ISSN 2201-2982
2017/3

IELTS Research Reports Online Series

Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory

Andrea Révész, Marije Michel and Minjin Lee

IELTS™
BRITISH COUNCIL
idp IELTS AUSTRALIA
CAMBRIDGE ENGLISH Language Assessment
Part of the University of Cambridge

Funding

This research was funded by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. The grant was awarded in 2014-15.

Publishing details

Published by the IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia © 2017.

Introduction

This study by Andrea Révész of University College London and her colleagues was conducted with support from the IELTS partners (British Council, IDP: IELTS Australia, and Cambridge English Language Assessment), as part of the IELTS joint-funded research program. Research funded by the British Council and IDP: IELTS Australia under this program complement those conducted or commissioned by Cambridge English Language Assessment, and together inform the ongoing validation and improvement of IELTS.

A significant body of research has been produced since the joint-funded research program started in 1995, with over 110 empirical studies receiving grant funding. After undergoing a process of peer review and revision, many of the studies have been published in academic journals, in several IELTS-focused volumes in the *Studies in Language Testing* series (http://www.cambridgeenglish.org/silt), and in *IELTS Research Reports*. Since 2012, in order to facilitate timely access, individual research reports have been made available on the IELTS website immediately after completing the peer review and revision process.

When language tests require test-takers to engage the same processes and produce the same products as they would in the real world, it makes it easier to determine that they indeed have the language skills needed. The study detailed in this report provides evidence of that, investigating the cognitive processes involved in producing IELTS Academic Writing Task 2 responses.

Mental processes cannot be observed directly, of course, and for many years, researchers depended on self-reports to gain insight into these. New tools have become available more recently, however, such as eye-tracking and keystroke-logging technology, which capture external behaviour that can provide more clues about internal processes. The present study is unique in being the first to combine these different methodologies—in addition to a battery of working memory tests—in order to develop a well-triangulated view of what goes on in candidates' heads while doing one part of the IELTS Writing test.

The study found that test-takers' writing processes—from planning to execution to monitoring—reflect those of L1 writers and are aligned with the focus of the assessment. That is, evidence in support of the cognitive validity of the IELTS Writing test.

But it is important to go beyond the headline finding to see the insights that the new methodologies make possible. For example, writers sometimes pause during the process of writing, and the researchers were able to distinguish different types of pauses, determined by where the writer was looking during that period of time, and the impact this had on the writer's subsequent production. Candidates who looked off-screen during pauses produced syntactically less complex sentences, whereas those who focused on the task instructions produced more complex structures. It is not difficult to think about or infer the different processes accompanying each behaviour above, but it takes the combination of methodologies used to provide evidence of these differences.

This report is very much worth reading, then, not just because of what it shows about the cognitive validity of the IELTS Writing test, but also for the way it demonstrates a fruitful way forward for the conduct of studies in this area. This study merely scratches the surface, and we look forward to the depths of insight that studies such as this will bring us in the future.

Dr Gad Lim
Principal Research Manager
Cambridge English Language Assessment

Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory

Abstract

This project examined the cognitive processes and online behaviours of second language writers while performing IELTS Academic Writing Test Task 2, and the ways in which the online behaviours of test-takers relate to the quality of the text produced. An additional aim was to assess whether writing behaviours and text quality are influenced by individual differences in phonological short-term memory and executive control functions.

Thirty participants, Mandarin users of L2 English from a UK university, performed a version of Task 2 of the IELTS Academic Writing Test. The online writing processes of the participants were captured by recordings of participants' eye-movements and logs of their keystrokes. After a short break, a subset of the participants took part in a stimulated recall session, as part of which participants were requested to describe their thought processes during task performance, prompted by the playback of the recordings of their keystrokes. Participants were administered an extensive battery of working memory tests (Chinese Digit Span, Chinese Non-word Span, Colour Shape Task, Corsi Block Forward-Backward, Stop Signal Task, and Operation Span).
The essays produced were scored in terms of IELTS rating criteria, and analysed for linguistic complexity (lexical, syntactic and discourse complexity) and accuracy relying on computer-based and expert analyses.

The results demonstrated that the IELTS Academic Writing Task 2 elicits a wide range of cognitive processes and writing behaviours, which are well aligned with the intended aim of the IELTS Academic Writing test. A number of links were also observed between the measures of writing behaviours and text quality, some of which included the IELTS total score and subscores. However, working memory was found to be related to only a few measures of writing behaviours and text quality indices.

Authors' biodata

Andrea Révész

Andrea Révész is a Senior Lecturer in Applied Linguistics at the UCL Institute of Education, University College London. Her research interests lie in the areas of second language acquisition, and second language instruction and assessment. Within these areas, her main interests include: cognitive aspects of second language acquisition, role of individual differences, and task-based language teaching and assessment (with particular emphases on speaking, listening, and writing). Her work has appeared in journals such as *Applied Linguistics, Applied Psycholinguistics, Language Awareness, Language Learning, The Modern Language Journal, Studies in Second Language Acquisition* and *TESOL Quarterly*. Andrea serves as Associate Editor of the journal, *Studies in Second Language Acquisition,* and is Vice-President of the International Association for Task-based Language Teaching (TBLT).

Marije Michel

Marije Michel is a Lecturer in Language Teaching at Lancaster University. She is interested in second language acquisition and assessment in multilingual educational settings. In particular, she investigates cognitive and interactive aspects of task-based language pedagogy with an emphasis on the role of task complexity and online peer interaction for second language development. Recently, she also started using eye-tracking methodology to investigate second language writing processes during synchronous computer-mediated communication. As a post-doctoral fellow, she co-designed and -validated the German online language training competence assessment tool SprachKoPF. Marije has published in *Studies in Second Language Acquisition, the European Journal of Applied Linguistics, Language Awareness, The Modern Language Journal, the International Review of Applied Linguistics* and *Language and Cognitive Processes*.

Minjin Lee

MinJin Lee is a doctoral candidate in Applied Linguistics at the UCL Institute of Education, University College London. Her research interests include second language acquisition, instructed second language acquisition, and task-based language teaching, with particular emphasis on the role of cognitive individual difference factors in learning a second language.

Table of contents

List of tables

1 Introduction

The end products of writing tasks have been the object of a considerable amount of research in the areas of second language (L2) assessment and second language acquisition (e.g., Cushing Weigle, 2002; Polio, 2012, for reviews). However, relatively little empirical research exists that examines the cognitive processes and writing behaviours in which L2 users engage while performing writing tasks in second language testing or instructed settings (Révész, 2014). So far, it has also been underexplored how the cognitive processes and writing behaviours in which L2 writers engage may relate to the quality of the end products of writing and how they might be influenced by individual differences in working memory capacity.

The aim of this study was to bring together these three areas of language testing and learning: research on the L2 writing process; the L2 writing product; and the role of individual differences in cognitive abilities. To address these goals, we examined the cognitive processes and online behaviours of second language writers with first language (L1) Mandarin background while performing one version of Task 2 of the IELTS Academic Writing Test, and the ways in which the cognitive processes and online behaviours of test-takers might relate to the quality of the texts produced. We also assessed whether the nature of the writing process and the writing product are influenced by individual differences in various components of working memory capacity. We utilised an innovative combination of research methods, employing eye-tracking methodology, online keystroke logging, retrospective stimulated recall, and computer-based text analyses.

In addition to contributing to research on L2 writing and assessment, the project aimed to help establish the cognitive validity (Shaw & Weir, 2007) of Task 2 of the IELTS Academic Writing test. Cognitive validity is concerned with "whether the tasks proposed by a test designer elicit mental processes resembling those which a language user would actually employ when undertaking similar tasks in the world beyond the test" (Field, 2011, p. 67). Field (2009) suggests two ways in which cognitive validity may be established. First, researchers can compare the processes in which L2 users engage under testing conditions with those that L2 users adopt under non-testing conditions. Second, native speaker real-life performances can be set as a criterion against which the processes in which L2 users engage are compared.

This study adopted the second approach by comparing the cognitive processes of L2 users performing a version of Task 2 of the IELTS Academic Writing test with the processes native writers employ when carrying out real-life writing tasks, as described in Kellogg's (1996) well-established model of writing.

2 Background

2.1 Investigating second language writing processes

Kellogg's (1996) model of writing views writing as an interactive process, which involves three sub-processes: formulation, execution, and monitoring. Formulation entails the planning of content and translating content into linguistic form. While planning, writers typically retrieve ideas from long-term memory or from the task input, and then devise a coherent plan for the text content. Translating ideas into linguistic form involves three key sub-processes: lexical retrieval, syntactic encoding, and expressing cohesion. In the execution phase, a handwritten or typed text is produced using motor movements. The last stage, monitoring, ensures that the text produced is an appropriate reflection of the writer's intended content. If discrepancies are identified between the text and the content planned, then L2 writers carry out revisions. The stages of formulation, execution, and monitoring constantly interact, resulting in a complex array of cognitive operations.

There is substantial amount of research investigating the processes in which L1 writers engage, and the results overall confirm the writing stages outlined in Kellogg's model. However, considerably less research has been conducted on L2 writing processes and how these may be linked to the outcomes of writing. Additionally, the small amount of research available has typically utilised a single method to tap writing processes, instead of triangulating a variety of sources to increase construct validity. For example, in some studies, researchers have relied solely on introspective protocols such as the think-aloud procedure to explore the cognitive processes in which L2 writers engage (e.g., Roca de Larios, Manchon, Murphy & Marin, 2008). Other researchers have exclusively utilised online computer recording of L2 writers' keystrokes and mouse movements to obtain information about online writing processes (see Leijten & Van Waes, 2013; Spelman Miller, Lindgren & Sullivan, 2008 for reviews).

Although these studies have yielded useful insights, there are clear advantages to combining various data sources in investigating writing processes (Leijten & Van Waes, 2013; Wengelin et al., 2009). For example, by triangulating data from keystroke-logging and eye-tracking methodology, researchers can not only observe the writing behaviour of language users, but also the reading activities in which writers engage. Among other things, information about writers' eye-movements might help reveal causes for pausing, such as re-reading the writing prompt or the text already produced. Clearly, the integration of these two types of data have the potential to capture the writing process more fully and, thereby, allow for making more valid inferences about the underlying cognitive processes involved in writing. Despite the advantages of combining eye-tracking and keystroke-logging, this approach still entails an important limitation: it affords no direct insights into the conscious cognitive operations of the writers during task performance. This issue could potentially be addressed by triangulating introspective protocols with eye tracking and keystroke-logging methodology.

To the best of our knowledge, these three methods (keystroke logging, eye tracking, and introspection) have not yet been utilised together in the context of L2 writing research and L2 testing. However, a small number of studies exist that have successfully triangulated introspective and keystroke-logging data to capture cognitive operations during L2 writing. For example, Stevenson, Schoonen and de Glopper (2006) used keystroke-logging together with the think-aloud procedure to test the hypothesis that, in the foreign language writing of secondary school students, attention to linguistic processes may inhibit higher level conceptual processing. Van Weijen (2009) employed the same combination of methods to compare online L1 and L2 writing processes, with a view to contrasting the cognitive activities of planning, generating ideas, and formulation when composing in one's native as compared to the second language. In both studies, the use of combined data sources allowed the researchers to arrive at more detailed and accurate - thus more valid - descriptions of the cognitive activities of L2 writers. Similarly, in the context of L2 reading, Bax (2013) and Brunfaut and McCray (2015) found that the eye-tracking and introspective methodology can be effectively utilised together to explore the cognitive validity of the IELTS and Aptis reading tests respectively.

Against this background, one aim of this research was to explore, via the joint application of eye-tracking, keystroke-logging, and introspective methodology, the cognitive writing processes in which test-takers engage when performing Task 2 of the IELTS Academic Writing Test. Our intention was twofold. First, we aimed to further our understanding of L2 writers' processing behaviours. Second, our goal was to confirm the cognitive validity of this assessment by comparing the processes in which test-takers engage with those posited in Kellogg's well-established model of writing.

2.2 The second language writing process and product

A second aim of the present study was to explore potential links between the processes in which L2 writers engage and the outcomes of their writing. While this relationship has received considerable attention in L1 writing research (e.g., Breetvelt, Van den Bergh & Rijlaarsdam, 1994), so far only a few studies have been dedicated to exploring this association in the context of L2 writing. Also, the results of the existing research are mixed. Stevenson et al. (2006) examined whether type of revision behaviour predicts text quality. The participants were 22 secondary school students, who wrote two essays in L1 Dutch and two essays in L2 English. It was expected that, for L2 writing, there would emerge a negative relationship between lower-level revisions (word-level changes) and quality of text content, since L2 writers are likely to allocate more attention to lower-level writing processes, resulting in less attention left to be dedicated to higher-level cognitive operations including revisions. This prediction, however, was not borne out. The researchers found no relationship between revision type and text quality.

Spelman Miller, Lindgren and Sullivan (2008) also set out to investigate whether writing behaviours predict text quality. They looked, not only into revision behaviours, but also pausing and fluency. The participants were high school L2 English writers with Swedish as first language. The study took three years, with the researchers collecting one writing piece from each student participant each year. The speed of writing was expressed in terms of fluency (number of words per minute), burst (number of typed characters between pauses and/or revisions), and fluency during burst (total writing time between pauses and/or revisions). Pausing was assessed by calculating mean pause length and pause time (proportion of pausing to total writing time), where the threshold for pausing was defined as two seconds. Amount of revision (deletions or insertions) was also examined. Text quality was determined in terms of weighted subscores computed for content, grammatical and lexical range, accuracy, and fluency. Two fluency measures, burst and fluency during burst, were identified as strong predictors of text quality, but none of the pausing or revision indices accounted for significant variation in text quality.

To sum up, previous research indicates that text quality is related to fluency but not to revision or pausing. Clearly, more research is warranted to further explore the generalisability of these patterns.

2.3 Working memory and second language writing

The third aim of this research was to investigate the extent to which individual differences in working memory are related to the nature of the cognitive operations involved in L2 writing and the quality of the written text produced. One rationale for investigating individual difference variables in relation to these constructs is that any links (or lack of them) can help deduce information about the cognitive processes which were in operation during the actual writing performance. As DeKeyser (2012, p. 190) explains, one way to understand cognitive "processes which are hard or impossible to observe is to infer them from the way individual difference variables interact with linguistic...variables". In the context of language assessment, this would appear to suggest that research examining associations between test performance and individual differences in cognitive abilities, such as working memory, can assist in establishing cognitive validity. In particular, if a relationship is found between working memory indices, writing behaviours, or text quality indices, this helps make inferences about the processes in which writers engage during test performance. For example, if task-switching ability (an executive function) is found to have a link with measures of writing behaviours and text quality, this implies that test-takers needed to rely on this ability while carrying out the testing task.

The most widely accepted model of working memory today was developed by Baddeley and Hitch (1974). This model defines a multi-component memory system comprising a central executive and two domain-specific sub-systems, the phonological loop and the visual-spatial sketchpad. Later, a fourth component, the episodic buffer, was added (Baddeley, 2000). The phonological loop is responsible for the temporary retention and manipulation of verbal information, whereas the visual-spatial sketchpad is specialised for storing and handling visual and spatial information. The central executive controls complex cognitive operations, such as: focusing, dividing and switching attention; activating and inhibiting processing routines; and regulating the information flow from the short-term storage subsystems and from long-term memory. The episodic buffer integrates multi-dimensional information to form episodes. The phonological loop, the visual-spatial sketchpad, and the central executive are all limited in capacity.

The role of individual differences in working memory capacity has been the subject of a growing number of studies in the field of SLA (e.g., Kormos & Sáfár, 2008; Révész, 2012; see Williams, 2012 for a review). Yet, only a very limited amount of research has looked into the relationship between working memory and second language writing. As Kormos (2012) notes, this lack of attention is surprising because the success of various stages of the writing process is heavily reliant on the availability of adequate working memory resources. For example, greater phonological short-term memory (PSTM) span is likely to assist in forming longer and more complex syntactic structures, since PSTM determines the amount of verbal information one can store in memory.

Strong visual-spatial short-term memory may benefit planning and editing processes while composing (Kellogg, 1996), since it is responsible for storing visual and spatial units. Finally, individuals with superior central executive will probably better handle increased demands on parallel processing when various stages of writing (e.g., planning and typing) run simultaneously. Although writing tends to involve less time pressure and pose fewer demands on parallel processing than producing speech, certain components of the writing process are still likely to operate in a parallel fashion and thus need to be coordinated (Kormos, 2012).

The few studies that have investigated whether working memory is related to L2 writing success confirmed a role for working memory. Kormos and Sáfár (2008) revealed that scores achieved by L2 learners in the writing section of the Cambridge First Certificate Examination had moderate, positive correlations with their phonological short-term memory spans. However, the same scores were not found to have a relationship with complex working memory capacity. Similarly, in studying bilingual writers, Adams and Guillot (2008) identified a significant link between PSTM and spelling performance, but complex working memory capacity showed no relationship with text quality. Evidently, more research is needed to confirm the nature of the link between different components of working memory and writing performance. Also, research is warranted to explore how different stages of the writing process may be linked to working memory, since, to the best of our knowledge, this association has not been researched to date.

3 Research questions

In light of the above, this proposed project intended to investigate the following research questions.

1. *What is the nature of the cognitive processes in which L2 writers of L1 Mandarin background engage when completing a version of Task 2 of the IELTS Academic Writing Test?*
2. *What is the nature of the online writing behaviours which L2 writers of L1 Mandarin background display when completing a version of Task 2 of the IELTS Academic Writing Test?*
3. *To what extent is text quality related to cognitive writing processes and online writing behaviours, for a version of Task 2 of the IELTS Academic Writing Test?*
4. *To what extent are phonological short-term memory, visual short-term memory, and executive control related to online writing behaviours and text quality, for a version of Task 2 of the IELTS Academic Writing Test?*

In the present study, L2 cognitive writing processes were operationalised in terms of participants' stimulated recall comments describing their internal cognitive processes. Online writing behaviours were operationalised as indices of fluency, pausing, and revision obtained via keystroke logging and recordings of participants' eye-movements when engaged in pausing behaviours. Text quality was defined as: (a) the linguistic complexity of the written texts as determined by automated text analysis software; and (b) task response, coherence and cohesion, lexical resource, and grammatical range and accuracy using IELTS rating criteria.

4 Methodology

4.1 Design

Thirty L2 English writers performed a version of Task 2 of the IELTS Academic Writing Test. Their online writing processes were recorded with a Tobii TX60 mobile eye-tracking system and the keystroke logging software Inputlog 6.1.5 (Leijten & Van Waes, 2013). After a short break, 12 participants were also requested to describe their thought processes during task performance via stimulated recall. All participants were administered a background questionnaire, and a battery of working memory tests.

4.2 Participants

The participants were 30 international students from a UK university, with IELTS entrance criteria of an overall score of 7.0. In terms of proficiency level, therefore, the targeted population was similar to students expected to take the IELTS test. Mandarin Chinese L2 users of English were recruited to control for the potential effect of first language on the findings. Most of the participating students were female (n=27). Their age ranged from 18 to 34 years with a mean of 26.60 (SD=3.69). The majority were studying towards a Master's level degree (n=24), five students were enrolled in a PhD program, and one participant was completing a Bachelor's degree.

4.3 Instruments and procedures

4.3.1 IELTS Test

A computer-based version of Task 2 of the IELTS Academic Writing Test was used to assess second language writing processes, behaviours, and outcomes. The IELTS essay prompt that students were asked to address was as follows:

> *Going overseas for university study is an exciting prospect for many people. But while it may offer some advantages, it is probably better to stay home because of the difficulties a student inevitably encounters living and studying in a different culture.*
>
> To what extent do you agree or disagree with this statement? Give reasons for your answer and include any relevant examples from your knowledge or experience.
>
> Write at least 250 words.

4.3.2 Stimulated recall procedure

Once the stimulated recall participants had completed the IELTS writing test, they were asked to describe the thought processes in which they engaged while performing the task as part of a stimulated recall protocol session. Participants were encouraged to pause the recording at any time they wished to describe the thoughts they had at any particular point during the test. In addition, the researcher paused the recording whenever participants paused, made a revision (e.g., substitution or deletion), or went back to parts of the text they had earlier produced. The stimulated recall sessions were carried out in English. Participants did not seem to experience difficulty in describing their thoughts in English, given their high level of proficiency.

4.3.3 Tests of working memory

Three components of Baddeley's (2000) working memory model were assessed: phonological short-term memory, visual short-term memory and executive control.

- Phonological short-term memory was assessed by a Mandarin Chinese non-word span (NW) and a Mandarin Chinese digit span test (DS).
- Visual short-term memory was gauged by the Forward Corsi Block (CBF) Task.
- Executing functioning was measured using the Backward Corsi Block (CBB), Operation Span (OSPAN), Colour Shape (CS), and Stop Signal Tasks (SST).

The order of the working memory tests was counterbalanced across participants.

4.3.3.1 Non-word span test

The Chinese non-word span test was adopted from Zhao (2013). The test was constructed using 48 one-syllable Chinese non-words. All of these pinyins could be pronounced but had no equivalent Chinese characters. They were randomised in order to form sequences containing 2 to 9 non-words, and presented to participants at a rate of one non-word per second. The participants' task was to recall each sequence immediately after they had heard the word 'okay'. The test entailed three trials for each sequence length. The test started with a short practice phase including two- and three-non-word sequences, followed by the actual test. Participants' non-word span was defined as the longest sequence for which they were able to recall at least one of the three sequences correctly.

4.3.3.2 Digit span test

The digit span test was also adopted from Zhao (2013), and had a similar design to the non-word span test. The test asked participants to recall sequences of 2 to 9 digits. The digit sequences were random generations of numbers from 11 to 99. The numbers were presented to the participants at a rate of one number per second. Each sequence ended with the word "okay", after which participants were asked to recall the sequence. The test entailed three trials for each sequence length, following a short practice phase of two- and three-digit sequences. Participants' digit span was determined as the highest number of digits they were able to recall at least once for a certain sequence length.

4.3.3.3 Corsi block tasks

The Forward Corsi Block task was included in the test battery to measure visual-spatial short-term memory capacity. The test was administered using Inquisit Lab 4. As part of this task, patterns of nine blocks were presented to the participants on the computer screen. For each trial, 2 to 9 blocks were highlighted, and the participants were asked to click the blocks in the same order as they had previously seen them highlighted. The number of the highlighted blocks gradually increased from 2 to 9. There were two trials for each sequence length.

The Backward Corsi Block task was included to assess the updating function of executive control. The test had the same format as the Forward Corsi Block task, the only difference was that the participants were asked to click the blocks in the reverse order of how they had previously been highlighted. The score for the two versions of the task was calculated based on the number of trials and the block span (i.e., the highest number of blocks that the participants could recall correctly at least on one of the two trials). It has been suggested that this total score is a more reliable index than the block span alone as it “takes into account the performance on both trials of an equal length” (Kessels et al., 2000, p. 254).

4.3.3.4 Automated operation span task (OSPAN)

The updating function of executive control was assessed by the OSPAN test (Turner & Engle, 1989) through the Inquisit Lab 4 platform. This task required participants to solve mathematic operations while keeping sets of English letters in memory. First, a math operation appeared on the screen that participants had to solve, then a letter was displayed. This was repeated until participants were asked to click the letters in the same sequence as they had previously been presented. Set sizes ranged from 3 to 7, where a set was defined in terms of the number of letters to be recalled. The test included three sets for each set size. Participants were presented with various set sizes in a random order. They were also asked to solve the math problems as quickly and accurately as possible. A 85% accuracy rate was set as a criterion following traditional scoring procedures (Unsworth et al., 2005). Participants' performance was expressed in terms of the absolute OSPAN score. This index is computed based on only those sets for which all letters are recalled accurately. Thus, following Unsworth et al. (2005), if a participant recalled all three letters in a set size of three, all four letters in a set size of four, but only three in a set size of five, their OSPAN score was 7 (3+4+0).

4.3.3.5 Colour shape task

To assess task-switching ability, the colour shape task was utilised (Miyake, Emerson, Padilla & Ahn, 2004). This task was also administered using Inquisit Lab. Participants were instructed to evaluate either the colour (e.g., green vs. red) or the shape (e.g., circle vs. triangle) of a stimulus consisting of coloured shapes, which were presented on the computer screen. In non-switching blocks, the participants only had to make a decision about the colour or the shape. In switching blocks, on the other hand, they were required to make a decision about either a colour or a shape according to a cue letter (C or S), which appeared on the screen. For instance, the participants had to indicate whether the colour was red or green in response to the cue letter “C”. In contrast, they were asked to identify the shape when the cue letter “S” appeared. In analysing the data, as a preliminary step, reaction times were trimmed to exclude values outside two standard deviations above or below the mean reaction time. Switching cost was expressed as the difference in mean reaction times between the two non-switching and two switching blocks (e.g., Altgassen et al., 2014; Friedman et al., 2006; Gold et al., 2013; Miyake et al., 2004).

4.3.3.6 Stop signal task

As a measure of inhibitory control, the stop signal task was included in the battery of working memory tests, and was presented via Inquisit Lab. An arrow stimulus was displayed on the computer screen, and the participants had to respond by pressing “D" on the keyboard if the arrow pointed to the left and “K” if the arrow pointed to the right. Participants, however, were instructed to withhold their response if the arrow was accompanied with an auditory signal (a beep). The mean reaction time (SSRT) was used to assess inhibitory control (Congdon et al., 2012; Enticott, Ogloff & Bradshaw, 2006). This index was computed after reaction times (SSRT) were trimmed to two standard deviations above or below the mean.

4.4 Data collection

All the participants attended one individual session. The session took approximately 2.5 hours for the non-stimulated recall participants and 4 hours for the stimulated recall participants. First, participants were asked to read the information sheet about the study, and to sign the consent form if they wished to participate in the research. Then, they completed a short background questionnaire. Next, the eye-tracker, a mobile Tobii X2-60 with a temporal resolution of 60 Hz, was calibrated. The eye-tracker was mounted to a 23" screen, with the participants sitting about 60cm away from the centre of the screen. A 9-point calibration grid was used to calibrate participants' eyes, and the experiment was presented with the help of Tobii Studio 3.0.9 software (Tobii Technology, n.d.). Once the eye-tracker had been calibrated, participants were asked to write the IELTS essay. After a short break, the stimulated recall participants were familiarised with the stimulated recall procedure, and then asked to describe their thought processes while writing the IELTS essay. The rest of the participants completed the working memory tests. The stimulated recall participants were administered the working memory tests after a short break following the stimulated recall session.

4.5 Data analysis

4.5.1 Analysis of stimulated recall comments

The analysis of the stimulated recall protocols involved five phases. First, the stimulated recall comments were transcribed. Second, one of the researchers reviewed the test-takers' comments describing the cognitive processes in which they engaged during writing and identified emergent categories. Third, the resulting categories were grouped into more general categories informed by Kellogg's (1996) model of L2 writing (see Tables 1 and 2 for examples of coding categories for pausing and revision respectively). Fourth, another researcher double-checked the micro-categories that emerged from the data and the more general categories that were formed. The percentage agreement between the first and second coder was 97% for coding micro-categories and 100% for identification of general categories. Finally, the comments falling into specific categories were added up to form a frequency count for each participant.

Table 1: *Examples for stimulated recall comments: Pausing*

Process/Subprocess	Example
PLANNING	Do I agree or disagree? Which position should I take? Which one is easy to write? Which side is easier to take?
Content	I was thinking what examples I was going to write here. What point should I make? I am thinking what kind difficulties they encounter...so I pause and think about difficulties.
Organisation	At that time, I was keeping on the eye on the word count. I found my word count is almost 250. I didn't have much space to develop my argumentation too much. I remembered that I wrote 'first of all', then... there should be 'secondly' or 'furthermore'. I realised that maybe I have space for only one opinion in detail. I was thinking how to structure the essay...I don't type all the main points for each paragraph. I would give different paragraphs for different topics.
FORMULATION **Lexical Retrieval**	Because I've already used the word 'discussions' so I was trying to think of another word which has the same meaning. I wanted to say 'if not facing the difficulties'. But I didn't think the expression is precise. I wanted to find another expression.
Syntactic Encoding	Uh...I was thinking whether I should treat 'study abroad' as a singular or plural form. Yes, because when I just first thought of using the word 'nationality', I thought in my own language there would not be any articles. Yeah so, but I think about the grammatical structure in English I may have to add the article.
Cohesion	I was thinking about linking words I should use. 'Secondly' is boring one. Should I use that? When I was writing this, all the paragraph was in my head. So I was thinking how to connect it better.
Unspecified	How to say. I mean very often I can figure out how to write smoothly in a simple way. I read lots of papers and I was greatly impacted by their way of expressing. I was trying to say a sentence a little bit in a complicated way…so it looks professional and academic. I had a meaning in my mind that this is very small population of this kind of students...it couldn't represent whole population so I was thinking about wording.
MONITORING	I want to maybe go back to the beginning and check one time and whether I should include anything. I finished the last paragraph and I went back to read whole essay. I review from the beginning...checking any grammar mistakes... I am proofreading.

Table 2: *Examples for stimulated recall comments: Revision*

Process/Subprocess	Example
PLANNING **Content**	I know I wanted to write a personal case of myself. So I wanted to start a sentence to bring my case to the essay. But later, you can see I regret afterwards. I deleted it. Ah yes....'cause at that time, I realised that critical thinking is one of the most important thing in research and in academic writing. Suddenly, I think about this point and I think it is important. I think it is good to point out critical thinking between UK education and Chinese education.
Organisation	I realised I type like I'm doing free writing. According to instruction it's like IELTS writing task so I suddenly remembered because I didn't take IELTS test before but I remember there must be some....may be some kind of structure I have to follow for that kind of formal writing, so I was thinking whether the way I am writing would not meet that kind of format required for the test. So I thought for a while and so I stopped and changed and deleted something. Yeah because I...when I was thinking about the second idea... I thought the structure would be improved if I tried maybe explained earlier by maybe dividing my ideas with maybe one key sentence or maybe different aspects, so that marker may be able to follow from that sentence, like, maybe for the first part it will be about maybe studying and for the second part is about living.
FORMULATION **Lexical Retrieval**	I didn't want to use 'competitiveness' or 'competence' because I used them before. I chose another word 'capacity'. Because I think it is little bit difficult for me to express the meaning of 'transfer'. In Chinese, it is transfer but, in this case, if I use 'transfer', I don't think it is appropriate. I used 'overcome' difficulties it will be easier for examiners to understand my meaning.
Syntactic Encoding	In the former sentence, I think I mentioned two things, first thing is I have never cooked before and the second thing is I have to think about how much money I spent. That means I talk about two things. Maybe I need change into plural. Because when I wrote this sentence, I didn't notice the tense and I examined it again and put the past tense.
Cohesion	Because I think for the first sentence I used all in singular form but if I use singular form and I have to use 'he' or 'she' for every sentence, so it may not be very convenient or it may not look that good, so maybe I am I was thinking whether I should change into plural so I can use 'they'. First, I used 'while' because I wanted to compare in the UK where I am forced to be independent and in China where I used to depend on parent and friends. First, I used 'while' but finally 'but' is a better connection word so I used 'but'.
Unspecified	I just I tried to rephrase the sentence to make it more academic. I revise the sentence into...I think...more proper way but I don't know it is enough. Actually I was not satisfied with the last sentence. I tried to revise it.

4.5.2 Analysis of online writing behaviours

To measure speed fluency, we utilised four measures: total writing time divided by total number of words/characters excluding pauses (minutes per word and characters per word), and number of words/characters occurring between pauses (words per P-burst and characters per P-burst). The threshold for pausing behaviour was set at two seconds, following conventions in writing research (Wengelin, 2006). Pausing behaviour was expressed in terms of number of pauses and mean length of pauses. Pauses were also categorised according to whether they occurred within words, or between words, sentences and paragraphs. Revision behaviours, such as deletions and substitutions, were measured by comparing the number of words/characters in the final text before and after the revision. Additionally, revisions were classified depending on whether they involved revisions below the word, word, below clause, clause or sentence level.

4.5.3 Analysis of eye-tracking data

In order to gain further insights into the nature of participants' online writing behaviours, we merged data from the recordings of the participants' eye movements and the pausing and revision patterns captured in the files produced by the keystroke logging software (Leijten & Van Waes, 2013). Next, we identified pauses in the Inputlog files using the two second pause threshold, and matched these pausing points to the corresponding positions in the eye-movement data with the help of Tobii Studio 3.0.9 software. Finally, we reviewed participants' gaze behaviours during pauses; in particular, we categorised participants' eye movements in terms of whether they remained within the word/expression, clause, sentence, or paragraph preceding the point of inscription. Occasionally, participants went back to the instructions or did not view the computer screen while they paused, these instances were coded as instruction and off-screen respectively.

4.5.4 Analysis of learner texts

The 30 texts produced by the participants were scored by an IELTS rater in terms of task response, coherence and cohesion, lexical resource, and grammatical range and accuracy, using IELTS rating criteria.

The texts of the test-takers were also analysed in terms of linguistic complexity (lexical, syntactic, and discourse complexity) and accuracy. Jarvis (2013) argued that lexical complexity or diversity entails at least six types of sub-constructs: volume (i.e., text length), evenness (i.e., distribution of token across types), dispersion (i.e., mean distance between tokens of the same type), rarity (i.e., frequency of words in the language), variability (i.e., type-token ratio corrected for text length), and disparity (i.e., proportion of semantically related words). In a project investigating the relationships among these facets of lexical diversity, Jarvis (2013) observed that volume, evenness, and dispersion correlate strongly. Thus, we decided to assess lexical diversity in terms of rarity, variability, and disparity, as participants in this study were asked to write texts of the same length (see also Mazgutova & Kormos, 2015).

Using the New General Service List (New-GSL, Brezina & Glabasova, 2013), rarity was expressed as proportion of the most frequent 500 (New-GSL 500), 501-1000 (New-GSL 1000), and 1001-2500 (New-GSL 2500) words in the texts. In addition to rarity of words, the proportion of words that were part of formulaic expressions in the texts was calculated. Specifically, we identified formulaic sequences in the 1,000, 2,000, 3,000, 4,000, and 5,000 word frequency bands (K1–K5) with the help of Martinez and Schmitt's (2012) Phrase list, which includes the 505 most frequent non-transparent formulae using the British National Corpus as a reference point.

Lexical variability was assessed using Malvern and Richards' (1997) D-formula and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010). The value D is estimated utilising a probabilistic mathematical model which creates a series of randomly sampled tokens to form a type-token ratio curve against increasing token size. MTLD refers to the mean length of word strings that conform to a certain threshold of type-token ratio. The indices of D and MTLD were obtained with the help of Coh-Metrix 3.0 (McNamara et al., 2005).

Following Jarvis (2013), disparity was operationalised as a latent semantic analysis (LSA) index, which we also obtained using Coh-Metrix 3.0. This LSA measure indicated the conceptual similarity between each sentence and every other sentence in the essays by analysing the semantic overlap among the lexical items the sentences.

The syntactic complexity of the texts was assessed in terms of three types of indices: complexity by subordination; phrasal complexity; and overall complexity (Norris & Ortega, 2009). Complexity by subordination was operationalised as the proportion of clauses per t-units. To measure phrasal complexity, the number of words was divided by the total number of clauses for each text. As an additional measure of phrasal complexity, the mean number of complex nominals per t-unit was calculated. Overall complexity was expressed in terms of the ratio of words to t-units and the Coh-Metrix 3.0 structural similarity index. Except for this measure, all indices were obtained by the program SynLex.

To assess the discourse complexity of the 30 essays, cohesion indices were also obtained with the help of the Coh-Metrix 3.0 program (McNamara et al., 2005). In particular, the texts were analysed for the use of various types of connectives. Connectives promote cohesion by providing cues about relationships between ideas presented in a text (Halliday & Hasan, 1976), and can be classified according to the type of cohesion they create, for example, whether they represent causal (e.g., because), logical (e.g., therefore), additive (e.g., and), or contrastive (e.g., however) relationships (Halliday & Hasan, 1976). We employed Coh-Metrix 3.0 to generate an incidence score for these type of connectives.

Accuracy was assessed in terms of the number of errors participants produced per 100 words. Errors were identified by one of the researchers, and 20% of the data were also double-coded by a native speaker with a background in language teaching. Intercoder agreements was found to be high (91%).

4.5.5 Statistical analyses

Research questions 1 and 2 were addressed by computing descriptive statistics based on the categories that emerged from the simulated recall comments and the data obtained about writing behaviours through the keystroke-logging and eye-tracking software.

Research questions 3 and 4 were answered by running Spearman correlational analyses. Correlations of .25, .40, and .60 large were considered small, medium, and large following Plonsky and Oswald (2014). Given that we ran a large number of correlations, we specified a conservative alpha level of .01.

5 Results

5.1 What is the nature of the cognitive processes in which L2 writers engage?

Table 3 summarises the stimulated recall comments, which were elicited to reveal the cognitive processes underlying participants' pausing behaviour while carrying the IELTS Academic Writing Task 2. As Table 3 demonstrates, the largest percentage of participants' stimulated recall comments referred to translation processes (48%), followed by comments describing planning operations (35%) and monitoring behaviours (11%). It is important to note, however, that this overall trend does not apply to all of the stimulated recall participants. In fact, four L2 writers paused more frequently in order to engage in planning rather than translation processes.

The distribution of planning and translation subprocesses yielded clearer trends. All but one participant mentioned planning content (29%) more frequently than planning organisation (6%) as a reason for pausing. As regards translation, all the students reported problems with lexical retrieval (33%) more often than with syntactic encoding (13%), and the majority referred to syntactic coding with greater frequency as compared to cohesion (3%).

Table 3: *Reasons for pausing: Summary of stimulated recall comments (N=12)*

	Planning				Translation					Monitoring		Don't remember		Total overall*	
	Con	Org	Tot		Lex	Syn	Coh	Tot							
Par	n	n	n	%	n	n	n	n	%	n	%	n	%	n	%
Par 2	4	0	4	25%	6	5	0	11	69%	0	0%	1	6%	16	100%
Par 6	0	4	4	40%	2	1	1	4	40%	0	0%	2	20%	10	100%
Par 10	2	0	2	25%	4	1	1	6	75%	0	0%	0	0%	8	100%
Par 15	9	0	9	64%	2	0	0	2	14%	2	14%	1	7%	14	100%
Par 17	3	2	5	31%	5	1	0	6	38%	3	19%	2	13%	16	100%
Par 21	6	0	6	29%	6	4	1	11	52%	4	19%	0	0%	21	100%
Par 23	2	0	2	9%	9	4	1	14	64%	5	23%	1	5%	22	100%
Par 24	5	0	5	50%	2	1	0	3	30%	0	0%	2	20%	10	100%
Par 25	7	1	8	24%	23	1	1	25	74%	1	3%	0	0%	34	100%
Par 26	6	2	8	44%	6	1	1	8	44%	2	11%	0	0%	18	100%
Par 28	9	2	11	55%	3	3	0	6	30%	2	10%	1	5%	20	100%
Par 29	10	1	11	41%	3	5	0	8	30%	4	15%	4	15%	27	100%
Tot	**63**	**12**	**75**	**35%**	**71**	**27**	**6**	**104**	**48%**	**23**	**11%**	**14**	**6%**	**216**	**100%**

Par = participant, Con = content, Org = organisation, Lex = lexical retrieval, Syn = syntactic encoding, Coh = cohesion, Tot = total; *Due to rounding some totals do not add up to 100.

Table 4 presents the summary of the stimulated recall comments which were elicited to describe participants' thoughts when they engaged in revision. Contrary to what was found for pausing, the distribution of revision-related comments was uniform across participants. All the participants referred to translation-related processes (70%) considerably more often than planning mechanisms (14%).

The trends observed for the sub-processes were similar to those we found for pausing. The stimulated recall comments yielded more reference to planning content (15%) than planning organisation (2%). Also, the participants attributed the largest percentage of their revision behaviours to lexical retrieval-related processes (37%), followed by revisions targeting morphosyntactic constructions (23%) and features associated with cohesion (10%).

Table 4: *Reasons for revision: Summary of stimulated recall comments (N=12)*

	Planning				Translation					Don't remember		Total overall*	
	Con	Org	Tot		Lex	Syn	Coh	Tot					
Par	n	n	n	%	n	n	n	n	%		%	n	%
Par 2	4	3	7	23%	8	8	3	21	68%	3	10%	**31**	**100%**
Par 6	0	1	1	25%	1	1	0	3	75%	0	0%	**4**	**100%**
Par 10	4	0	4	12%	9	3	8	24	73%	5	15%	**33**	**100%**
Par 15	4	0	4	33%	1	3	0	6	50%	2	17%	**12**	**100%**
Par 17	2	0	2	20%	1	4	0	7	70%	1	10%	**10**	**100%**
Par 21	1	0	1	5%	7	1	4	15	68%	6	27%	**22**	**100%**
Par 23	2	0	2	6%	10	9	3	24	75%	6	19%	**32**	**100%**
Par 24	4	1	5	12%	14	9	3	26	60%	12	28%	**43**	**100%**
Par 25	1	0	1	8%	16	7	2	12	92%	0	0%	**13**	**100%**
Par 26	11	0	11	22%	16	14	3	30	60%	9	18%	**50**	**100%**
Par 28	2	0	2	4%	25	7	4	37	82%	6	13%	**45**	**100%**
Par 29	3	0	3	15%	7	8	1	16	80%	1	5%	**20**	**100%**
Tot	**38**	**5**	**43**	**14%**	**115**	**74**	**31**	**221**	**70%**	**51**	**16%**	**315**	**100%**

Par = participant, Con = content, Org = organisation, Lex = lexical retrieval, Syn = syntactic encoding, Tot = total; *Due to rounding some totals do not add up to 100.

5.2 What is the nature of the online writing behaviours which L2 writers display?

Table 5 presents the descriptive statistics for the fluency, pausing, and revision behaviours of the participants. First, we consider the fluency indices. As shown in Table 5, participants, on average, produced 20 words and 100 characters per minute excluding pauses (M=.05 min per word, M=.01 min per character), and typed almost 4 words (M=3.75) and more than 20 characters (M=20.47) between pauses.

Turning to pausing behaviours, participants paused for the shortest period within words (M=5.19 s), followed by between words (M=5.34 s) and sentences (M=5.77 s). Pause length was the longest between paragraphs (M=6.33 s). The majority of pauses occurred between words (M=1.08). Considerably smaller number of pauses were observed within words (M=.10), and between sentences (M=.05) and paragraphs (M=.01).

Finally, the analysis of revision behaviours revealed that participants kept 79% of the words and 74% of the characters in the final draft, from among all the words/characters they had produced during the entire writing process. Participants made most revisions below the word level (M=90.73). They revised full words (M=40.07) and units smaller than clauses (M=43.97) on considerably fewer occasions. Full clauses (M=3.07) and units longer than clauses (M=2.60) were rarely revised.

Table 6 presents the descriptive statistics for the location of eye-movements during pauses. As Table 6 indicates, participants most frequently did not look at the screen when they paused (M=.11). As regards on-screen eye-fixations, participants stayed within the clause (M=.09) or paragraph (M=.09) in most cases. The next most frequent category was when eye-movements remained within the sentence (M=.08), followed by eye-gazes fixating at points within the word or expression (M=.07), the instruction (M=.06), or elsewhere on the screen (M=.05).

Table 5: *Descriptive statistics for fluency, pausing, and revision behaviours (N=30)*

	M	SD	95% CI
Fluency			
Minutes per word	.05	.01	[.04, .05]
Minutes per character	.01	.002	[.009, .01]
Words per P-burst	3.75	2.47	[2.96, 4.65]
Chars per P-burst	20.47	12.73	[16.40, 24.90]
Pause length (s)			
Total	5.59	1.13	[5.18, 5.99]
Within words	5.19	1.80	[4.55, 5.86]
Between words	5.34	1.80	[4.73, 5.99]
Between sentences	5.77	2.72	[4.89, 6.75]
Between paragraphs	6.33	3.84	[5.05, 7.90]
Pause frequency per 100 words			
Total	.43	.21	[.36, .51]
Within words	.10	.07	[.08, .13]
Between words	1.08	.20	[1.02, 1.16]
Between sentences	.05	.01	[.04, .06]
Between paragraphs	.01	.005	[.007, .01]
Revision overall			
Words product/process	.79	.10	[.75, .83]
Chars product/process	.74	.11	[.69, .78]
Revision by location per 100 words			
Below word	90.73	70.45	[66.77, 117.82]
Full word	40.07	37.87	[28.60, 54.86]
Below clause	43.97	45.10	[30.10, 60.29]
Full clause	3.07	3.34	[1.97, 4.33]
Sentence	2.60	4.29	[1.33, 4.23]

Table 6: *Descriptive statistics for location of eye-gazes per 100 words (N=30)*

	M	SD	95% CI
Word or expression	.07	.06	[.05, .10]
Clause	.09	.07	[.07, .12]
Sentence	.08	.07	[.06, .10]
Paragraph	.09	.05	[.07, .11]
Instruction	.06	.06	[.04, .08]
Elsewhere	.05	.04	[.04, .06]
Off-screen	.11	.10	[.08, .15]

5.3 To what extent is text quality related to online writing behaviours?

5.3.1 Relationships between IELTS scores and online writing behaviours

The descriptive statistics for the IELTS scores are presented in Table 7. As Table 7 shows, participants' mean total IELTS writing score was close to 7 (M=6.88), with participants achieving the highest sub-score in the category task response (M=7.37), followed by lexical resource (M=6.83), grammatical range and accuracy (M=6.73), and coherence and cohesion (M=6.57).

Table 7: *Descriptive statistics for IELTS*

	M	SD	95% CI
Task response	7.37	1.38	[6.90, 7.83]
Coherence and cohesion	6.57	.90	[6.27, 6.90]
Lexical resource	6.83	1.15	[6.43, 7.23]
Grammatical range and accuracy	6.73	1.01	[6.40, 7.10]
Total	**6.88**	**.95**	**[6.59, 7.22]**

Table 8 presents the results of the Spearman correlations which were run to assess the relationships between the IELTS scores and measures of writing behaviours. As shown in Table 8, three correlations involving fluency, four involving frequency of pause, and five involving the location of eye-fixations were found to be significant. All of the fluency measures included the measure minutes per word. Participants achieved higher ratings on task response, lexical resources, and the total when they produced more words per minute (excluding pauses). The effect sizes were medium (lexical resource: rho=-.53; IELTS total: rho=-.53) or large (task response: rho=-.61). Turning to pausing, participants who paused more often within words received lower ratings in the categories of task response (rho=-.51), lexical resource (rho=-.53), and the total (rho=-.50). Those with lower task response ratings also paused more frequently between paragraphs (rho=-.53). All of these significant relationships for pausing were of medium size.

Finally, the eye-tracking data revealed that those participants who returned to points within the paragraph they were writing with greater frequency while they paused, wrote less successful IELTS essays in terms of task response (rho=-.50).

In addition, participants looking away from the screen more frequently during pauses produced essays that were rated as less successful in terms of task completion (rho=-.49), lexical complexity (rho=-.55), accuracy (rho=-.51) and overall quality (rho=-.60). The strength of these relationships was of medium size, except for a large effect size for the link between number of off-screen eye-gazes and the IELTS total score.

Table 8: *Spearman correlations between IELTS scores and writing behaviours (N=30)*

	Task response	Coh. & cohesion	Lexical resource	Grammar & accuracy	IELTS total
Fluency					
Minutes per word	**-.61****	-.26	**-.53****	-.38*	**-.53****
Minutes per character	-.22	.14	-.20	-.25	-.18
Words per P-burst	.37*	.09	.40*	.41*	.41*
Chars per P-burst	.33	.04	.36	.42*	.38*
Pause length (s)					
Total	-.07	-.05	-.09	-.18	-.14
Within words	-.17	-.02	-.20	-.26	-.20
Between words	-.13	-.11	-.26	-.27	-.23
Between sentences	-.18	-.14	-.23	-.24	-.24
Between paragraphs	-.11	-.21	-.24	-.32	-.26
Pause frequency per 100 words					
Total	-.37*	-.04	-.39*	-.34	-.36*
Within words	**-.51****	-.11	**-.53****	-.45*	**-.50****
Between words	.08	.30	.17	.25	.21
Between sentences	-.13	-.04	-.14	-.08	-.06
Between paragraphs	**-.53****	-.10	-.30	-.37*	-.42*
Revision overall					
Words product/process	.07	-.16	-.02	-.10	-.06
Chars product/process	.08	-.13	.05	-.05	-.03
Revision by location per 100 words					
Below word	-.14	.01	-.29	-.21	-.14
Full word	-.09	.13	-.14	-.07	.00
Below clause	-.05	.23	-.04	.05	.08
Full clause	-.21	.25	-.07	.02	.04
Sentence	-.24	.14	-.02	.06	.00
Location of eye-gazes					
Word or expression	-.25	-.03	-.13	-.19	-.18
Clause	-.05	.32	.07	-.01	.07
Sentence	-.33	.07	-.03	-.09	-.15
Paragraph	**-.50****	-.10	-.33	-.32	-.38*
Instruction	-.08	.33	.08	.05	.00
Elsewhere	-.28	.52	-.29	-.30	-.29
Off-screen	**-.49****	.00	**-.55****	-.51**	**-.60****

**. Correlation is significant at the 0.01 level (2-tailed).

To sum up, writers with lower fluency, more frequent pausing within words, and more frequent off-screen eye-gazes and within the current paragraph during pauses, wrote less effective IELTS essays in terms of task response. Decreased fluency, more frequent pausing within words, and off-screen eye movements were also associated with lower lexical complexity and overall lower IELTS scores. Finally, less fluent writing behaviour and more off-screen gazing also predicted IELTS scores that were lower overall.

5.3.2 Relationships between linguistic complexity and online writing behaviours

Table 9 gives the descriptive statistics for the linguistic complexity measures. As shown, the majority of words (M=74.00) in the final texts were among the 500 most frequent words according to the new-GSL list, but the essays also contained some less frequent, off-list words (M=8.58). Participants, too, included formulaic expressions(M=9.47) in the texts, most of which came from the K1-K3 range according to the Phrase list. The language of the essays was, on average, lexically varied, with considerable semantic overlap among the words. Participants also produced syntactically complex language. The average length of t-units was nearly 18 words (M=17.81), including more than one clause (M=1.67). The texts utilised a large number of connectives (M=101.87); additive (M=55.07) and logical connectives (M=49.95) were particularly frequent in the IELTS essays.

Table 9: *Descriptive statistics for linguistic complexity (N=30)*

	M	SD	95% CI
Lexical diversity			
New-GSL 500	74.00	4.29	[72.52, 75.59]
New-GSL 1000	9.93	2.37	[9.10, 10.78]
New-GSL 2500	7.48	1.75	[6.92, 8.11]
Off-list words	8.58	3.00	[7.62, 9.74]
Phrase List Total	9.47	4.97	[7.60, 11.33]
Phrase List 1000	3.97	2.55	[3.10, 4.87]
Phrase List 2000	2.27	1.57	[1.73, 2.87]
Phrase List 3000	2.10	1.73h	[1.47, 2.77]
Phrase List 4000	.97	1.16	[.60, 1.37]
Phrase List 5000	.17	.59	[.00, .43]
MTLD	89.70	16.87	[83.79, 95.97]
D-value	91.55	14.35	[86.84, 96.52]
LSA	.17	.04	[.15, 18]
Syntactic complexity			
Structural similarity	.10	.02	[.09, 1.00]
Words/t-unit	17.81	3.57	[16.52, 19.11]
Words/clause	10.80	1.58	[10.27, 11.37]
ComNom/t-unit	2.09	.58	[1.88, 2.28]
Clause/t-unit	1.67	.35	[1.55, 1.80]
Discourse complexity			
All connectives	101.87	19.07	[94.98, 108.32]
Causal connectives	31.92	13.13	[27.25, 36.49]
Logical connectives	49.95	14.42	[45.02, 55.00]
Contrastive connectives	15.87	8.53	[12.94, 19.11]
Additive connectives	55.07	14.86	[48.94, 60.13]

Table 10 summarises the results of the Spearman correlations carried out between the lexical diversity measures and indices of writing behaviours. Eight significant relationships were identified. Participants who produced fewer words and characters per P-bursts included a significantly larger number of words from the New-GSL 1000 word list (rho=-.56 and rho=-.53, respectively). Also, participants who paused longer in total, as well as within and between words, used New-GSL 1000 words with greater frequency (rho=.49, rho=.48, rho=.48 respectively). Greater overall frequency of pausing was significantly and positively correlated with the number of words from the New-GSL 1000 word list (rho=.56). Amount of sentence-level revision also had a positive relationship with the number of off-list words occurring in the IELTS essays (rho=.55). Finally, more frequent eye-movements targeting the word or expression just produced during pauses were associated with more extensive use of New-GSL 1000 words in the texts (rho=.47). In sum, participants who exhibited lower writing fluency; used longer, more frequent pauses; and looked more often on the word/expression that they had just typed during pauses, included a larger number of frequent words in their IELTS essays. On the other hand, infrequent words were more often utilised by writers if they made more higher-level revisions.

The Spearman correlations computed between the syntactic complexity measures and indices of writing behaviours are summarised in Table 11. The analyses yielded five significant correlations. A negative relationship emerged between the number of words produced per minute and the number of clauses per t-unit (rho=-.48), that is, more fluent writers wrote IELTS essays with greater clausal complexity. More frequent pausing between words was also correlated with the structural similarity index (rho=.53) and words per t-unit (rho=-.52), indicating that those who paused more between sentences used less diverse syntactic structures and produced, in general, less syntactically complex language. Significant links were also identified between phrasal complexity and frequency of eye fixations; the more participants fixated on the instructions during pauses, the more likely they were to produce essays with greater phrasal complexity (rho=.47). Finally, off-screen viewing behaviour while pausing was negatively correlated with subordination complexity, that is, participants who looked away from the screen while pausing wrote essays with fewer subordinate clauses (rho=-.49). All effects sizes were in the medium range. To sum up, greater syntactic complexity was found to be associated with greater speed fluency, less frequent pausing, more gazes on the instruction during pauses, but fewer off-screen eye fixations.

Table 12 gives the results of the Spearman correlations we ran between the discourse complexity measures and indices of writing behaviours. Six significant relationships were found. Participants who used clausal connectives more often also wrote fewer words (rho=-.47) and characters between pauses (rho=-.47). Also, more extensive use of clausal connectives was associated with more frequent pausing overall and within words (rho=.51; rho=.48). Interestingly, however, participants who paused less frequently between words produced a greater number of contrastive connectives (rho=-.51). Finally, overall use of connectives was positively related to how often participants looked at the instruction during pauses (rho=.58). These significant correlations were all of medium effect size. In summary, greater discourse complexity, expressed in terms of use of causal connectives, was linked to decreased speed fluency and more frequent pausing, more complex discourse operationalised as use of contrastive connectives was associated with fewer number of pauses, more gazes on the instruction predicted more extensive use of connectives.

Table 10: *Spearman correlations between lexical diversity and writing behaviours (N=30)*

	New GSL 500	New GSL 1000	New GSL 2500	Off-list words	Phrase total	Phrase 1000	Phrase 2000	Phrase 3000	Phrase 4000	Phrase 5000	MTLD	D-value	LSA
Fluency													
Minutes per word	-.11	.16	-.25	.07	-.38*	-.37*	-.04	-.21	-.21	-.15	-.11	-.13	.05
Minutes per char	-.29	.24	-.01	.29	-.36*	-.24	-.27	-.24	-.32	.06	-.16	-.12	.25
Words per P-burst	.22	**-.56****	.23	.06	.40*	.33	.04	.00	.01	-.20	.00	.09	-.01
Chars per P-burst	.14	**-.53****	.27	.13	.32	.29	.03	-.03	.00	-.23	.01	.11	.01
Pause length													
Total	-.23	**.49****	-.07	-.01	-.19	-.23	-.13	.09	-.24	-.05	-.10	-.19	.14
Within words	-.17	**.48****	-.13	-.06	-.03	-.22	-.04	.21	-.01	.13	.01	.01	-.04
Between words	-.15	**.48****	-.22	-.03	-.04	-.29	-.20	-.01	-.21	-.02	-.08	-.17	-.03
Between sentences	-.28	.27	.13	.12	-.27	-.26	.12	.31	-.13	-.11	.20	.14	.27
Between paragraphs	-.02	.39*	-.18	-.17	-.06	-.07	-.06	.08	-.13	-.01	-.12	-.14	-.06
Pause frequency													
Total	-.15	**.56****	-.27	-.08	-.38*	-.43*	-.13	-.14	-.28	-.04	-.22	-.21	.14
Within words	-.10	.43*	-.28	-.04	-.35	-.35	-.09	-.23	-.24	-.09	-.23	-.26	.09
Between words	-.42*	.22	.30	.26	-.09	.01	.09	-.16	-.27	-.06	.02	.07	.29
Between sentences	-.21	-.12	.16	.32	-.31	-.10	-.21	-.23	-.38*	-.18	.00	.18	.09
Between paragraphs	-.47*	.26	.03	.44*	-.38*	-.29	-.10	-.28	-.24	-.19	.05	.13	.08
Revision overall													
Words prod/proc	.27	-.10	-.16	-.22	.19	.14	-.14	.22	.25	.25	.07	-.04	-.36*
Chars prod/proc	.11	.01	-.03	-.14	.30	.23	-.07	.30	.27	.32	.11	-.04	-.32
Revision by location per 100 words													
Below word	.22	-.21	-.16	-.06	-.20	-.19	.07	-.12	-.24	-.21	-.09	-.03	-.01
Full word	.36	-.29	-.33	-.09	-.21	-.17	-.01	-.14	-.24	-.15	-.12	.04	-.04
Below clause	-.03	-.20	.02	.18	-.05	-.03	.14	-.02	-.19	-.19	-.02	-.11	.04
Full clause	-.06	-.25	-.05	.31	-.26	-.29	-.13	-.11	-.03	-.20	.12	.14	.08
Sentence	-.24	-.28	.03	**.55****	-.32	-.18	-.19	-.26	-.24	-.16	.08	.23	-.07
Location of eye-gazes													
Word or expression	-.26	**.47****	-.02	.07	-.11	-.18	.12	-.01	.00	-.18	-.16	-.23	.11
Clause	-.04	.31	.07	-.23	-.15	-.19	-.05	-.08	-.04	-.01	-.14	-.10	.23
Sentence	.11	.18	-.17	-.33	-.26	-.29	-.18	-.18	-.10	-.24	-.16	-.18	.27
Paragraph	-.03	.03	.02	-.11	-.28	-.43*	-.02	-.11	-.09	-.17	.09	.11	.04
Instruction	-.19	.35	-.05	.01	-.19	-.06	-.08	-.26	-.31	.17	-.09	-.18	.11
Elsewhere	-.17	.35	-.11	.05	-.28	-.37*	.00	-.06	.02	-.01	-.11	-.15	.02
Off-screen	-.26	.20	-.16	.16	-.15	-.19	.02	.12	-.16	.09	.05	.01	-.15

**. Correlation is significant at the 0.01 level (2-tailed).

Table 11: *Spearman correlations between syntactic complexity and writing behaviours (N=30)*

	Structural similarity	Words/ t-unit	Words/ clause	ComNom/ t-unit	Clause/ t-unit
Fluency					
Minutes per word	.22	-.33	.21	-.28	**-.48****
Minutes per character	.22	.01	.22	.10	-.16
Words per P-burst	-.11	.05	-.26	-.11	.25
Chars per P-burst	-.09	.00	-.27	-.14	.21
Pause length (s)					
Total	.05	-.02	.16	.15	-.13
Within words	.16	-.18	.10	-.12	-.25
Between words	.09	-.04	.11	.18	-.13
Between sentences	.11	-.20	.23	.05	-.35
Between paragraphs	.19	-.39*	.12	-.01	-.41*
Pause frequency					
Total	-.04	-.12	.18	.05	-.26
Within words	.06	-.23	.12	-.15	-.33
Between words	.00	.15	.16	.08	.03
Between sentences	**.53****	**-.52****	-.26	-.34	-.30
Between paragraphs	.11	-.30	.11	-.33	-.37
Revision overall					
Words product/process	-.11	.08	-.02	.13	.08
Chars product/process	-.12	.21	.13	.24	.11
Revision by location per 100 words					
Below word	.06	-.16	-.23	-.15	.04
Full word	-.08	-.12	-.30	-.11	.13
Below clause	.07	.05	.17	-.05	-.02
Full clause	.32	-.26	-.13	-.34	-.15
Sentence	.09	-.15	.07	-.16	-.15
Location of eye-gazes					
Word or expression	.02	.01	.06	.09	-.09
Clause	-.10	.15	-.11	.12	.12
Sentence	-.14	.09	.00	.12	-.01
Paragraph	-.03	-.18	.16	-.11	-.29
Instruction	-.24	.16	**.47****	.13	-.18
Elsewhere	.03	-.25	.04	-.08	-.25
Off-screen	.25	-.34	.23	.01	**-.49****

Table 12: *Spearman correlations between discourse complexity and writing behaviours (N=30)*

	All connectives	Causal connectives	Logical connectives	Contrastive connectives	Additive connectives
Fluency					
Minutes per word	.26	.26	.22	-.01	.33
Minutes per character	.04	.33	.05	-.28	-.02
Words per P-burst	-.17	**-.47****	-.34	.15	-.08
Chars per P-burst	-.19	**-.47****	-.35	.17	-.08
Pause length					
Total	.25	.31	.33	.09	.19
Within words	.31	.18	.21	-.01	.39*
Between words	.14	.28	.41*	.10	.14
Between sentences	.04	.38*	.45*	.09	.04
Between paragraphs	-.09	.26	.32	.23	-.08
Pause frequency					
Total	.27	**.51****	.45*	.06	.18
Within words	.18	**.48****	.41*	.02	.09
Between words	-.10	-.11	-.34	**-.51****	-.02
Between sentences	-.28	-.26	-.40*	-.34	-.04
Between paragraphs	.42*	.33	.44*	.14	.40*
Revision overall					
Words product/process	.08	.09	.33	.31	-.03
Chars product/process	.11	.10	.29	.15	-.01
Revision by location per 100 words					
Below word	-.18	-.09	-.18	-.03	-.15
Full word	-.30	-.14	-.23	.11	-.30
Below clause	-.18	-.26	-.38*	-.46*	-.04
Full clause	-.39*	-.27	-.34	-.37*	-.14
Sentence	-.21	-.19	-.32	-.34	-.08
Location of eye-gazes					
Word or expression	-.11	.18	.21	.03	-.12
Clause	-.08	.22	.20	-.06	-.24
Sentence	-.02	.25	.10	-.09	-.06
Paragraph	.15	.14	.05	.11	.21
Instruction	**.58****	.20	.32	-.19	.44*
Elsewhere	.08	.20	.13	.07	.13
Off-screen	.10	.22	.24	.01	.26

**. Correlation is significant at the 0.01 level (2-tailed).

5.3.3 Relationships between accuracy and online writing behaviours

Table 13 gives the descriptive statistics for our accuracy measure, errors per 100 words. Participants, on average, produced highly accurate texts, they only committed three errors per 100 words (M=.03).

Table 13: *Descriptive statistics for accuracy (N=30)*

	M	SD	95% CI
Errors per 100 words	.03	.02	[.02, .03]

As Table 14 shows, none of the Spearman correlations conducted between this measure of accuracy and writing behaviours were found to be significant, indicating that whether participants wrote more fluently, produced shorter and fewer pauses, or revised less did not predict the accuracy level of their texts.

Table 14: *Spearman correlations between accuracy and writing behaviours (N=30)*

	Accuracy
Fluency	
Minutes per word	.31
Minutes per character	.01
Words per P-burst	-.05
Chars per P-burst	-.03
Pause length	
Total	-.17
Within words	.01
Between words	-.04
Between sentences	-.02
Between paragraphs	.05
Pause frequency	
Total	.03
Within words	.21
Between words	-.19
Between sentences	.10
Between paragraphs	.26
Revision overall	
Words product/process	-.03
Chars product/process	-.08
Revision by location per 100 words	
Below word	.15
Full word	.06
Below clause	.03
Full clause	.23
Sentence	.14
Location of eye-gazes	
Word or expression	.03
Clause	-.13
Sentence	.05
Paragraph	.19
Instruction	-.24
Elsewhere	.05
Off-screen	.22

5.4 What is the nature of the relationship of phonological short-term memory, visual short-term memory, and executive control to online writing behaviours and text quality?

Table 15 gives the descriptive statistics for the working memory measures, while Tables 16 and 17 present the results of the Spearman correlations between the various working memory tests and the indices of writing behaviours and text quality. Three significant, medium-size correlations were found between working memory skills and writing behaviours: participants with more superior task-switching ability paused for shorter periods between sentences (rho=.59); those who had better ability to update information paused less frequently between paragraphs (rho=-.51); and those who had less superior visual short-term memory gazed on the instructions more frequently during pauses (rho=-.52).

Table 17 demonstrates that three significant links emerged between the working memory and text quality indices. Two significant, medium-size correlations included the measure of task-switching ability. Participants who were less able to switch between tasks produced a greater number of New-GSL 1000 words (rho=.46), and used more logical connectives (rho=.48). An additional, strong relationship was found between participants' non-word span scores and their use of words from the New-GSL 1000 list. Those who had better span scores included a larger number of New-GSL 1000 words (rho=.60).

Table 15: *Descriptive statistics for working memory measures (N=30)*

	M	SD	95% CI
Phonological short-term memory			
Non-word span	3.34	1.26	[2.90, 3.79]
Digit span	3.83	.71	[3.59, 4.10]
Visual-spatial short-term memory			
Corsi block forward	58.80	22.69	[50.57, 66.76]
Executive control			
Corsi block backward	57.53	12.42	[53.30, 61.77]
Operation span task (updating)	51.33	18.37	[44.97, 58.37]
Colour shape task (task-switching ability) (ms)	481.89	361.81	[364.37, 618.57]
Stop signal task (inhibitory control) (ms)	299.85	58.17	[277.10, 320.34]

Table 16: *Spearman correlations between working memory measures and writing behaviours (N=30)*

	NWS	DS	CBF	CBB	OSPAN	CST	SST
Fluency							
Minutes per word	.06	.20	-.37*	-.04	-.37*	.00	.11
Minutes per character	-.13	.23	-.31	-.31	-.29	.27	.05
Words per P-burst	.00	-.19	.37*	.01	.19	-.41*	-.16
Chars per P-burst	.04	-.14	.38*	.04	.21	-.41*	-.13
Pause length							
Total	.27	.29	-.06	.12	-.10	.42*	-.08
Within words	.23	.18	.01	.15	-.10	.25	.09
Between words	.23	.17	-.06	.04	-.11	.44*	-.06
Between sentences	.31	.28	.00	.23	-.17	**.59****	-.11
Between paragraphs	.34	.14	.10	.14	-.31	.40*	-.35
Pause frequency							
Total	.09	.20	-.34	.07	-.24	.32	.16
Within words	.08	.23	-.44*	-.10	-.22	.17	.09
Between words	.39*	-.28	-.05	-.21	-.24	.14	.00
Between sentences	.15	.10	.11	-.01	-.33	-.06	-.32
Between paragraphs	.35	-.02	-.14	.03	**-.51****	.38*	.09
Revision overall							
Words product/process	-.31	.08	.05	.10	.39*	.08	.12
Chars product/process	-.22	.04	.06	.03	.41*	.12	.08
Below word	-.19	.34	-.25	.00	.01	-.24	.13
Full word	-.27	.20	-.07	-.02	-.01	-.38*	.12
Below clause	-.17	.25	-.23	-.18	.12	-.29	.08
Full clause	-.19	.20	.01	-.05	.05	-.31	.05
Sentence	.02	.16	-.18	-.03	.04	-.09	.09
Location of eye-gazes							
Word or expression	.28	.01	.05	-.16	-.31	.35	-.13
Clause	-.08	-.17	-.02	.14	-.22	.23	.14
Sentence	-.23	-.28	-.16	.12	-.25	.01	.41*
Paragraph	-.24	.12	-.39*	.07	-.36	.00	.08
Instruction	.04	-.02	**-.52****	-.08	-.16	.43*	.40*
Elsewhere	.01	.33	-.29	-.13	-.08	.25	.03
Off-screen	.13	.23	-.09	.01	-.19	.06	-.45*

**. Correlation is significant at the 0.01 level (2-tailed).

NWS = non-word span, DS = digit span, CBF = Corsi block forward, CBB = Corsi block backward, OSPAN = operation span, CST = colour shape task, SST = stop signal task

6 Summary and discussion

6.1 What is the nature of the cognitive processes in which L2 writers engage?

This study utilised the stimulated recall procedure to tap into the cognitive processes in which L2 writers engage when performing one version of the IELTS Academic Writing Task 2. In particular, L1 Mandarin participants were prompted to describe what they were thinking when they paused and revised their texts. In line with Kellogg's (1996) model of writing, the stimulated recall comments revealed that, as predicted by the model, participants engaged in planning, translation and monitoring processes.

As summarised in Table 3, nearly half of the pauses were associated with translation processes, with participants referring most frequently to problems related to lexical retrieval followed by syntactic encoding and cohesion. Slightly more than a third of the comments mentioned planning operations, the large majority of which were concerned with planning content. Only a small percentage of comments made reference to organisation. According to the stimulated recall comments, approximately 10% of the pauses were underlain by monitoring processes. As compared to pausing, a considerably larger number of revision-related stimulated recall comments mentioned translation processes, with 70% of the comments being associated with linguistic encoding. Similar to pausing, however, participants made lexical revisions most frequently, followed by revisions to morphosyntactic and cohesive features. Only 14% of the revision-related comments referred to planning, most of which concerned planning the content of the essay.

These findings, overall, suggest that the IELTS Academic Writing Task 2 has cognitive validity in the sense that the cognitive processes in which L2 writers engaged while completing the task reflected the processes which L1 writers typically employ, as captured in Kellogg's (1996) well established model of writing.

Our results also suggest that the cognitive processes of L2 writers completing the IELTS Academic Writing Task 2 are well aligned with the intended focus of the assessment. The *IELTS Candidate Guide* states that the aim of the academic writing test is to assess test-takers' ability to write an appropriate response in terms of content, organisation, and accuracy and range of lexis and grammar. The participants in the present study did engage in cognitive writing processes reflecting these focus areas.

Table 17: *Spearman correlations between working memory and text quality measures (N=30)*

	NWS	DS	CBF	CBB	OSPAN	CST	SST
IELTS scores							
Task response	.07	-.03	.15	.02	.35	.00	-.01
Coh and cohesion	-.20	-.01	-.16	-.27	-.01	.09	.12
Lexical resource	-.04	-.37*	.16	-.08	.14	.04	.20
Gram range and acc	.12	-.33	.19	.08	.30	-.17	.25
Total	.02	-.19	.18	.01	.24	-.06	.17
Lexical diversity							
New-GSL 500	-.31	-.15	.07	.10	-.05	-.20	.03
New-GSL 1000	**.60****	.11	-.06	.15	-.07	**.46****	.04
New-GSL 2500	-.06	.03	.11	.11	.33	-.17	.06
Off-list words	.17	.30	-.03	-.25	.07	.04	-.18
Phrase List Total	-.05	.04	-.05	.00	.20	-.10	.07
Phrase List 1000	.03	-.08	-.10	-.07	.16	.10	.13
Phrase List 2000	.14	.22	-.01	.05	-.01	-.05	-.11
Phrase List 3000	-.02	.28	.12	.15	.27	-.09	-.02
Phrase List 4000	-.12	.02	.09	.07	.35	**-.46***	-.12
Phrase List 5000	-.16	-.08	-.25	-.32	-.06	.28	.13
MTLD	-.28	.22	.07	-.02	.20	-.17	-.01
D-value	-.36	-.03	.03	-.07	-.05	-.18	.10
LSA	.33	-.21	-.10	.24	-.15	.19	.24
Syntactic complexity							
Structural similarity	.07	.17	.13	.06	-.18	-.05	-.09
Words/t-unit	.08	-.17	-.11	-.25	.27	.10	.16
Words/clause	.23	.15	-.24	-.16	.30	.16	.14
ComNom/t-unit	.03	-.14	.00	-.14	.14	.12	.17
Clause/t-unit	-.13	-.30	.06	-.19	.14	-.10	.01
Discourse complexity							
All connectives	.14	.21	-.36	-.09	-.01	.35	-.05
Causal connectives	-.06	-.02	.01	-.20	-.15	.30	-.02
Logical connectives	.16	.12	.07	.06	.03	**.48****	.08
Cont. connectives	.05	.14	.29	.29	-.11	-.01	-.23
Additive connectives	.21	.34	-.30	.15	.09	.11	-.05
Accuracy							
Errors per 100 words	.02	.12	-.07	.01	-.19	.09	-.22

**. Correlation is significant at the 0.01 level (2-tailed).

NWS = non-word span, DS = digit span, CBF = Corsi block forward, CBB = Corsi block backward, OSPAN = operation span, CST = colour shape task, SST = stop signal task

6.2 What is the nature of the online writing behaviours which L2 writers display?

Keystroke logging and eye-tracking methodology were employed to examine the online writing behaviours of L2 writers when carrying out the IELTS Academic Writing Task 2. More specifically, we assessed the speed fluency, the length and frequency of pausing across locations, the total amount of revision overall and by location, and the location of eye-movements during pauses. Participants, on average, wrote 20 words and 100 characters per minute excluding pauses, and typed almost 4 words and more than 20 characters between pauses. Pauses were shortest within words; followed by pauses between words, sentences, and paragraphs. Most of the pauses occurred between words. Of the total words and characters they produced during the writing process, participants kept 79% of their words and 74% of their characters in the final draft. The majority of revisions occurred at the word level.

It is worth comparing the results for pausing to those of Spelman Miller (2000), as our study looked at a similar population of L2 writers. The two studies yielded similar trends. Spelman Miller also observed that pause length gradually increased as text level unit increased; and pauses were most frequent between intermediate constituents, a category parallel to pauses between words. Pause bursts were also found to be in a similar range, reaching almost 4 words per minute in both studies. Notably, this is lower than the rate identified by Spelman Miller for native writers.

Similar to the stimulated recall comments, the keystroke logging indices, as well as the eye-gaze data, provide further confirmation of the cognitive validity of the IELTS Academic Writing Task 2. As predicted by Kellogg's model of writing, the task prompted test-takers to engage in differential cognitive processes, including both lower- and higher-level writing operations. This was reflected in the fact that participants paused and revised at various text level units and gazed at various levels of previously produced texts. Pausing and revision at lower and higher level of text units have been shown to be associated, respectively, with lower and higher-level writing processes (cf. Révész, Kourtali, & Mazgutova, 2017; Stevenson et al., 2006).

6.3 To what extent is text quality related to online writing behaviours?

A series of Spearman correlations were conducted between the text quality measures and indices of writing behaviours to establish relationships between the process and product measures. A number of significant links were observed, which are summarised in Table 18, grouped according to measures of writing behaviours.

Less fluent writing, expressed in terms of minutes per word, was associated with lower IELTS task response, lexical resource, and total scores, as well as with lower syntactic complexity (subordination). Lower fluency, defined as words and characters per P-burst, was also related to decreased lexical complexity (more frequent use of New-GSL 1000 words) and more extensive use of causal connectives.

Longer pauses in total, within words and between words, predicted less sophisticated vocabulary use, i.e., more extensive use of New-GSL 1000 words. More frequent pausing in total was also related to a larger percentage of New-GSL 1000 words in the texts, i.e., less sophisticated use of lexis. Those who paused more often overall, too, produced more causal connectives. More extensive pausing within words was found to be associated with lower IELTS task response, lexical resource and total scores.

Increased pausing within words also predicted more frequent use of causal connectives. Interestingly, however, larger number of pauses between words was linked to decreased production of contrastive connectives. More frequent pauses between sentences were associated with lower syntactic complexity, more specifically, greater structural similarity and shorter t-units. Finally, those who paused more between paragraphs wrote less effective essays in terms of IELTS task response criteria. Participants who engaged in more sentence-level or higher level revisions also produced more sophisticated lexis.

Where participants looked while pausing was also found to predict some aspects of text quality. Those participants who gazed at the previously produced word or expression more often while pausing wrote essays with less sophisticated lexis. Greater number of eye-movements staying within the same paragraph predicted lower IELTS task response scores. Looking back on the instruction during pauses, however, was associated with greater phrasal complexity (words per clause) and more extensive use of connectives. Finally, the more participants looked away from the screen, the lower IELTS task response, lexical resources, accuracy and total scores they received. They also produced less complex sentences with fewer clauses.

In summary, the following broad trends were observed. First, less fluent writing was associated with lower IELTS scores; less sophisticated language use, and more extensive use of causal connectives. Second, more frequent pausing between lower textual units was linked to the use of less sophisticated lexis. Third, greater frequency of pauses predicted lower IELTS scores; less sophisticated lexis; lower syntactic complexity; and larger number of causal but fewer contrastive connectives. Fourth, more higher-order revisions predicted more sophisticated lexis. Finally, gazing at the previous word/expression, paragraph, and off-screen during pauses was linked to lower text quality, whereas re-visiting the instruction predicted higher syntactic and discourse complexity. See Table 18.

These results run counter to the findings of Stevenson et al. (2006) who found no links between revision behaviours and text quality. However, this might have been due to the fact that Stevenson et al. utilised broader measures of text quality (content and language quality ratings), which might not have been sensitive enough to detect some links. Our findings partially replicate those of Spelman Miller et al. (2008) since we identified a positive link between fluency and text quality. Unlike Spelman Miller et al., however, we also observed significant associations between text quality and some indices of pausing. Like Stevenson et al. (2006), Spelman Miller et al. (2008) employed a broad measure of text quality (composite score of content, range, complexity and accuracy), which again might account for the discrepancy between the results of the two studies.

Table 18: *Significant links between writing behaviours and text quality*

Writing behaviour	Text quality	rho
Fluency		
Minutes per word	IELTS task response	-.61
	IELTS lexical resources	-.53
	IELTS total	-.53
	Clause/t-unit	-.48
Words per P-burst	New-GSL 1000	-.56
	Causal connectives	-.47
Chars per P-burst	New-GSL 1000	-.53
	Causal connectives	-.47
Pause length		
Total	New-GSL 1000	.49
Within words	New-GSL 1000	.48
Between words	New-GSL 1000	.48
Pause frequency		
Total	New-GSL 1000	.56
	Causal connectives	.51
Within words	IELTS task response	-.51
	IELTS lexical resources	-.53
	IELTS total	-.50
	Causal connectives	.48
Between words	Contrastive connectives	-.51
Between sentences	Structural similarity	.53
	Words per t-unit	-.52
Between paragraphs	IELTS task response	-.53
Revision		
Sentence	Off-list words	.55
Location of eye-gazes		
Word or expression	New-GSL 1000	.47
Paragraph	IELTS task response	-.50
Instruction	Words per clause	.47
	All connectives	
Off-screen	IELTS task response	-.49
	IELTS lexical resource	-.55
	IELTS accuracy	-.51
	IELTS total	-.60
	Clause per t-unit	-.49

6.4 To what extent are phonological short-term memory, visual short-term memory, and executive control related to online writing behaviours and text quality?

To address the relationship of the working memory measures to the indices of writing behaviours and text quality, we carried out another series of Spearman correlations and found a small number of significant links. These are summarised in Table 19, grouped according to working memory measures. First, those with higher phonological short-term memory produced more New-GSL 100 words. Second, participants who had superior visual-spatial span gazed at the instructions less frequently while pausing. Third, updating ability was associated with less frequent pausing between paragraphs. Finally, less advanced task-switching skills predicted longer pauses between sentences and the use of less sophisticated lexis and fewer connectives.

These results overall run counter to the patterns observed by Kormos and Sáfár (2008) and Adams and Guillot (2008), who both observed a positive link between phonological short-term memory and text quality. Our results also differ in that we did find significant, positive correlations between executive control and some of the text quality measures. These differences might have been due to the distinct background of the participants in the studies, as well as the different measures of working memory and text quality utilised. Further research is needed to clarify the associations of working memory to writing behaviours and text quality.

In addition to the stimulated recall, keystroke-logging, and eye-gaze data, the working memory results supply evidence for the cognitive validity of the IELTS Academic Writing Task 2. In line with Kellogg's model of writing, phonological short-term memory, visual spatial sketchpad and executive functioning were all related to some of the measures of text quality or writing behaviours in the expected direction. This suggests that these working memory components, as described in Kellogg's model, were drawn on during the writing process.

Table 19: *Significant relationships of working memory measures to writing behaviours and text quality*

Working memory measure	Writing behaviour/ Text quality measure	rho
Phonological short-term memory		
Non-word span	New-GSL 1000	.60
Visual-spatial short-term memory		
Corsi block forward	Eye-fixations at instruction during pauses	-.52
Executive control		
Operation span task (updating)	Pause frequency between paragraphs	-.51
Colour shape task (task-switching ability)	Pause length between sentences	.59
	New-GSL 1000	.46
	Logical connectives	.48

7 Conclusion

The results of this study provide evidence from various data sources for the cognitive validity of a version of Task 2 of the IELTS Academic Writing Test. Following Field (2009), we set out to establish cognitive validity by comparing the processes in which L2 test-takers engage with those that native writers adopt when they complete real-life writing tasks. First language writing processes are well documented and theorised, thus we were able to rely on a model of first language writing, Kellogg's (1996) model, as a baseline for this comparison. The stimulated recall comments demonstrated that the cognitive processes elicited by Task 2 of the IELTS Academic Writing Test are well aligned with the writing stages and sub-stages captured in Kellogg's model. Parallel to the stimulated recall comments, the keystroke logging indices of pausing and revision, along with the eye-gaze data, supply further confirmation that the IELTS Academic Writing Task 2 encourages test-takers to engage in cognitive processes that resemble those that native writers adopt, including both lower and higher-level writing processes. This was reflected in the fact that participants paused and revised at various text level units and gazed at various levels of previously produced texts, similar to first language writers as documented in a number of studies (e.g., Stevenson, 2006). Finally, we found evidence that components of working memory that are assigned a role in Kellogg's model (phonological short-term memory, visual spatial sketchpad and executive control) are implicated when L2 users complete Task 2 of the IELTS Academic Writing Test. Together, these findings provide evidence from various sources that the type of writing processes in which test-takers engaged in this study reflect those that first language writers employ when they produce written pieces.

Although this study yielded some interesting insights, it has a number of limitations, which should be addressed in future research. First, a major limitation of this study has to do with the relatively homogeneous background of the participants, both in terms of L1 (Mandarin) and level of L2 English, which, for some of the measures, resulted in little variation among text quality, keystroke-logging and eye-gaze indices. This inevitably restricted the chance of finding correlations between the measures of writing behaviours, working memory, and text quality. Thus, a potential avenue for future research would involve exploring the research questions addressed here for a wider range of L1 backgrounds and proficiency levels. A second limitation of this project lies in the fact that only one version of the IELTS Academic Writing Task 2 was used to elicit writing performances. Another interesting area of follow-up research would be to repeat the study utilising several versions of this test, as well as different types of writing assessments, in order to test the generalisability of the findings. Third, ideally we would have collected stimulated recall data from all our participants. In future research, researchers could collect introspective data from a larger group of participants, which would enable for inferential statistics to be conducted. Finally, it would be worthwhile to explore how additional individual differences among test-takers, such as anxiety, creativity, and personality, might influence writing processes and products.

References

Adams, A.M. & Guillot, K. (2008). Working memory and writing in bilingual students. *International Journal of Applied Linguistics*, vol 156, pp 13-28.

Altgassen, M., Vetter, N.C., Phillips, L.H., Akgün, C. & Kliegel, M. (2014). Theory of mind and switching predict prospective memory performance in adolescents. *Journal of Experimental Child Psychology,* vol 127, pp 163-175.

Baddeley, A.D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences,* vol 4, pp 417-423.

Baddeley, A.D. & Hitch, G.J. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation*, (pp 47-90). New York: Academic Press.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing,* vol 30, pp 441–465.

Brunfaut, T. & McCray, G. (2015). Looking into test-takers' cognitive processes whilst completing reading tasks: A mixed-method eye-tracking and stimulated recall study. *ARAGs Research Reports*, vol. 1, no. 1. London: British Council.

Breetvelt, I., Van den Bergh, H. & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: When and how? *Cognition and Instruction,* vol 12, pp 103-123.

Brezina, V. & Gablasova, D. (2013). Is There a Core General Vocabulary? Introducing the New General Service List. *Applied Linguistics,* vol 36, pp 1-22.

Congdon, E., Mumford, J.A., Cohen, J.R., Galvan, A., Canli, T. & Poldrack, R.A. (2012). Measurement and reliability of response inhibition. *Frontiers in Psychology*, vol 3, pp 1-10.

Cushing Weigle, S. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

DeKeyser, R.M. (2012). Interactions between individual differences, treatments, and structures in SLA. *Language Learning,* vol 62, pp 189-200.

Enticott, P.G., Ogloff, J.R. & Bradshaw, J.L. (2006). Associations between laboratory measures of executive inhibitory control and self-reported impulsivity. *Personality and Individual Differences,* vol 41, pp 285-294.

Field, J. (2009). The cognitive validity of the lecture listening section of the IELTS listening paper. *IELTS Research Reports Vol 9*. Canberra: IELTS Australia and London: British Council.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Examining speaking*. Cambridge: Cambridge University Press.

Friedman, N.P., Miyake, A., Corley, R.P., Young, S.E., DeFries, J.C. & Hewitt, J.K. (2006). Not all executive functions are related to intelligence. *Psychological science,* vol 17, pp 172-179.

Gold, B.T., Kim, C., Johnson, N.F., Kryscio, R.J. & Smith, C.D. (2013). Lifelong bilingualism maintains neural efficiency for cognitive control in aging. *The Journal of Neuroscience,* vol 33, 387-396.

Halliday, M.A.K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning,* vol 63, pp 87-106.

Kellogg, R.T. (1996). A model of working memory in writing. In C.M. Levy & S. Ransdell (Ed.s), *The science of writing: Theories, methods, individual differences and applications* (pp 57-71). Mahwah, NJ: Lawrence Erlbaum.

Kessels, R.P., Van Zandvoort, M.J., Postma, A., Kappelle, L.J. & De Haan, E.H. (2000). The Corsi block-tapping task: standardization and normative data. *Applied Neuropsychology,* vol 7, pp 252-258.

Kormos, J. (2012). The role of individual differences in L2 writing. *Journal of Second Language Writing,* vol 21, pp 390-403.

Kormos, J. & Sáfár, A. (2008). Phonological short-term memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition,* vol 11, pp 261-271.

Leijten, M. & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication,* vol 30, pp 358-392.

Malvern, D. & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Ed.s), *Evolving models of language (*pp 58–71). Clevedon, England: Multilingual Matters.

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics,* vol 33, pp 299- 320.

Mazgutova, D. & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing,* vol 29, pp 3-15.

McCarthy, P.M. & Jarvis, S.A. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behaviour Research Methods,* vol 42, pp 381-392.

McNamara, D.S., Louwerse, M.M., Cai, Z. & Graesser, A. (2005). Coh-Metrix (Version 1.4) [computer software]. Retrieved from http//:cohmetrix.memphis.edu

Miyake, A., Emerson, M.J., Padilla, F. & Ahn, J.C. (2004). Inner speech as a retrieval aid for task goals: The effects of cue type and articulatory suppression in the random task cuing paradigm. *Acta psychologica,* vol 115, pp 123-142.

Norris, J.M. & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics,* vol 30, pp 555-578.

Polio, C. (2012). Second language writing. In S. Gass & A. Mackey (Ed.s), *Handbook of second language acquisition* (pp 319-334). New York: Routledge.

Plonsky, L. & Oswald, F.L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning,* vol 64, pp 878-912.

Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning,* vol 62, pp 93–132.

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics,* vol 35, pp 87-92.

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

Roca de Larios, J., Manchon, R.M., Murphy, L. & Marin, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing,* vol 17, pp 30-47.

Shaw, S. & Weir, C.J. (2007). *Examining writing: Research and practice in assessing second language writing*, Cambridge University Press, Cambridge

Spelman Miller, K., Lindgren, E. & Sullivan, K.P.H. (2008). The psycholinguistic dimension in second language writing: opportunities for research and pedagogy. *TESOL Quarterly,* vol 42, pp 433-454.

Stevenson, M., Schoonen, R. & Glopper, K. (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing,* vol 15, pp 201-233.

Turner, M.L. & Engle, R.W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language,* vol 28, pp 127-154.

Unsworth, N., Heitz, R.P., Schrock, J.C. & Engle, R.W. (2005). An automated version of the operation span task. *Behavior Research Methods,* vol 37, pp 498-505.

Van Weijen, D. (2009). Writing processes, text quality, and task effects. *Empirical studies in first and second language writing,* Vol 201. Utrecht: LOT.

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. *Computer key-stroke logging and writing: methods and applications (Studies in Writing),* vol 18, pp 107-130.

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

Williams, J.N. (2012). Working memory and SLA. In S. Gass & A. Mackey (Ed.s), *Handbook of second language acquisition (*pp 427-441). New York: Routledge.

Zhao, Y. (2013). Working memory and corrective recasts in L2 oral production. *Asian Journal of English Language Teaching,* vol 23, pp 57-82.