK. E-mail: aganem@essex.ac.uk

Abstract

The current study focused on emergent processes during real-time second language (L2) writing activity in an English as a foreign language university context, examining differences in these processes across individual capacities. Participants included 22 adult Japanese learners of L2 English and their tutor. The data were collected using digital screen capture and eye-tracking technologies while the learners wrote a 35-minute argumentative essay. Supplementary stimulated retrospective recalls were also conducted to document the learners' and the tutor's reflections on the writing event. Results revealed clear differences in L2 writing activity at different periods in time as well as differences in cognitive activity which appear to be mediated by L2 proficiency. Importantly, the obtained patterns differed depending on whether duration or frequency data were considered. These findings thus demonstrate the need to broaden the study

of the temporal dimension of L2 writing and to consider more nuanced mixed-methods approaches in future work.

Keywords English; second language; writing processes; multimodal data; temporal dimension; digital screen capture; eye-tracking

Introduction

Despite the interest and, ultimately, necessity to fully understand the cognitive processes that underlie second language (L2) writing, the temporal dimension of L2 writing has been neglected. For decades, scholars have been preoccupied with understanding the type of processes and strategies that writers deploy while composing (e.g., Bereiter & Scardamalia, 1987; Flower & Hayes, 1980; Hayes, 2012). To that end, L2 writers are normally asked to externalize cognitive activity through either concurrent or retrospective verbalization. These verbalized accounts have been valuable for developing descriptive taxonomies and for gaining insights into learners' perspectives on their use of strategies, but they are problematic for the study of emergent real-time writing activity. How long and when during the composing period, for example, do L2 writers dedicate time to constructing a text or rereading or revising their draft? We believe that answering questions such as these about the temporal dimension of writing is as important as categorizing the processes (Manchón, Roca de Larios, & Murphy, 2009; Tillema, 2012; van Weijen, 2009), but to achieve this goal it is necessary to foreground a number of methodological concerns. We propose that, in order to investigate how L2 writers allocate time to various L2 processes in real time, researchers need to make use of a range of data, tools, and techniques and to broaden their analytical focus. In this study, we gathered data using digital screen capture, digital video recording, and eye-tracking software, as well as stimulated retrospective recall.

The overarching aims of our project were (a) to investigate cognitive processes emergent during real-time L2 writing activity in an English as a foreign language (EFL) university context and (b) to investigate relative differences in these processes across varying individual developing capacities as reflected in the participants' L2 proficiency level and the quality of their resulting texts. Ultimately, a better understanding of the online, emergent process of writing will contribute to L2 writing theory building. In this article, we report on the allocation of time to different writing processes throughout the composing period among L2 writers at different levels of proficiency. We use the term *processes*[1] in a broad sense to refer to strategies or actions, that is, externalized cognitive activity undertaken by L2 writers in order to perform a writing task. L2 writing activity involves problem-solving processes, including those that can be performed with various degrees of automaticity during text construction. These processes involve strategic action, such as purposefully using a bilingual dictionary to tackle a specific formulation problem, for example, to convert a verbal thought generated in writers' first language (L1) or in their L2 into L2 written language during text construction.

Historical Foundations of L2 Writing Process Research

Writing Processes

The study of writing processes in L2 contexts has built upon work and models developed within L1 writing, in particular, the influential cognitive models advanced by scholars such as Bereiter and Scardamalia (1987) and Flower and Hayes (1980, 1981), who described the process of writing in terms of four elementary mental macro processes and a series of subprocesses that constantly interact while a writer composes a text. The elementary processes are planning, formulating (also known as translating), reviewing, and monitoring (Flower & Hayes, 1981, p. 370; Hayes, 2012). The latest model additionally includes searching, in recognition of writers'

use of external sources, such as online dictionaries, to access information during the process of writing (Leijten, Van Waes, Schriver, & Hayes, 2014, p. 325). An important aspect of this model is the element of goal setting by writers, which is described as occurring primarily, although not exclusively, as part of the planning process. The main role of monitoring as a component of the model is to enable writers to coordinate the overall writing processes in a reiterative cycle that evolves as writers' goals change throughout the writing task. Building either directly or indirectly on these general models, scholars set out to investigate L2 writing processes, and this has resulted in several descriptive taxonomies (e.g., Sasaki, 2000; Wong, 2005) and in theory building (see Grabe, 2001; for comprehensive reviews of writing strategies, see Manchón, Roca de Larios, & Murphy, 2007, and also Leki, Cumming, & Silva, 2008).

In general terms, studies of L2 writing processes have reported findings which have focused either on the entire range of processes used by L2 writers to tackle a particular writing task or on how learners deploy specific strategies for particular macro processes, such as planning, formulation, or revision. The present study falls into the former category and, therefore, we endeavored to capture the full range of processes our L2 writers engaged in while composing. Research in this area has produced informative taxonomies that describe L2 writers' strategic behavior. The underlying foundations in this work have tended to be the general processes of planning, formulating, revising, and monitoring. However, as illustrated in Table 1, the level of detail, the range of processes categorized in individual taxonomies, and their specific focus have varied considerably.

< Table 1 near here>

Temporal Dimensions of Writing

Limiting analyses of writing processes to frequency counts has been criticized for treating composing processes as static entities and neglecting a much less studied, yet key, aspect in L2 writing: the temporal dimension of composition (Roca de Larios, Manchón, Murphy, & Marín, 2008, p. 32; see also Schoonen, Snellings, Stevenson, & van Gelderen, 2009). Indeed, the temporal distribution of cognitive processes throughout task execution matters (Leijten & Van Waes, 2006; Roca de Larios, Marín, & Murphy, 2001; Tillema, 2012; van den Bergh & Rijlaarsdam, 1999). The processes writers engage in at a certain moment in time are seen as a reflection of how they (re)conceptualize the task throughout the writing process. Therefore, the study of temporality in writing is crucial in investigations of complex cognitive activity: "[time] is an (observable) indicator of another conceptual variable, the changing task situation" (Rijlaarsdam & van den Bergh, 1996, p. 107). Thus, the complex and demanding task of producing text necessarily results in competition between various cognitive and linguistic concerns that interact with task requirements and, in the case of L2 writing, make L2 proficiency (discussed below) a particularly relevant aspect to consider (Galbraith, 2009).

Temporal aspects of writing have been an important component of a comprehensive project in the Netherlands which compared L1 (Dutch) and L2 (English) writers, revealing differences between L1 and L2 writing (Tillema, 2012; van Weijen, 2009). For instance, Schoonen et al. (2009) reported on an investigation of text production processes in real time which, in line with the researchers' expectations, showed that participants writing in their L2 spent more time solving language problems than when writing in their L1, suggesting that more attention was devoted to linguistic processing in the L2 context. However, studies that have investigated L2 writing processes across a range of L2 proficiency levels are scarce. Most of the L2 studies available to date have considered specific processes, such as when most planning or

revising takes place depending, for example, on writing expertise or skill (for review, see Roca de Larios et al., 2008).

In a particularly relevant study, Roca de Larios et al. (2008) investigated L2 writing activity during a complete 1-hour composing period and treated L2 proficiency as a mediating factor. Their participants were 21 Spanish EFL students equally distributed at three levels of proficiency. Drawing on concurrent think-aloud protocols as participants wrote an argumentative essay, the researchers operationalized the temporal distribution of activities by dividing the composition time into three periods: beginning, middle, and end stages (see also Manchón et al., 2009). They then analyzed the time spent by their participants talking about specific writing processes in the transcribed protocols (across the categories of reading prompt, task conceptualization, planning, formulation, evaluation, revision, and providing metacomments) and reported that, regardless of L2 proficiency level, there was a predominance of comments in the formulation category (which peaked in the second period) over the comments in the planning, evaluating, and revising categories. Furthermore, planning was concentrated in the first period, and revising gradually increased from beginning to end. When L2 proficiency was considered, the main finding was the association of recursiveness of various processes throughout the writing task at higher L2 proficiency levels, a result which, based on van den Bergh and Rijlaarsdam (1999), the researchers interpreted as a greater ability to make strategic decisions. In more detail, higher L2 proficiency was associated with overall diversification of activity, with planning concentrated in the first period and progressively decreasing; revision patterns were the opposite, such that formulation for the highest L2 proficiency level peaked in the second period. In contrast, lower L2 proficiency was associated with more time spent on formulation at the beginning of composition. Tillema (2012), who also studied writing processes

throughout the full writing activity in L1 Dutch secondary students learning L2 English, did not find an effect for L2 proficiency on time spent on formulation or on any other L2 writing processes except process planning, such as participants talking about rereading their text or expressing a regret for not having made an outline of their essay prior to writing.

Finally, although lower levels of L2 proficiency have been associated with poorer text quality (e.g., see Sasaki & Hirose, 1996; Schoonen, van Gelderen, de Glooper, Hulstijn, Simis, Snellings, & Stevenson, 2003), we know much less about the influence of temporal aspects on L2 writing processes and text quality. Once again, some studies have focused on selected processes, such as formulation (Roca de Larios, et al., 2001) or revision (Stevenson, Schoonen, & de Glopper, 2006), but further work is needed to better understand relationships between text quality and the full range of processes used by L2 writers while composing. Two studies which addressed these issues are those by van Weijen (2009) and Tillema (2012), both of which identified a relationship between variations of occurrence of certain cognitive processes and text quality at different stages of writing. Nonetheless, our understanding of these issues is still limited.

To summarize, previous work has laid the foundations for the study of L2 writing processes. However, a methodological issue prevalent in studies of L2 writing processes is that the findings have overwhelmingly relied on what participants reported in terms of strategy use. Popular tools for data collection have included questionnaires, interviews, process logs, and particularly such introspective techniques as concurrent think-aloud and stimulated retrospective recalls. Pitfalls associated with some of these techniques have been extensively documented and acknowledged (e.g., Hyland, 2009; Janssen, Van Waes, & van den Bergh, 1996). The few studies which included direct observation of L2 writing processes, such as videorecording or even screen

capture (e.g., Sasaki, 2000; Zamel, 1983) have used these valuable sources of data to elicit retrospection (e.g., as prompts for stimulated recall) and, as succinctly put by Park and Kinginger (2010), "Retrospective data is [sic] not so much a precise reproduction of the composing process as a reinterpretation of it… Reliance on the participant's account taken at face value, thus, can and does undermine the validity of research" (p. 31). Therefore, in order to properly understand the attentional and temporal dimensions of L2 writing activity, it is indispensable to study that activity as it unfolds over time and, importantly, to do so by relying on direct observations in addition to verbally mediated data. Technological advances can support this endeavor.

Technological Affordances and Current Perspectives on L2 Writing Processes

Scholars have begun to exploit cutting edge technologies for the study of L2 writing processes (Latif, 2008; Van Waes, Leijten, & Neuwirth, 2006). Three such technologies include keystroke logging, digital screen capture, and more recently, eye-tracking tools. Keystroke logging software produces time stamps for keyboard use, cursor movements, and mouse clicks in order to reconstruct and facilitate interpretation of writing activity (for an in-depth review of this technology see Leijten & Van Waes, 2013). Digital screen capture is defined as "a specialized software application used to record and save for future viewing an audiovisual trail (image or video) of the specific actions that are visible and audible as a person interacts with a screen in a digital environment" (Hamel, Séror, & Dion, 2015, p. 11; see also Degenhardt, 2006). Eye-tracking technology records and measures eye movements while a person is looking at a computer screen (e.g., the place, sequence, and length of gaze). During reading, eye-tracking measures both the moments when the eyes are relatively stationary (fixations) as well as the rapid movements (saccades) made from one fixation to another (Elgort, Brysbaert, Stevens, & Van Assche, 2017; Roberts & Siyanova-Chanturia, 2013). A unifying advantage of these

technologies is that they are unobtrusive and, therefore, support ecological validity in L2 writing research by not interfering with the composing process. The evolution of the text as well as all actions on the desktop (e.g., use of a browser) can be fully captured and replayed for subsequent analysis.

Research based on keystroke data has produced interesting accounts of L1 writing in particular as well as comparisons between L1 and L2 writing behaviors. Studies have tended to focus either on the investigation of specific writing processes, such as pauses or revision, or on the interpretation of processing activity to determine, for example, levels of writing fluency (see Miller, Lindgren, & Sullivan, 2008). This technology has also supported, for instance, theoretical and methodological developments relating to the measurement, conceptualization, and understanding of pausing and (re)reading behaviors during writing (Van Waes, Leijten, Lindgren, & Wengelin, 2016). In contrast, a very limited number of studies have made use of digital screen capture to investigate L2 writing patterns as they unfold in real time and have also recorded and studied on-screen activity beyond keyboard and mouse use (e.g., use of external resources, such as the Web). Emerging work has suggested that digital screen capture has the potential to strengthen research design and contribute to studies of complex cognitive activity. For example, Park and Kinginger (2010) investigated the composing activity of a Chinese advanced English L2 learner at an American university. The researchers used an innovative combination of data sources: computer screen recordings, corpus-based query analysis, and reflections to propose an analytical framework based on hypothesis-testing behavior. Their data analysis identified three recursive steps followed by their participant: hypothesis testing through a query, analysis and evaluation of search results, and revision. The researchers suggested that

the problem-solving nature of the composing process observed led to changes in the state of knowledge of their participant.

In an interesting case study also using digital screen capture as a tool for data collection, Séror (2013) provided a brief overview of two L2 French (B2 CEFR level) learners' writing processes whose L1 was English (see also Hamel & Séror, 2016). Although descriptive in nature and only based on the analysis of a fragment within a composition, Séror's findings demonstrated that various processes can be observed through the recordings made while writers are composing in their target language. The study also documented a considerable difference in the amount of time the learners spent using online resources, such as dictionaries, translation sites, and online grammar checkers (25% vs. 8.5% of the analyzed data for these two students).

The use of eye-tracking devices has become well established in the broader field of psycholinguistics and in the study of reading activity (e.g., Dussias, 2010), but it is still in its infancy in L2 writing research. Emerging developments in methodological approaches to the study of writing activity are ongoing. A case in point is the integration of eye-tracking and Inputlog, a leading keystroke logging technology, in order to more accurately define and measure reading activity during writing (e.g., De Smet, Leijten, & Van Waes, 2014).

The Current Study

Although investigation of learners' cognition while composing is not an easy feat, scholars have come a long way in the development of taxonomies that support the conceptualization and modelling of the writing process; some of those taxonomies reflect its multidimensional nature involving cognitive, metacognitive, and affective strategies. More recently, work has begun on the study of the temporal aspects of L2 composing (e.g., Van Waes & Leijten, 2015), but research in this area in particular is still scarce. The challenges that researchers face and that

become apparent when one examines the previous literature in this field of study often relate to methodological issues. One of these is the overreliance on elicited verbalizations as the principal source of data, which are problematic when attempting to accurately assess L2 writing activity in real time. In particular, concurrent think-aloud protocols have been criticized on the grounds of both reactivity (i.e., the act of verbalizing might influence the task itself) and the risk of cognitive overload (van den Haak, de Jong, & Schellens, 2003). In order to gain more accurate accounts of the processes underlying L2 composition, it is therefore necessary to make use of various sources of both quantitative and qualitative data—mixed methods research (Leijten & Van Waes, 2013). The value of this approach lies in its power to effectively capture patterns in L2 writing processes at various levels: macro (e.g., whole composing time), meso (e.g., subperiods within composing time), and micro (e.g., moment-by-moment changes in time).

To study the real-time evolution of a writing event across a range of L2 proficiency levels, we considered it necessary to gather data through a combination of tools. Inevitably, methodological decisions involve a careful balance between research objectives and practical considerations, not least because of rapid and continuous advances in technology. Thus, as discussed in detail below, we decided to use digital screen capture as our primary source of data. Digital video recording, eye-tracking, as well as stimulated retrospective recall served as complementary sources of information to strengthen the coding, analysis, and interpretation of data and to add, when relevant, a qualitative dimension to our understanding of the phenomena. Finally, for our study to more accurately reflect contemporary L2 writing practices, our participants had access to a range of online resources while they were composing their essays: a monolingual dictionary, a bilingual dictionary, a thesaurus, and a Web browser. As Leijten and Van Waes (2013, p. 383) have stated, “the interaction with multiple sources—intentionally and

unintentionally—has become an inherent part of most writing processes." The study addressed the following sets of research questions:

1. Do different writing processes differ in temporal duration and/or frequency throughout the composing period? If so, do these processes differ depending on the sequential time period?
2. Are there any relationships between L2 proficiency (as measured by a C-test) and temporal duration and/or frequency for different L2 writing processes? If so, do these relationships differ depending on the sequential time period?
3. Are there any relationships between essay quality (operationalised as blind-rated essay scores) and/or essay length, temporal duration, and frequency for different L2 writing processes? If so, do these relationships differ depending upon the sequential time period?

Method

Participants

The participants were 22 EFL students (6 males, 16 females) from two Japanese universities and an EFL tutor/researcher (the 2nd author). The tutor was qualified (DTEFLA, MA (ELT), PhD (applied linguistics), Cambridge CELTA teacher-trainer) with over 25 years English language teaching experience in five countries. Students were recruited to the study from the tutor's own academic reading & writing classes (known for less than 1 year) as well as other classes from the two universities (these participants were therefore not known to the tutor). Participation in the study was entirely voluntary, with students offered a one-off payment of 1,000 Yen (approx. 9 dollars) for their participation in the project - recruitment was carried out through announcements at the beginning of classes or meetings. The participants' L1 was Japanese in all cases except for three, whose L1s were Mandarin, Korean, and Spanish. Participants' ages

ranged from 18 to 40 years (M = 21.4, Mdn = 20, SD = 5.58), with length of time learning English from 6 to 14 years and self-reported writing expertise assessed as elementary to advanced. Participants' L2 proficiency levels ranged from elementary to advanced (see Appendix S1 in the Supporting Information online).

Tools and Procedure

All data (obtained with full written consent) were collected by the tutor/researcher on a computer in his office and on an individual basis in three stages: a precomposition stage, a composition stage, and a stimulated retrospective recall stage. In the precomposition stage, participants completed a 116 item C-test in order to estimate their English proficiency (see Gilmore, 2011). The C-test is similar to a traditional cloze test except that it involves deleting the second half of every second word in a text and the text starting and ending with an intact sentence (Grotjahn, 2010; Klein-Braley & Raatz, 1984). It has been found to be a superior measure of general language proficiency compared to the standard cloze test (Dörnyei & Katona, 1992).

In the composition stage, participants were first familiarized with the hardware and software, including a word processor and the various online resources mentioned above, and the eye-tracker was calibrated. Then, participants were given 10 minutes planning time (not included for analysis in this study)[2] for writing an IELTS style argumentative essay on the topic: "Education should be free for everyone. To what extent do you agree or disagree with this statement?" An argumentative essay was chosen because this is a preferred type of task for the investigation of writing processes given its potential for knowledge transforming and problem solving demands (Roca de Larios et al., 2008). The essay topic was chosen for its familiarity and engagement potential.

Each participant was given 35 minutes to write the essay. The eye-tracking suite Tobii T60/Studio 2.2[3] was used to produce the core source of data, that is, visual records (from digital screen capture) of the whole L2 writing event (13 hours in total) with eye gaze data overlaid. This technology renders a powerful visualization of L2 writing processes, including the use of online resources (Latif, 2008; Park & Kinginger, 2010) by integrating eye gaze data with on-screen activity, recorded sound, keystrokes, and mouse clicks. Eye movements are known to be driven by both bottom-up processes (properties of the visual signal) and top-down effects (the task, affective state, prior knowledge, or semantic context) (Couronné, Guérin-Dugué, Dubois, Faye, & Marendaz, 2010) and are therefore an excellent way of unobtrusively tracking the moment-to-moment cognitive processes involved in L2 text construction. Digital video recordings of the participants' interaction with the computer and their paper notes were also collected in order to capture all possible activity during composing time (see Figure 1 for a summary of all elements captured for subsequent analyses).

< Figure 1 near here>

In the third and final stage, we conducted stimulated retrospective recalls to increase the accuracy of data coding and improve interpretation of L2 writing behaviors as well as to provide data for complementary qualitative analysis. Following general guidelines (Gass & Mackey, 2000), the retrospective stimulated recall protocol (based on the digital screen capture video with gaze data replay, as shown in Figure 1) was initiated after a 10 minute break subsequent to the writing activity and while the writing event was still fresh in participants' memories. The participants received the following instructions:

> We will now watch your composition video, and I would like you to talk me through what was going on in your mind as you were writing your essay. You can press the pause

button whenever you want to make a comment, and if I pause, I would also like you to tell me what you were thinking at the time.

The stimulated retrospective recalls were also recorded using Tobii Studio 2.2, yielding a dataset of a total of 27 hours 6 minutes of recordings. All retrospective stimulated recall data were transcribed in full to produce written protocols for subsequent analysis.

Data Analysis

C-Test and Essay

The C-test was scored using the exact word scoring method (Weir, 1990), assigning 1 point to each correct answer (maximum score = 116 points). The essay was blind-rated by three native speaker teachers with language testing training and experience. Using the IELTS Task 2 writing band descriptors and scoring procedures,[4] the raters scored each essay on four dimensions: task achievement, coherence and cohesion, lexical resources, grammatical range and accuracy. These four ratings were then averaged to provide a global score for each composition. The result for each participant was obtained by calculating the mean of the global score given by each of the three expert raters (see Appendix S1). Interrater reliability was excellent (Cronbach's $\alpha = .97$). Essay samples from the participants at the highest and lowest L2 proficiency levels are provided in Appendix S2 in the Supporting Information online. C-test scores and essay scores correlated strongly, $r = .75$, $p < .001$.

Digital Screen Capture Data

In preparation for analysis, the digital screen capture videos of real-time L2 writing behaviors were simultaneously segmented into episodes and coded for several processes using the ELAN v.4.8.1 annotation software (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). Once all files are segmented and coded, ELAN produces descriptive statistics, for example, frequency

and temporal duration (length in seconds) of episodes, which formed the basis for subsequent statistical analyses. An episode was defined as a segment of video which contained only one L2 writing process, such as text construction or revising. A new episode reflected a writer's switch to a different L2 writing process (see van den Bergh & Rijlaarsdam, 2001, and Tillema, 2012, p. 41). The simultaneous procedure of video segmentation into episodes and coding the episodes followed a recursive process between the authors (Bernard & Ryan, 2010). The coding scheme was adapted from van Weijen, van den Bergh, Rijlaarsdam, and Sanders (2009) and Stevenson et al. (2006). After intercoder reliability was established ($\kappa = .83$) based on 10% of the data (Landis & Koch, 1977), all video files were segmented and coded by one of the authors. Intracoder reliability was subsequently checked using a random sample of 5% of the data ($\kappa = .93$). Finally, once all data had been coded, a second cycle of intercoder reliability based on a random sample of 10% of the data was conducted ($\kappa = .86$).

Episode Coding Categories

We established six categories for coding episodes into discrete writing processes (with all examples taken from the data):

1. Text construction: Period when students were producing new text, that is, typing the actual words on the computer, for example, "I agree."
2. Revising: Period when any previously written text was modified at word, sentence, or text level. Revisions could occur both at the point of inscription, for example, when a participant typed the letters *th* but then immediately deleted them, and at a point in the text previous to the point of inscription (see Stevenson et al., 2006, p. 206), for example, as a participant deleted *elementary* and then typed *public*. We coded as text construction the moment the participants began producing new text unless the new text was part of an

addition to a previously written sentence, for example revising "[m]oreover the parents' income" to "[m]oreover depending on the parents' income."

3. Rereading: Period when students were rereading segments of their previously written text, as evident through the combination of digital screen capture and eye-tracking data (see Figure 1). For example, immediately after an episode of text construction, a participant's gaze lifted from the keyboard to the screen, and a series of 24 fixations and saccades clearly showed a rereading pattern. The moment when the writer redirected her gaze towards the keyboard marked the end of the rereading episode. The retrospective stimulated recall protocols provided further data to support coding decisions.
4. Use of external resources: Period when students left the word processor in order to access external resources, for example, monolingual dictionary, bilingual dictionary, thesaurus, Web browser, or paper notes, as evident through screen capture data, video recorded data, and retrospective stimulated recall protocols.
5. Pausing:[5] Period when activities described in the previous coding categories ceased temporarily. In other words, a pausing episode marked a transition between two L2 process episodes. This happened, for example, when the video recorded data showed a participant looking down as if thinking or when a participant was looking in the direction of the screen, and his/her gaze was fixated off-text as shown by the eye-tracking data and he/she was neither rereading nor typing. The retrospective stimulated recall protocols provided further data to support coding decisions.
6. Other: Period when students were doing something other than the above, for example, looking at the word processor tool bars, searching at the bottom of the screen before

opening a Web browser, or performing random eye movements around the screen not focused on the text.

Following Roca de Larios et al. (2008), "composition time was operationalized by measuring each individual [L2 process] category in seconds and adding up the total amount of time for all of them" (p. 37). To study the temporal distribution of L2 processes throughout the entire duration of text construction, however, we produced a more fine-grained analysis by dividing the total amount of time for each composition into five periods instead of three (see also Tillema, 2012).

The key process measures were quantified as follows: Each participant's personal total time was divided into five equal length periods. Episode frequency was calculated for each participant within each period for each process and expressed as a percentage of total episodes within each period. This served to remove any effects arising from the fact that, although the same opportunity was available to all participants, some of them evidenced more episodes overall than others (M = 298.8, *range* = 108–551). Total duration of episodes was similarly calculated for each participant within each period for each process and expressed as a percentage of total time within each period. This served to remove effects arising from the fact that, although the same writing time was allowed for all (35 minutes) and the majority were close to using it all, not everyone did (M = 33.3 minutes, *range* = 14.8–35.0).

Normality of the data was checked using the one sample Kolmogorov-Smirnov test with Lilliefors correction: 82% of the data passed the test, which was deemed sufficient to proceed to analyze the data with parametric statistics using repeated-measures ANOVAs. To address Research Question 1 (RQ1), the design was treated as containing two repeated-measures factors: the five process types and the five equal sequential time periods. Language proficiency measured

by the C-test was added as a covariate for answering Research Questions 2 and 3 (RQs 2 and 3). Episode frequency and total duration were treated as continuous dependent variables for answering RQ1 and RQ2 but as covariates for answering RQ3, where text quality score and text length were the dependent variables. Post hoc tests were conducted where this was necessary using the Bonferroni correction. Where sphericity was violated, Greenhouse-Geisser corrections were applied to determine the statistical significance of *F* values. The models tested were in some cases necessarily incomplete due to the design of the study. For instance, where percentage frequencies of episodes of different process types calculated within each period were the dependent variables, it was impossible to test for a main effect of period or for an interactive effect of proficiency and period because every period had the same overall percent of episode occurrences (100%) and the same participant proficiency.

Results

Research Question 1

The initial analyses targeted the first set of research questions: Do different writing processes differ in temporal duration and/or frequency throughout the composing period? If so, do these processes differ depending on the sequential time period? An initial overall analysis (five processes by five periods) targeting the frequency of episodes showed that, taking all periods together, there was a significant main effect of process type, $F(4, 84) = 28.78$, $p < .001$, $\eta_p^2 = .58$. Furthermore, there was a significant interaction effect between process type and period, $F(6.03, 126.62) = 5.87$, $p < .001$, $\eta_p^2 = .22$. A parallel analysis of the total duration of episodes also yielded both a significant main effect of process type, $F(4, 84) = 5.25$, $p = .001$, $\eta_p^2 = .20$, and a significant two-way interaction, $F(5.64, 114.44) = 6.05$, $p < .001$, $\eta_p^2 = .22$.

With respect to the main effect of process type (illustrated in Figure 2), although there was a broad similarity between the two results, there were also important differences. First, the differences between processes were more marked for the frequency data compared to the duration data. This may be seen from the larger effect size in the frequency analysis, compared to the effect size in the duration analysis (η_p^2 = .58 vs. .20). Furthermore, the frequency differences between six of the 10 pairs of processes were significant, although for duration only three of the 10 pairs of processes differed significantly (for details, see Appendix S3 in the Supporting Information online). Most notably, text construction exceeded all process types except revising in terms of frequency and all except external resources in terms of duration. Second, the order of processes from most frequent and longest in duration to least frequent and shortest in duration agreed in all respects except for the position of using external resources, which was lowest in the frequency but third longest in the duration analyses. Regardless of whether frequency or duration was considered, however, text construction took the highest position followed by revising, although use of external resources came close to revising in terms of duration.

< Figure 2 near here>

With respect to the significant interaction, where the effect size was greater in the duration than frequency data (see details in Appendix S4 in the Supporting Information online), we first considered the frequency data, comparing the processes at each separate time period and then comparing time periods for each separate process. As Table 2 illustrates, there were significant differences between frequencies associated with different processes in every successive separate time period (see Figure 3). As the effect size measures indicate, there was a noticeable trend (except in Period 5) for the differences between process frequencies to decrease across successive periods with some convergence towards 20%, which would be the chance

percentage considering all five processes. Furthermore, as confirmed through follow-up pairwise comparisons of process types across time (summarized in Appendix S5 in the Supporting Information online), there was a transition over time from text construction significantly exceeding all other processes in Period 1 to using external resources emerging as the least used process, compared to all others, in Period 5.

< Table 2 near here>

< Figure 3 near here>

As shown in Table 3, the follow-up pairwise comparisons of time periods for each process type revealed that revising, pausing, and using external resources showed no significant tendency to rise or fall over time. The other two processes exhibited significant linear trends, meaning that they tended to become more frequent (rereading) or less frequent (text construction) successively over time. Indeed, the temporal sequence trends in both cases had greater effect sizes than those that were associated with the tests of differences between periods regardless of the order they were in.

< Table 3 near here>

In the corresponding analyses using total episode duration as a dependent variable (see Table 4 and Figure 4), the effect sizes for episode duration differences at each time period exhibited the same general tendency to reflect more differentiation in Period 1 and progressively less in successive periods, converging on 20% (i.e., chance), with some divergence again in Period 5. In this case, unlike the frequency result, the five process types were not significantly different in Period 4 and barely so in Period 5. Text construction in Period 1 took up a significantly greater percentage of time than did all other processes, except using external resources, and rereading occupied significantly less time than all others, except use of external

resources. This pattern then decayed over subsequent periods, and in Period 5, strikingly, rereading emerged as occupying the greatest rather than the least percentage of time.

< Table 4 near here>

< Figure 4 near here>

The results for time differences between periods for each process taken separately had a good deal in common with those for frequency (see Table 5). There was once again no significant variation in pausing, revising, or use of external resources dependent upon period. What dominated the duration results was the dramatic falling pattern exhibited for time devoted to text construction and the corresponding rise in rereading time. The effect sizes for these trends were in this instance similar to those for the frequency data.

< Table 5 near here>

Research Question 2

The next analyses addressed the second set of research questions: Are there any relationships between L2 proficiency (as measured by the C-test) and temporal duration and/or frequency for different L2 writing processes? If so, do these relationships differ depending on the sequential time period? We performed an overall analysis for each dependent measure (frequency and total duration of episodes) with the five process types and the five successive periods as repeated-measures factors and the C-test scores used as a covariate. This allowed us to test if C-test scores correlated with (a) frequency or (b) amount of time spent on episodes, and, if so, whether these relationships differed depending on what process was involved or what period the episodes occurred in. For both dependent variables, there was a significant interaction between process type and proficiency, indicating that relationships between C-test scores and episode frequency or duration differed for different process types but not for different periods: $F(4, 145) = 8.15$, $p <$

.001, η_p^2 = .09, in the analyses of frequency, and $F(4, 156) = 5.72$, $p < .001$, η_p^2 = .10, in the analyses of duration.

In order to examine the nature of these differences, we calculated follow-up Pearson correlations between proficiency and each dependent measure for each process separately (see Table 6). The strongest correlation was a significant negative association between proficiency and use of external sources. The effect sizes were similar for both dependent measures: More proficient students engaged less in use of external resources, both in terms of frequency and duration. In contrast, there was a significant positive correlation between text construction and L2 proficiency, again for both episode frequency and duration. The only other significant association involved revising, which was obtained only for episode frequency. More proficient students devoted a higher percentage of their episodes, but not more of their time, to revising than did less proficient students. Overall, then, quantitative analyses of proficiency (as measured by a C-test) in relation to episode frequency and duration, showed that the two dependent measures behaved similarly but yielded by no means exactly the same findings.

< Table 6 near here>

Research Question 3

The final analyses targeted the third set of research questions: Are there any relationships between essay quality (operationalised as essay score) and/or essay length, temporal duration, and frequency for different L2 writing processes? If so, do these relationships differ depending upon the sequential time period? Because the aim of these questions was to examine if any writer process variables were related to either of the essay product variables (essay quality score and/or essay length), included in the analyses were all potentially meaningful predictor variables and

principal interactions between them (writer L2 proficiency, frequency and duration of episodes, five processes, and five periods).

For essay score as the outcome variable, the analyses (summarized in Appendix S6 in the Supporting Information online) showed that only proficiency had a significant impact on essay score, $F(1, 490) = 606.63$, $p < .001$, $\eta_p^2 = .56$. It seems that no variation in episode occurrence or time given to different processes overall, or differentially in different periods, had any obvious impact on essay quality when considered against writer proficiency. Indeed, even when proficiency was omitted from the analysis, the variables of episode duration and frequency still showed no significant relationship with essay quality scores. For essay length as the outcome variable, however, there were effects beyond that of proficiency, despite its clear impact, $F(1, 477) = 175.02$, $p < .001$, $\eta_p^2 = .27$. In this case, there was a significant effect for variation between processes in duration and in episode frequency.

These findings are best understood from the follow-up Pearson correlations (summarized in Table 7). These analyses suggested, first of all, that the more time that participants allocated to text construction, the longer was the text that they produced ($r = .31$). The frequency of episodes devoted to text construction also had a significant (but weaker) relationship with text length ($r = .19$). This makes sense as, presumably, what leads to a longer text is time spent writing rather than the number of instances when writing occurs. Put differently, multiple instances can be brief and interspersed with other processes, yet what matters for text length is the total writing time. Revising also had a positive association with text length, both in terms of frequency and duration. Conversely, using external resources had a strong negative relationship with essay length, presumably because this activity takes time away from the production of text and is also the behavior more associated with lower proficiency writers. Indeed, this relationship was

stronger in the duration data (r = –.49) than in the frequency data (r = –.32), and it was the strongest association, except for that for proficiency (r = .52). Finally, pausing, in frequency terms, also had a negative association with text length, again because a writer is not producing new text when pausing, and pausing may also be the behavior typical of lower proficiency students. The same might be expected for rereading, but in fact rereading had no significant association with text length.

< Table 7 near here>

Discussion

This study set out to investigate the L2 writing activity of 22 EFL writers as they engaged in the composition of an argumentative essay. Our investigation is unique in the EFL context in that its findings are based on the study of complex cognitive activity as it unfolded in observed real time writing. A mixed-methods research design allowed us to capture, and subsequently measure, the recursive and chronological evolution of the writing event in order to contribute to existing knowledge about L2 writers' behavior, on the one hand, and specific composing activity trajectories across a wide range of L2 proficiency levels, on the other.

Duration and Frequency of Different Writing Processes

Overall, during the 35 minute composing period, text construction and revising were the dominant processes. This first (intuitive) finding is reminiscent of Manchón et al.'s (2009) conclusion that formulation is a dominant process, although it was not as marked in the current dataset as in Manchón et al.'s study (see also van Weijen, 2009). However, direct comparisons between studies are problematic because of the differences in methodology. In particular, an important aim of the current design was to observe and measure real-time composing behavior which was not mediated by concurrent verbalization (think-alouds). Furthermore, the writers in

this study had access to online resources, such as dictionaries during the composition period. The second finding was that, with the exception of use of external resources, the remaining L2 writing processes were used more frequently but tended to be shorter. In other words, the overall pattern of use for the majority of processes could be described as "little and often," suggesting a complex approach to composing where the various writing processes are dynamically and contingently intertwined (Bereiter & Scardamalia, 1987; Flower & Hayes, 1981). These patterns were more marked in the frequency than in the duration data.

In an effort to address a central gap in current understanding of the intrinsic temporal dynamicity of L2 composing activity and in response to a call to attend to this neglected matter (Manchón et al., 2009), we also investigated the allocation of time to different L2 writing processes at different stages of the composition task. Our results show clear differences in L2 writing activity at different periods in time and confirm that occurrence of cognitive activities varies throughout task execution, as has long been demonstrated for L1 writing (see van den Bergh & Rijlaarsdam, 2001). Furthermore, we found that however one looks at the data, that is, either in terms of frequency or duration, there was much less variation of activity in Periods 1 to 3 than in Periods 4 and 5. In Periods 1 to 3, the dominant processes were text construction and revising, and activity was more diversified in Period 4. This pattern continued in Period 5 for four of the processes, whose use remained balanced but contrasted with rereading, which then became dominant.

If the trends for each process are looked at in more detail, our findings further support the observation that "the various composing activities… participants engaged in did not stand an equal chance of being activated at any given time in the composing process" (Manchón et al., 2009, p. 108). We found that text construction and rereading showed a significant linear

tendency across the five periods both for frequency and, even more markedly, for duration. As text construction decreased, rereading increased to become a dominant process towards the end of the composition period. In contrast, Manchón et al. (2009) found that formulation reached its peak in the middle of the composing period. However, they divided their composition time into only three periods, compared to our five, and revision gradually increased from the beginning to the end, which was the pattern that we found for rereading. Although Manchón et al. did observe instances of rereading in their data, this process does not seem, unfortunately, to have been included in their model other than for rereading of the essay prompt, so further comparisons in this respect were not possible. For revising, our findings are more in line with those of Tillema (2012), who found that its occurrence did not vary across the writing process.

The examination of digital screen capture, eye-tracking, and retrospective stimulated recall data allowed us to gain a deeper understanding of these patterns. For example, Figure 5 shows an intense burst of rereading by Participant 3 (P03) starting at 33:08 minutes into her nearly completed essay. Interestingly, she did not start rereading from the beginning of her essay, as she explained in the stimulated recall:

P03: Yeah so the first paragraph is where I wrote mmm simple idea my simple idea …but from the second paragraph I said more specific argument so I'm not worrying that logic is appropriate or grammar.

Tutor: Right okay so you thought it was more likely to be problems there okay.

< Figure 5 near here>

The eye-tracking data clearly show P03 spending the final few minutes of the allocated time rereading the complete essay, checking for errors and coherence. This rereading of the text also instigated a series of revision episodes, as she refined her work, a point taken up in the

recall, as seen in the following excerpt where she justified her reasons for changing *also* into *in addition*:

P03: So *also* it's not appropriate for academic writing and in this paragraph I do not use *in addition*

Tutor: Right so you thought you can make it more formal by putting *in addition* and you're not repeating so you think it was it's okay.

The multiple sources of data used in this investigation provided us with evidence of revision throughout the writing task, triggered by both composing and rereading activity, and this is consistent with the relatively stable pattern seen in our statistical analyses. Thus, we believe that through a mixed-methods approach we can better understand the complex interplay between the various writing processes. The final major finding was that differences observed between processes were more marked in terms of duration than frequency, as discussed previously. This was particularly apparent in relation to the external resources data which suggested that although participants did not consult external resources very frequently, when they did so, they spent considerable time using them (see also Séror, 2013).

Relationships Between L2 Proficiency and L2 Writing Processes

With respect to L2 proficiency levels represented by our participants, proficiency correlated positively with text construction, and, interestingly, more proficient writers devoted a higher percentage of their episodes to revision but not necessarily more of their time. More proficient writers also appeared to consult external resources less, both in terms of the frequency and the length of time that they spent on them. However, what more proficient writers actually did while consulting these resources was qualitatively different from the actions of lower proficiency writers, and this could only be captured through multimodal data and micro level analysis.

Figures 6 and 7 illustrate the contrast between two examples (which were nearly equal in duration) of the highest proficiency participant (P01) and the lowest proficiency participant (P22) consulting external resources. These showed the more sophisticated and highly regulated strategic behavior of P01, which also ultimately leads to more successful outcomes. This participant skipped to Page 8 of the search results for *hassei* (accrual/emergence) in an attempt to find a translation closer to her intended meaning; she then altered the search word to *umareru* (accrual/emergence), a synonym of *hassei*) to find a more appropriate translation, and she searched a monolingual dictionary to test her theory that *generate* and *generation* have different meanings. In contrast, P22 spent a similar amount of time online searching for a translation for *shinro* (career options/paths), which ultimately led to her producing the unnatural sentence, "After that, they choose the course for a dream." Both online lexical queries by participants were triggered by a need to translate a L1 concept into the L2: *accrual/emergence* for P01 and *career options/paths* for P22.

< Figure 6 near here>

< Figure 7 near here>

Thus, the complex and dynamic nature of L2 writing behavior clearly evident in studies such as the present one underscores and supports Rijlaarsdam and van den Bergh's (1996) suggestion that changes in process activity are likely to reflect different representations of the task as held by the participants at different moments in time, implying that they are contingent. In fact, some research has found that a factor underlying these different representations is L2 proficiency itself, suggesting that higher proficiency learners "appear to be able to strategically decide *what* attentional resources to allocate to *which* writing activities at *which* stages of the writing process" (Roca de Larios et al., 2008, p. 43). Tillema (2012), however, found only an

effect of proficiency on processes relating to metacognitive activity (e.g., process planning and evaluation of text) but not on any other cognitive activity. In our case, we did not find any statistical relationships between proficiency and writing processes at specific time periods. Nonetheless, when the temporal dimension was examined at a more nuanced (micro) level, as we demonstrated previously, it could clearly be seen that what higher L2 proficiency students do and achieve within a given time period appears to be radically different from what lower proficiency learners do and achieve.

Given the limited number of studies that have looked at the temporal dimensions of L2 writing, it is difficult to untangle the complex web of internal (e.g., motivation, working memory capacity, cognitive style) and external (e.g., task difficulty, topic, planning time) factors influencing L2 writing trajectories. Individual differences have long been studied and acknowledged in L2 research, although to a lesser extent in the subfield of L2 writing (but see Kormos, 2012). There have been interesting discussions of the potential impact of task difficulty and working memory on temporal strategic use of L2 writing processes (Miller et al., 2008). In fact, van Weijen (2009) reported that variation across individual writers has a larger impact on the use of writing processes than does variation between tasks. Evidently, more research is needed to gain further insights into these matters. More importantly, multiple, multimodal measures at macro, meso, and micro levels—as advocated, for example, by researchers using a complex systems approach (Gilmore, 2015; Larsen-Freeman & Cameron, 2008)—are required in order to better understand relationships between L2 writing activity and temporality.

Relationships Between Essay Quality and Length and Different L2 Writing Processes

Only L2 proficiency had a clear association with essay quality, operationalized as blind-rated essay scores (see also Schoonen et al., 2003). In other words, no other measures of duration or

frequency for various other writing processes had a relationship with essay quality. This was a surprising finding, given that variations in cognitive activity at different stages of composition were found to be associated with text quality (Manchón et. al., 2009; Roca de Larios et al., 2008; Tillema, 2012; van Weijen, 2009). We feel that the effects of L2 writing processes on text quality ultimately depend more on the deployment of the right strategy at the right time in a contingent way rather than on any predetermined patterns of process use during a particular period of task execution. For example, it would be over-simplistic to suggest that revisions occurring towards the end of a writing task necessarily lead to superior results. Writers may prefer to edit their texts regularly as they compose, or may simply be more accurate to begin with so that only minimal revision is required in the final stages. While temporality is a crucial aspect to consider when trying to understand text quality, the important question is: At what level (macro, meso, or micro) do certain temporality patterns emerge?

With regard to relationships between essay length and duration/frequency of different L2 writing processes, some clear relationships emerged. It was unsurprising to find that higher levels of text construction were positively correlated with essay length or that increased use of external resources was negatively correlated with it. More interestingly, there was a positive relationship between revising and essay length. Ad hoc analysis of the revising episodes from the digital screen capture suggested that one possible explanation for this finding was the relatively high percentage of revising episodes which included text additions (21.3%). It might also be possible that the revision process itself stimulates generation of new ideas, which in turn results in longer texts. Finally, there was a negative correlation between pausing and essay length, but only for frequency of episodes and not for episode duration. Thus, engaging in longer pauses does not necessarily result in shorter text, but pausing frequently might. In essence, longer

pauses might be used for conceptualization while writers formulate what they want to express next and thus result in more productive writing. However, the type of analysis which could throw further light on this issue is beyond the scope of this study.

Limitations

As with any research design, our methodological choices have inevitably resulted in certain limitations. In order to incorporate the array of data gathering tools necessary to best address the current research questions, we opted to rely primarily on Tobii Studio (which does not offer the keystroke logging sophistication of specialized programs, such as Inputlog) and ELAN. Continuous development and innovation, for example the now possible integration of eye-tracking technology and keystroke logging systems (Leijten & Van Waes, 2013), should further facilitate and strengthen research designs. One tradeoff of modern eye-tracking technologies—which are unobtrusive (e.g., allowing for free head movement) and therefore contribute to stronger ecological validity—is that eye gaze is often not captured in its totality. In this study, only approximately 50% of attempts by the Tobii T60 to record eye gaze were successful (with a sampling rate of 60 Hz, this equates to around 30 times per second). Head movements, leaning back in the chair away from the screen, or gaze aversion during demanding cognitive activities (Benedek, Stoiser, Walcher, & Körner, 2017) all have the potential to interfere with the tracking capability and affect data capture. These tracking rates were, however, adequate for our needs because most of the useful data (i.e., when participants were working on the screen) were captured. Another limitation of this study is that, although the obtained rich multimodal database includes composing activity of 22 L2 writers with a range of proficiencies, this activity was based on a single type of task. Van Weijen (2009) reported that, compared to L1 writers, L2 writers' behavior is surprisingly stable across tasks and, therefore, this issue may be of less

importance here. Nonetheless, caution should be exercised in the interpretation of these findings beyond this particular task type. Clearly, this is an area that needs further exploration.

Conclusion

The overarching aim of this study was to investigate real-time writing activity for a group of L2 writers from a range of proficiency levels. Ultimately, a better understanding of the online, emergent, process of writing should contribute to L2 writing theory building. As pointed out by Roca de Larios et al. (2008), the study of the temporal dimension of L2 writing could make a valuable contribution towards a fuller conceptualization of a writer's mental model, that is, the "whole set of conceptions and beliefs that underlie and guide writing performance"(p. 43). A second, and crucial, goal of this study was to respond to the call of Schoonen et al. (2009), urging colleagues to improve research design to increase validity. The current study, which has brought methodological issues to the fore, thus represents a step in that direction in three ways:

1. The first contribution of this study pertains to its use of data from multiple sources and modalities, in particular, digital screen recordings to capture real-time performance not mediated by verbalization. This reduces an overreliance on either concurrent or retrospective accounts of L2 writing activity as the main sources of procedural information. Furthermore, the current analyses incorporated video recordings showing L2 writers' interactions with the computer and their notes as well as eye-tracking which, together with stimulated retrospective verbalizations, complemented and supported data coding and analysis.
2. This study has demonstrated the importance of embracing a mixed-methods approach to the study of L2 writing in order to gain a more accurate understanding of the phenomena

under investigation. It also revealed the significance of broadening research into the temporal dimension of L2 writing to include both duration and frequency data.

3. A third contribution of this study pertains to its consideration of the use of external resources, such as online dictionaries, during L2 writing events. Access to online information is becoming an increasingly common element of writing processes (Leijten, Van Waes, Schriver, & Hayes, 2014) and therefore deserves to be included in research agendas if we hope to better understand real-world L2 writing behaviors (Hayes, 2012).

To conclude, we would like to emphasize that, as practitioners, our ultimate goal is to better understand the factors which might account for successful writing so that we can contribute to strengthening pedagogical practice and materials design (Hamel & Séror, 2016). A fuller characterization of the L2 writing process as it evolves in real time has the potential to support such an endeavor.

Final revised version accepted 15 November 2017

Notes

1 As Manchón, Roca de Larios, and Murphy (2007) have pointed out, there is an issue with the conceptualization of terms in the literature on writing processes, with authors using various terms to refer to similar phenomena (e.g., process, strategy, behavior, action, etc.) and failing to clearly define the notions.

2 This essay planning time was given to participants prior to the composition period in order to standardize this aspect of the investigation. It has been excluded from the current analyses because our focus here was on real-time composing behavior recorded with digital screen

capture. However, records of participants' planning notes were collected for analysis at a later date.

3 Tobii T60 and Studio 2.2 were used because they were both available at the university where this research was conducted and were judged to have the performance characteristics necessary to answer our particular research questions. Tobii T60 collects raw eye movement data points every 16.7 milliseconds, with each data point given a time stamp and x/y coordinates which are subsequently used to establish the location of the fixation. It is now also possible to integrate eye-tracking with Inputlog.

4 IELTS Task 2 writing band descriptors (public version) are available at: http://takeielts.britishcouncil.org/sites/default/files/IELTS_task_2_Writing_band_descriptors.pdf

5 When relying on automatic machine identification and analysis of pauses, for example, using Inputlog, a minimal pause threshold (e.g., 2 seconds) tends to be used for operationalization purposes; this was not necessary for us because we coded manually.

References

Benedek, M., Stoiser, R., Walcher, S., & Körner, C. (2017). Eye behavior associated with internally versus externally directed cognition. *Frontiers in Psychology*, *8*, 1092–1101. https://doi.org/10.1016/j.concog.2017.06.009

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Mahwah, NJ: Lawrence Erlbaum.

Bernard, H. R., & Ryan, G. W. (2010). *Analyzing qualitative data: Systematic approaches*. London: Sage.

Couronné, T., Guérin-Dugué, A., Dubois, M., Faye, P., & Marendaz, C. (2010). A statistical mixture method to reveal bottom-up and top-down factors guiding the eye-movements. *Journal of Eye Movement Research*, *3*, 1–13.

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

Degenhardt, M. (2006). CAMTASIA and CATMOVIE: Two digital tools for observing, documenting and analysing writing processes of university students. In L. Van Waes, M. Leijten, C. Neuwirth, & G. Rijlaarsdam (Eds.), *Writing and digital media* (pp. 180-186). Amsterdam: Elsevier.

De Smet, M., Leijten, M., & Van Waes, L. (2014, April). *Inputlog 6.0: Eye tracking. Defining reading during writing*. Paper presented at the Keystroke Logging Training School, Antwerp, Belgium.

Dörnyei, Z., & Katona, L. (1992). Validation of the C-test amongst Hungarian EFL learners. *Language Testing*, *9*, 187–206. http://doi.org/10.1177/026553229200900206

Dussias, P. E. (2010). Uses of eye-tracking data in second language sentence processing research. *Annual Review of Applied Linguistics*, *30*, 149–166. https://doi.org/10.1017/S026719051000005X

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2017). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 1-26. http://doi.org/10.1017/S0272263117000109

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*, 365–387. http://doi.org/10.2307/356600

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. G. Steinberg & E. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 31–50). Hillsdale, NJ: Erlbaum.

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

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum.

Gilmore, A. (2011). "I prefer not text": Developing Japanese learners' communicative competence with authentic materials. *Language Learning*, 61, 786–819. http://doi.org/10.1111/j.1467-9922.2011.00634.x

Gilmore, A. (2015). Language learning in context: Complex dynamic systems and the role of mixed methods research. In J. King (Ed.), *The dynamic interplay between context and the language learner*. Basingstoke, UK: Palgrave MacMillan, 194–224.

Grabe, W. (2001). Notes toward a theory of second language writing. In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 39–57). Mahwah, NJ: Lawrence Erlbaum.

Grotjahn, R. (2010). *The C-test: Contributions from current research*. Frankfurt, Germany: Peter Lang.

Hamel, M.-J., & Séror, J. (2016). Video screen capture to document and scaffold the L2 writing process. In C. Caws & M.-J. Hamel (Eds.), *Language-learner computer interactions* (pp. 137–162). Amsterdam: John Benjamins.

Hamel, M.-J., Séror, J., & Dion, C. (2015). *Writers in action: Modelling and scaffolding second-language learners' writing process*. Toronto, ON: Higher Education Quality Council of Ontario.

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

Hyland, K. (2009). *Teaching and researching writing* (2nd ed.). Harlow: Longman Pearson.

Janssen, D., Van Waes, L., & van den Bergh, H. (1996). Effects of thinking aloud on writing processes. In M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 233–250). Mahwah, NJ: Erlbaum.

Klein-Braley, C., & Raatz, E. (1984). A survey on the C test. *Language Testing*, *1*, 134–146. http://doi.org/10.1007/978-1-4899-0870-4_18

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159–174. http://doi.org/10.2307/2529310

Larsen-Freeman, D. & Cameron, L. (2008). *Complex systems and applied linguistics*. Oxford: Oxford University Press.

Latif, M. M. A. (2008). A state-of-the-art review of the real-time computer-aided study of the writing process. *IJES, International Journal of English Studies*, *8*, 29–50.

Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of on-line writing. In K. P. H. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing: Methods and applications* (pp. 73–94). Oxford: Elsevier.

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

Leijten, M., Van Waes, L., Schriver, K., & Hayes, J. R. (2014), Writing in the workplace: Constructing documents using multiple digital sources. *Journal of Writing Research*, 3, 285–337. http://doi.org/10.17239/jowr-2014.05.03.3

Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. New York: Routledge.

Manchón, R. M., Roca de Larios, J., & Murphy, L. (2007). A review of writing strategies: Focus on conceptualizations and impact of first language. In A. D. Cohen & E. Macaro (Eds.), *Language learner strategies* (pp. 229–250). Oxford: Oxford University Press.

Manchón, R. M., Roca de Larios, J., & Murphy, L. (2009). The temporal dimension and problem-solving nature of foreign language composing processes: Implications for theory. In R. M. Manchón (Ed.), *Writing in foreign language contexts: Learning, teaching, and research* (pp. 102–129). Clevedon, UK: Multilingual Matters.

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

Park, K., & Kinginger, C. (2010). Writing/thinking in real time: Digital video and corpus query analysis. *Language Learning & Technology*, *14*, 31–50.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly*, *19*, 229. http://doi.org/10.2307/3586828

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

Rijlaarsdam, G., & 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 C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences & applications* (pp. 107–125). Mahwah, NJ: Lawrence Erlbaum.

Roberts, L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and L2 processing. *Studies in Second Language Acquisition*, 35, 213-235. http://doi.org/10.1017/S0272263112000861

Roca de Larios, J., Marín, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language Learning*, *51*, 497–538. http://doi.org/10.1111/0023-8333.00163

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*, 30–47. http://doi.org/10.1016/j.jslw.2007.08.005

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, *9*, 259–291. http://doi.org/10.1016/S1060-3743(00)00028-X

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning*, *46*, 137–174. http://doi.org/10.1111/j.1467-1770.1996.tb00643.x

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, yeaching, and research* (pp. 77–101). Clevedon, UK: Multilingual Matters.

Schoonen, R., van Gelderen, A., de Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, R. (2003). First language and second language writing: The role of linguistic knowledge, speed of processing, and metacognitive knowledge. *Language Learning, 53*, 165–202. http://doi.org/10.1111/1467-9922.00213

Séror, J. (2013). Screen capture technology: A digital window into students' writing processes/Technologie de capture d'écran: une fenêtre numérique sur le processus d'écriture des étudiants. *Canadian Journal of Learning and Technology/La Revue canadienne de l'apprentissage et de la technologie*, *39*, 1–16. Retrieved from https://www.cjlt.ca/index.php/cjlt/article/view/26305/19487

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. http://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*. Utrecht, The Netherlands: LOT. Retrieved from https://www.lotpublications.nl/writing-in-first-and-second-language-writing-in-first-and-second-language-empirical-studies-on-text-quality-and-writing-processes

van den Bergh, H., & Rijlaarsdam, G. (1999). The dynamics of idea generation during writing. In M. Torrance & D. Galbraith (Eds.) *Studies in writing* (pp. 99–121). Amsterdam: Amsterdam University Press.

van den Bergh, H., & Rijlaarsdam, G. (2001). Changes in cognitive activities during the writing process and relationships with text quality. *Educational Psychology*, *21*, 373–385. http://doi.org/10.1080/01443410120090777

van den Haak, M., de Jong, M., & Schellens, P. J. (2003). Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology*, *22*, 339–351. http://doi.org/10.1080/0044929031000

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

Van Waes, L., Leijten, M., Lindgren, E., & Wengelin, A. (2016) (2nd ed.). Keystroke logging in writing research. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.) *Handbook of writing research* (pp. 410–426). New York: The Guildford Press.

Van Waes, L., Leijten, M., & Neuwirth, C. M. (Eds.). (2006). *Writing and digital media* (Vol. 17). Oxford: Elsevier.

van Weijen, D. (2009). *Writing processes, text quality, and task effects: Empirical studies in first and second language writing.* Utrecht, The Netherlands: LOT. Retrieved from https://dspace.library.uu.nl/handle/1874/33624

van Weijen, D., van den Bergh, H., Rijlaarsdam, D., & Sanders, T. (2009). L1 use during L2 writing: An empirical study of a complex phenomenon. *Journal of Second Language Writing*, *18*, 235–250. https://doi.org/10.1016/j.jslw.2009.06.003

Weir, C. (1990). *Communicative language testing*. New York: Prentice Hall.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In N. Calzolari, K. Choukri, A. Gagemi, B. Maegaard, J. Mariani, J. Odijk, & D. Tapias (Eds.), *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 1556-1559). Retrieved from www.lrec-conf.org/proceedings/lrec2006/pdf/153_pdf.pdf

Wong, A. T. Y. (2005). Writers' mental representations of the intended audience and of the rhetorical purpose for writing and the strategies that they employed when they composed. *System*, *33*(1), 29–47. https://doi.org/10.1016/j.system.2004.06.009

Zamel, V. (1983). The composing processes of advanced ESL students: Six case-studies. *TESOL Quarterly*, *17*, 165–187. http://doi.org/10.2307/3586647

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. L2 proficiency and Essay Quality Measures.

Appendix S2. Essay Samples of the Highest (P01) and Lowest (P22) L2 Proficiency Participants.

Appendix S3. Summary Statistics for Main Effect of Process.

Appendix S4. Descriptive Statistics for Episode Frequency and Duration, by Period and Process.

Appendix S5. Summary Statistics for Interactive Effect of Process and Period.

Appendix S6. Summary of Analyses Using Episode Duration and Frequency as Predictors of Essay Quality and Length.

Table 1 Variation in L2 writing process taxonomies

Researcher	Taxonomy
Raimes (1985)	Assessing, commenting, editing, planning structure, questioning, reading, repeating, writing, revising, silence, and writing
Raimes (1987); Wong (2005)	Metacognitive categories, such as questioning, goal setting, or self-assessment
Cumming (1989)	Attention to aspects of writing: language use, discourse organization, gist, intention, procedures for writing
	Problem-solving strategies: problem identification, engaging in a search routine, generating and assessing alternatives, assessing criterion, directed questions, and setting/adhering to a goal
Sasaki (2000)	Main categories: planning, retrieving, generating ideas, verbalizing, translating, rereading, evaluating, and others

Table 2 Differences between the five processes in frequency of episodes in each separate period

Period	*F*	*df*	*p*	η_p^2
1	55.23	2.66, 55.90	< .001	.725
2	23.26	3.00, 62.90	< .001	.526
3	18.43	4, 84	< .001	.467
4	7.60	4, 84	< .001	.266
5	10.72	3.00, 62.90	< .001	.338

Table 3 Differences between the five periods in frequency of episodes for each separate process type

	Differences between periods				Linear trend across periods 1 to 5			
Process type	*F*	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Text construction	6.82	4, 84	< .001	.245	21.66	1, 21	< .001	.508
Revising	1.31	4, 84	.274	.059	2.09	1, 21	.163	.091
Pausing	2.73	2.34, 49.07	.067	.115	0.48	1, 21	.497	.022
Rereading	9.74	2.88, 60.50	< .001	.317	22.60	1, 21	< .001	.518
Using external resources	1.84	4, 84	.128	.081	1.42	1, 21	.247	.063

Table 4 Differences between the five processes in total duration of episodes in each separate period

Period	*F*	*df*	*p*	η_p^2
1	15.26	1.98, 41.60	< .001	.421
2	7.24	1.98, 41.60	< .001	.256
3	4.21	4, 84	< .01	.167
4	1.98	4, 84	.104	.086
5	3.11	1.98, 41.60	.047	.129

Table 5 Differences between the five periods in total time duration of episodes for each separate process type

	Differences between periods				Linear trend across periods from 1 to 5			
Process type	*F*	*df*	*p*	η_p^2	*F*	*df*	*p*	η_p^2
Text construction	6.89	4, 84	< .001	.247	22.77	1, 21	< .001	.520
Revising	1.72	2.19, 45.93	.187	.076	0.41	1, 21	.530	.019
Pausing	1.61	2.96, 62.06	.196	.071	3.39	1, 21	.080	.139
Rereading	12.52	2.08, 43.63	< .001	.373	19.51	1, 21	< .001	.482
Using external resources	1.90	2.75, 57.81	.145	.083	0.01	1, 21	.971	.000

Table 6 Correlations of proficiency with episode frequency and duration by process type

	Episode frequency		Episode duration	
Process type	*r*	*p*	*r*	*p*
Text construction	.253	.013	.219	.022
Revising	.230	.016	.080	.408
Pausing	–.158	.099	–.076	.428
Rereading	.002	.979	.154	.108
Using external resources	–.342	< .001	–.321	< .010

Table 7 Correlations of episode frequency and duration with essay length by process type

	Episode frequency		Episode duration	
Process type	*r*	*p*	*r*	*p*
Text construction	.194	.043	.314	< .010
Revising	.517	< .001	.383	< .001
Pausing	–.293	< .010	–.178	.063
Rereading	–.055	.569	.105	.277
Using external resources	–.324	< .001	–.488	< .001

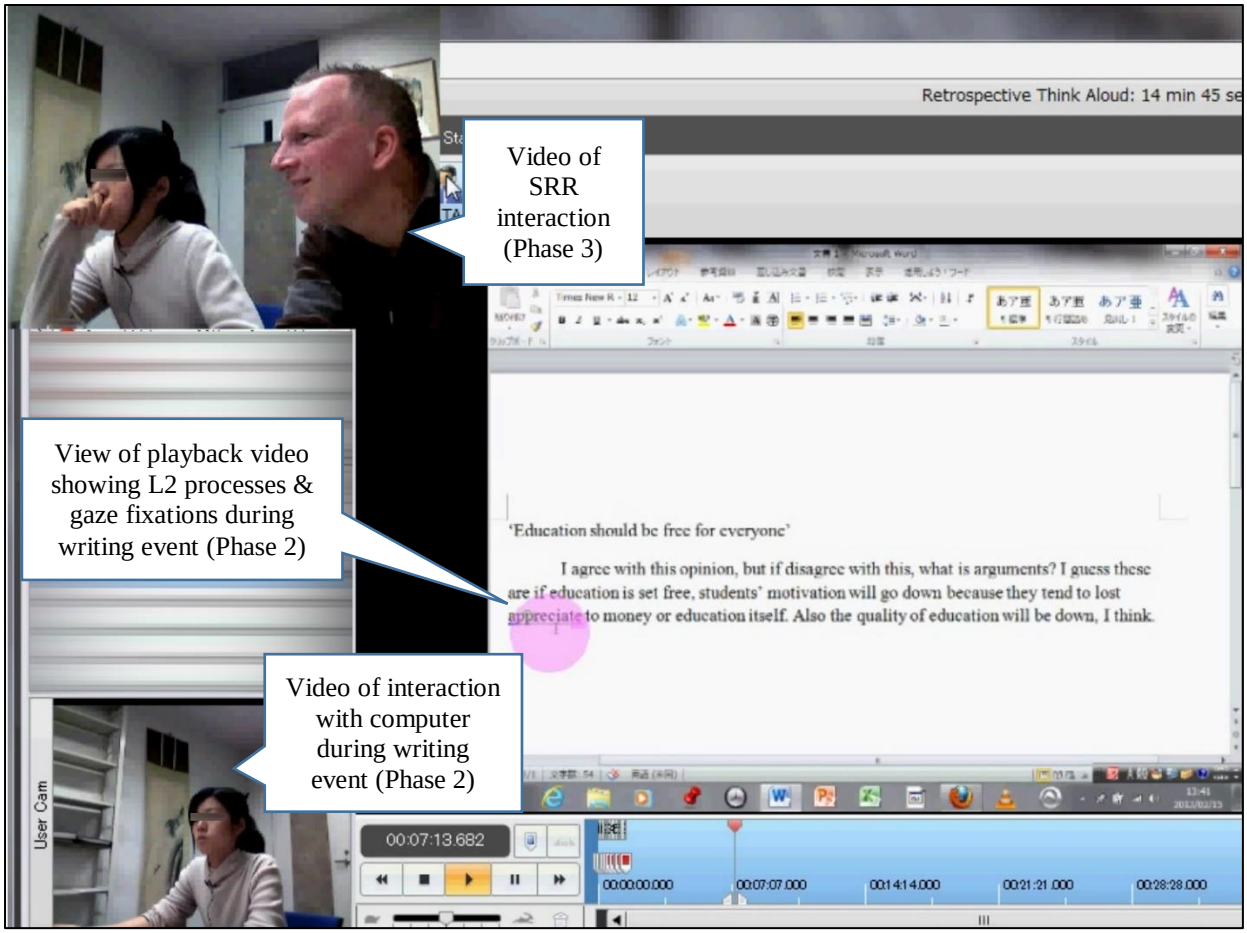

Figure 1 Overview of visual data available for analysis.

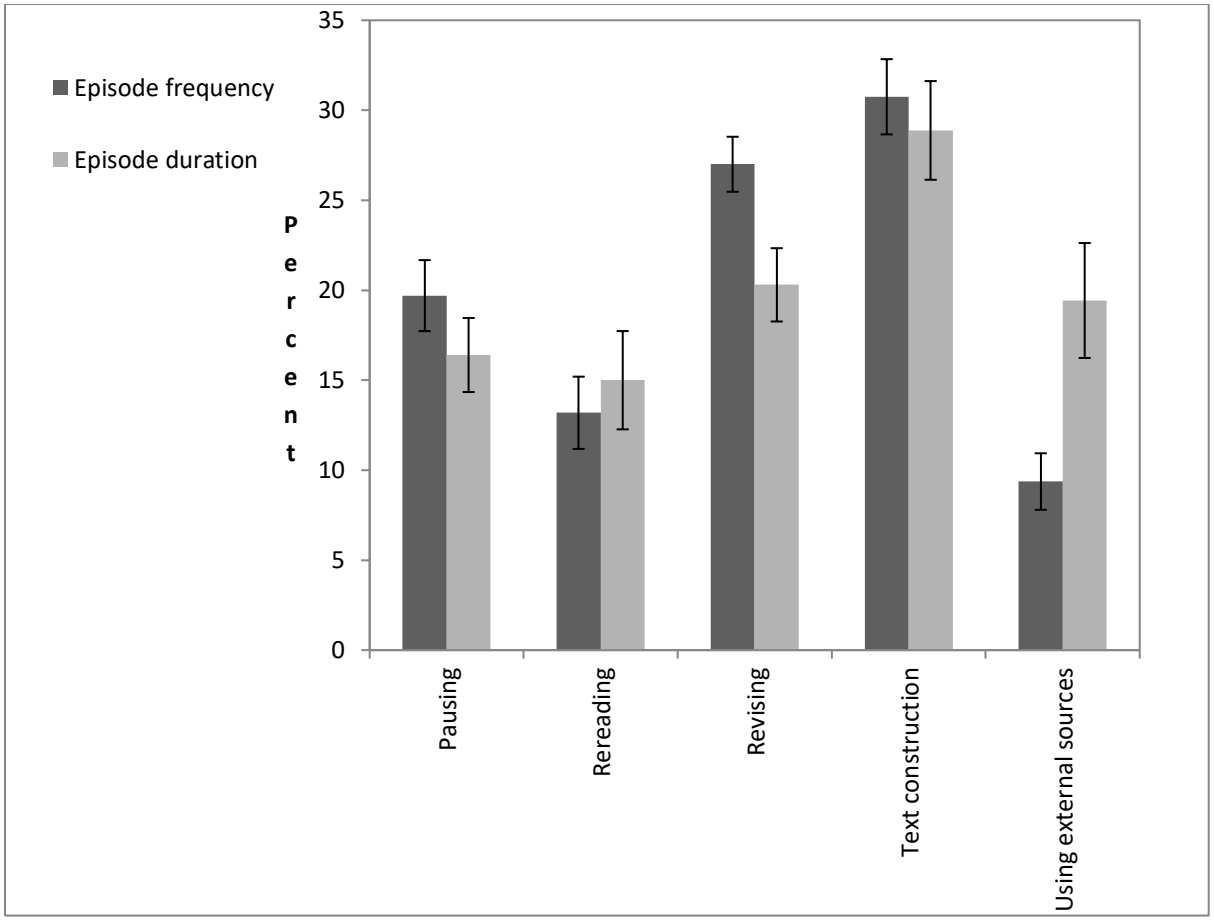

Figure 2 Frequencies and duration of episodes for each process type. Means over all periods are combined. Error bars include 95% confidence intervals.

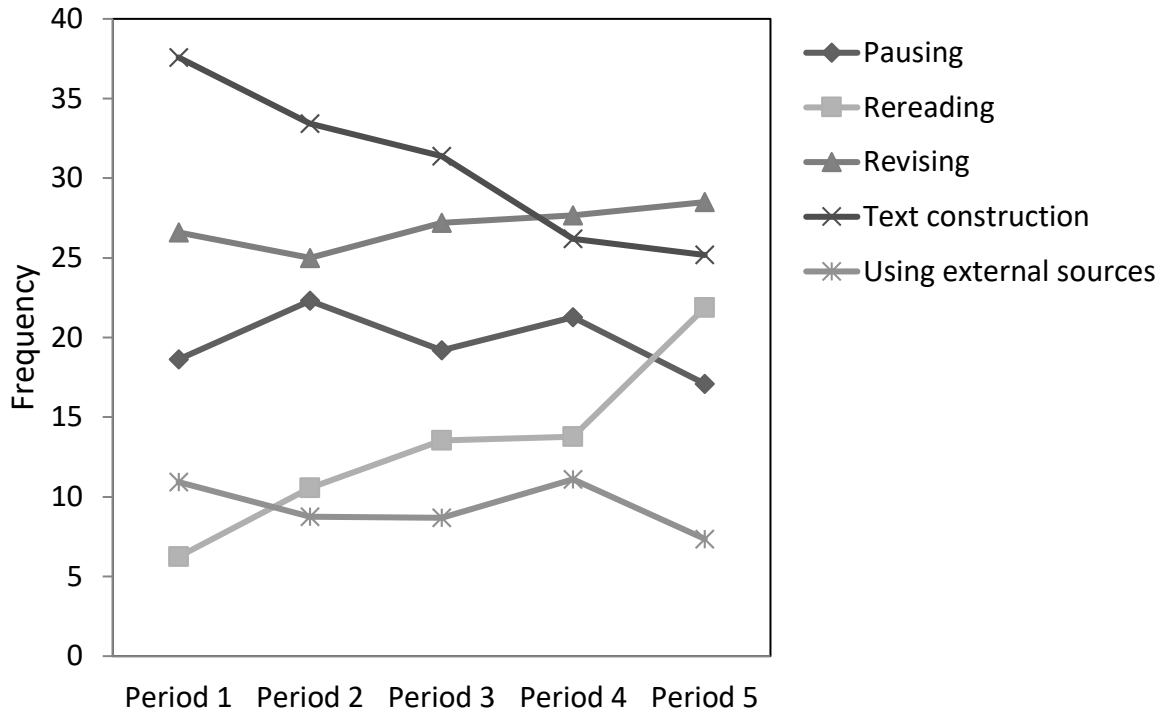

Figure 3 Frequency of episodes of each process type across successive time periods.

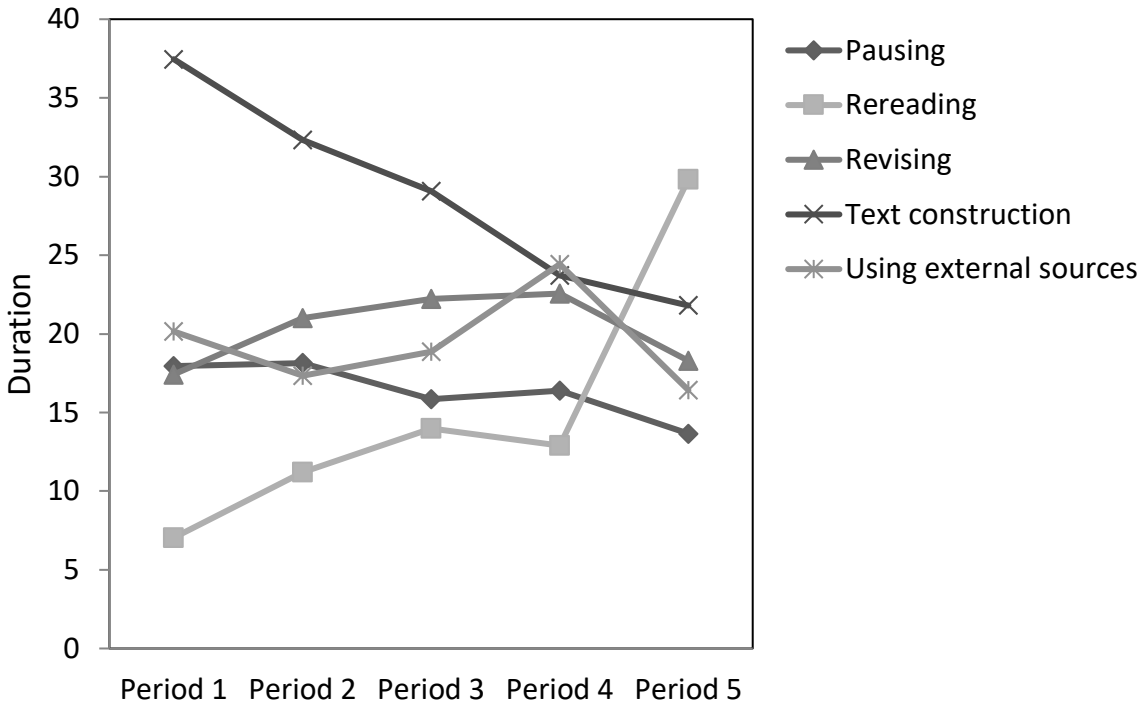

Figure 4 Total duration of episodes of each process type across successive time periods.

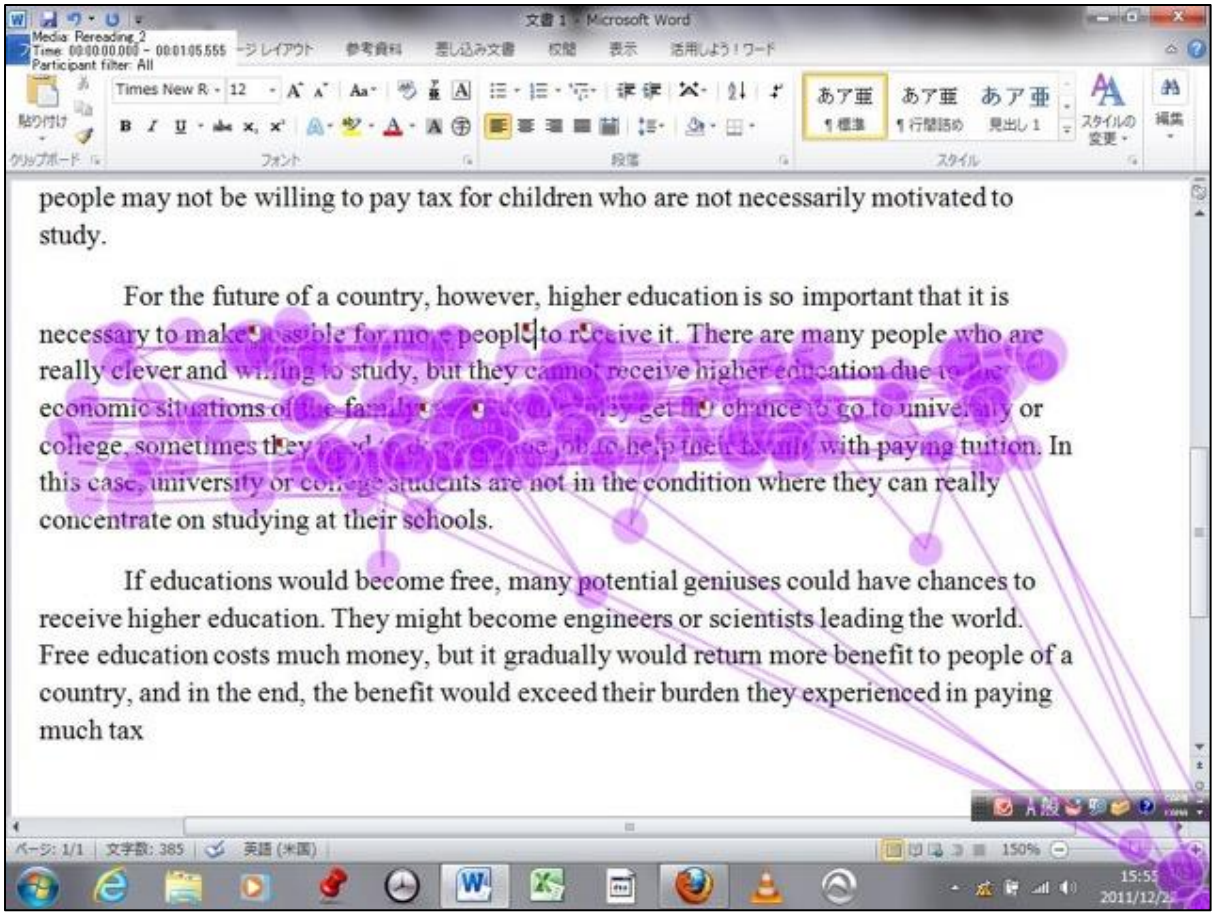

Figure 5 Intense rereading burst as shown by the eye-tracking data in a digital screen capture.

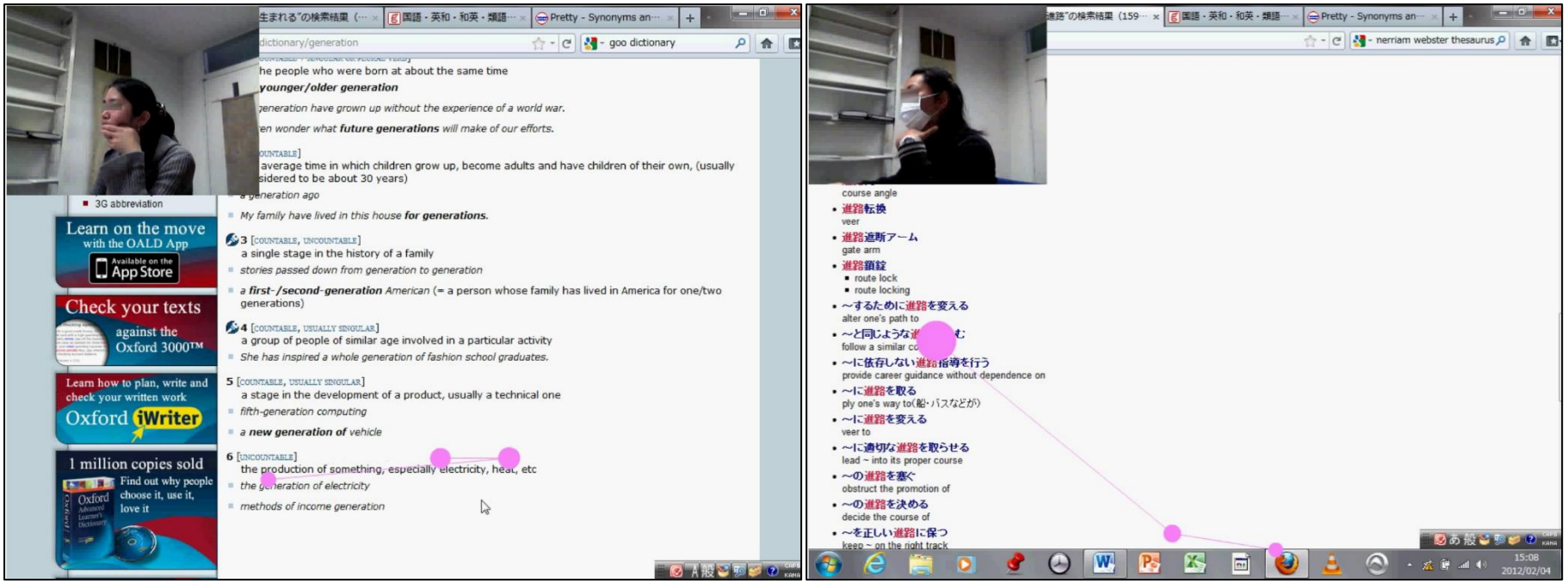

Figure 6 Screenshots of online behaviour for P01 (searching a monolingual dictionary for "generation") and P22 (searching a bilingual dictionary for *shinro*, "career options/paths").

High L2 proficiency (P01)			Time (secs)		Low L2 proficiency (P22)		
Action	**Outcome**	**Commentary**	**P01**	**P22**	**Action**	**Outcome**	**Commentary**
1. Query for *hassei* (accrual/emergence) in bilingual dictionary	Unsuccessful search	P01: it has lots of meanings but it's like the development of our feelings or something although I think I looked at a lot of definitions here	0	0	1. Query for *shinro* (career options/paths)		Tutor: what are you reading here? P22: *shinro* many many Tutor: many choices yeah
2. Skips to page 8 in the search results for *hassei*	Unsuccessful search	Tutor: skipped to page eight? P01: because I do that often and I thought that the first the second page I looked at was using *hassei* in a different sort of meaning from I was looking so I skipped to see how it's used	32	31	2. Multiple rereadings of one particular example using 'path' in the bilingual dictionary (indicated by eye gaze data and cursor movements)		P22: *kore dake* (just this one) only I think Tutor: do you think this is a good meaning? P22: mm mm mm
3. Second query, *umareru* (accrual/ emergence)	Unsuccessful search	P01: I changed it again to *umareru hassei* is more difficult word than *umareru* so I thought *umareru* might have a better translation than *hassei*	48	56	3. Movement of cursor to Word icon at the bottom of the screen (indicated by eye gaze data)		
4. Return to Word to write 'gene' (incomplete form of generation)	Triggers new search in monolingual dictionary	P01: [triggered by DSC video showing text construction of 'gene'] oh and then I thought of the verb generate and then I changed it to generation but I thought that generation had different meanings than generate	82	57	4. Returns to examining the translations for *shinro* in the bilingual dictionary, focusing on examples using 'course'		P22: *kono hen wa sugoi atteru* (these are really appropriate) good meaning
5. Query for 'generation' in monolingual dictionary	Successful search	P01: it says that the production of something especially electricity so I thought it wouldn't fit in my [essay] So I thought I couldn't use the word generation here	88	135	5. After eye gaze fixation on the word 'course', she quickly returns to Word to insert lexical item	Inappropriate text revision	
6. Return to word processor to delete 'gene'	Successful text revision	P01: then I couldn't find the right word so I think I use a different word [laughs]	113				

Figure 7 Online strategic behavior for highest and lowest proficiency students (episodes of nearly equal duration)