EMPIRICAL STUDY 

Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity

Andrea Révész,[a] Nektaria-Efstathia Kourtali,[a] and Diana Mazgutova[b]

[a]University College London and [b]Lancaster University

This study investigated whether task complexity influences second language (L2) writers' fluency, pausing, and revision behaviors, and the cognitive processes underlying these behaviours; whether task complexity affects linguistic complexity of written output; and whether relationships between writing behaviors and linguistic complexity are moderated by task complexity. Participants were 73 advanced L2 writers, who completed simple or complex essay tasks. Task complexity was operationalized as the absence versus presence of content support. Participants' writing behaviors were recorded via keystroke logging software. Four writers, drawn from groups performing simple and complex tasks, additionally engaged in stimulated recall. Content support was found to lead to less pausing, more revision, and increased linguistic complexity. When content support was absent, more frequent pauses and revisions were associated with less sophisticated lexis. These results, combined with stimulated recall comments, suggest that content support likely reduced processing burden on planning processes and thereby facilitated attention to linguistic encoding.

Keywords task complexity; second language writing; fluency; pausing; revision; linguistic complexity

Author Note

This study was supported by a grant from Trinity College London. The authors would like to thank Elaine Boyd, Roger Gilabert, Gareth McCray, and John Rogers for their insightful comments on earlier versions of this article. We are also grateful to the anonymous reviewers for their very helpful suggestions and the editor, Pavel Trofimovich, for his meticulous and constructive feedback on the manuscript,

This article has been awarded an Open Materials badge. All materials are publicly accessible in the IRIS digital repository at http://www.iris-database.org. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Correspondence concerning this article should be addressed to Andrea Révész, UCL Institute of Education, University College London, Room 623b, 20 Bedford Way, London WC1H OAL. E-mail: a.revesz@ucl.ac.uk

Introduction

The role of tasks in second language (L2) teaching, learning, and assessment has been the object of an increasing amount of research in recent years. This growing interest in tasks has largely been due to the fact that tasks offer an optimal platform for combining meaning- and form-based L2 instruction and assessment, while engaging learners in communicative tasks which have high face validity. Within this active area of research, a key objective has been to investigate the effects of task complexity (i.e., inherent cognitive demands of tasks) on L2 performance and development, with the goal of establishing task grading and sequencing

criteria in order to inform curricular decisions (Robinson, 2001; Skehan, 1998) and specifying factors contributing to the difficulty of L2 assessments (Brown, Hudson, Norris, & Bonk, 2002). To date, researchers have primarily been concerned with exploring the impact of task complexity on the skill of speaking. It is only relatively recently that the issue of how task complexity may affect L2 writing has begun to attract researchers' attention.

Similar to research on speaking, most studies on writing have focused on the relationship between task complexity and the products of task performance, in particular, the linguistic quality of written output (e.g., Byrnes & Manchón, 2014; Kormos, 2011; Kuiken & Vedder, 2007, 2008; Ong & Zhang, 2010). To date, few empirical studies have looked into how cognitive complexity of tasks may influence L2 writers' online behaviors (i.e., directly observable features of the writing process, such as pausing and revision phenomena) and the underlying cognitive processes in which writers engage during writing (e.g., planning, linguistic encoding). The scarcity of research on writing behaviors and associated cognitive processes leaves an important gap in task-based research (Macaro, 2014; Révész, 2014). Clearly, we cannot fully test models of writing (e.g., Kellogg, 1996) in relation to task complexity without examining the causal processes that task manipulations are predicted to generate. For theory-building purposes, it is therefore essential that we gain evidence about task-generated behaviors and the cognitive processes underlying those behaviors (Norris & Ortega, 2003; Révész, 2014); otherwise, we might risk falling into the trap of construct underrepresentation (Norris & Ortega, 2003).

To help address this research gap, this study examined whether task complexity influences fluency, pausing, and revision behaviors of L2 writers and associated cognitive processes, such as planning and linguistic encoding. In addition, this study explored whether relationships between text quality and fluency, pausing, and revision are moderated by task complexity. Within task-based research, this study is methodologically innovative in that it

employed a combination of research methods, including behavioral measures of online keystroke logging and introspective data obtained through stimulated recall. To expand on previous research, this research also investigated the effects of task complexity on the linguistic complexity of L2 writers' output.

We operationalized task complexity as the provision versus no provision of content (ideas) to include in an essay. Our rationale for selecting this task dimension was twofold. On the one hand, it allowed for testing predictions of cognitive models of writing, which attribute a crucial role to planning processes, including the generation of text content. On the other hand, investigating the effects of content support on writing performance is of practical importance. Content support is expected to make it more difficult for L2 writers to avoid complex morphosyntactic and lexical constructions, thereby creating more favorable conditions for linguistic development and assessment of linguistic performance (Kormos, 2011).

Theoretical Background

Kellogg's (1996) model of writing was selected as a theoretical basis for this investigation. This cognitive model, originally developed to account for first language (L1) writing, lends itself well to studying L2 writing processes, given that it makes detailed predictions about linguistic encoding processes, which (compared to L1 writing) are likely to generate considerable cognitive demands for L2 writers. Kellogg describes writing as an interactive and recursive process, which can be defined in terms of three subprocesses: formulation, execution, and monitoring. Formulation involves planning the content of the writing piece and translating it into linguistic form. During the course of planning, writers access ideas from long-term memory and/or task instructions, and organize these to form a coherent plan for the content of the written text. Translating ideas into linguistic form entails the subprocesses of lexical retrieval, syntactic encoding, and expression of cohesion. In the

execution stage, motor movements are employed to produce either a handwritten or typed text. Finally, monitoring involves ensuring that the resulting text is an adequate expression of the writer's intended content. If mismatches are identified between the product and the content planned, then writers may engage in revision. As the subprocesses of formulation, execution, and monitoring interact, complex patterns of processes emerge, which have been suggested to be influenced by a number of variables, including task complexity in the context of L2 writing (e.g., Kormos, 2011).

Kellogg's model, similar to other writing models, makes no direct predictions regarding the relationship between specific task manipulations and L2 writing processes and outcomes. These links, however, have received considerable theoretical and empirical attention in the area of L2 speech production (e.g., Skehan, 2014; Robinson, 2011). The two theoretical models that have been put forward to model task effects in speaking—Skehan's (1998) limited capacity model and Robinson's cognition hypothesis (2001)—have also been adapted to conceptualize task complexity research in writing (e.g., Kuiken & Vedder, 2008). Recently, however, several researchers have convincingly argued that these theoretical frameworks, originally developed for speaking, cannot directly be applied to L2 writing (Kormos, 2011; Kormos & Trebits, 2012; Manchón, 2014; Tavakoli, 2014). One principal reason for this is that there are key differences between the psycholinguistic processes involved in producing speech and in generating written texts. For example, writing is usually less constrained by time than speaking. Thus, writers face less difficulty in dividing their attentional resources among the various stages of the writing process against limitations of their working memory capacity. This means that writers can spend more time planning the content of their message and allocate more attentional resources to translation processes, such as lexical retrieval, syntactic encoding, and the expression of cohesion. Given the cyclical nature of writing, writers also have more opportunities to monitor their performance. Unlike

speakers who can only repair their immediately preceding utterance, writers can go back and revise any previously written part of their text.

Nevertheless, cognitive task demands and attentional limitations are also expected to have an impact on L2 writing behaviors and text quality (Kormos, 2011). Drawing on Kellogg's (1996) model of writing, it appears reasonable to assume that, when task complexity is increased, L2 writers will be less successful at coping with the enhanced demands placed on writing subprocesses due to possible limitations of their working memory. This will probably be reflected in the behaviors they exhibit during task work and in the quality of the output they produce.

Let us consider how the task complexity manipulation in the present study, which was operationalized as provision versus no provision of ideas to include in an essay, is likely to affect writing behaviors and linguistic complexity in light of Kellogg's (1996) model. It is expected that, under the complex condition when content ideas are not available, there will be more pressure on planning processes, as writers are required to access more ideas from long-term memory and to combine these ideas on their own. The absence of ideas will likely also put extra pressure on translating processes by prompting the use of more lexical and grammatical constructions, in comparison to the simple condition where some relevant vocabulary and grammar might appear as part of task instructions (cf. Kormos, 2011). This increased effort involved in planning and translating content is likely to lead to slower task processing. Decreased processing speed, in turn, is anticipated to manifest in a slower pace of writing and a greater number and length of pauses. Pausing might become particularly frequent and lengthy between larger units, such as clauses and sentences, since pauses at these locations are presumably more often associated with planning processes (Schilperoord, 1996). L2 writers might also make fewer language-related revisions because fewer attentional resources are left for monitoring language use. As a consequence, pausing and revision

behaviors might not emerge as predictors of greater linguistic complexity. In general, in the absence of initial content support provided as part of task instructions, writers might produce less linguistically complex texts.

In contrast, when suggestions for content are made available to writers, there is anticipated to be less pressure on planning processes, allowing writers to direct more resources to translation operations, such as lexical retrieval and grammatical encoding (cf. Kormos, 2011). Given the decreased effort required to create content, participants may be able to write faster, as well as pause less frequently and with shorter pause lengths at larger discourse units. However, revisions involving lower discourse units (below the word and clause level) which are typically associated with lexical and syntactic encoding processes, are expected to be more frequent, as writers can focus more on translating their ideas effectively. As a consequence, when ideas are given, those who pause and revise more at lower discourse units will probably write texts of superior lexical and syntactic complexity.

Previous empirical research provides little direct evidence about the validity of these predictions; nevertheless, it is worth considering these predictions in light of existing findings relevant to the effects of task complexity on L2 writers' output and their writing behaviors.

Previous Empirical Research

Task Complexity and L2 Writing Behaviors

Previous work on the link between task complexity and L2 writing behaviors is limited. There seems to exist a single small-scale study (Spelman Miller, 2000) which has directly examined L2 writing behaviors in relation to cognitive task complexity. Spelman Miller investigated whether L2 pausing behavior and fluency differed across computer-delivered descriptive and evaluative essay writing tasks. The evaluative task was assumed to pose greater cognitive demands in that it required critical synthesis and assessment of various viewpoints. While the participants (10 native and 11 L2 writers of English) composed the two

essays, their online keyboard activity was recorded. The resulting log files were analyzed for a number of fluency and pausing indices. Spelman Miller found that participants paused longer at higher levels of text units (i.e., clause and sentence completion points), with the longest pauses occurring between sentences. Furthermore, pauses were most frequent between intermediate constituents, such as noun and verb phrases. These findings were interpreted as suggesting that participants engaged in the most planning at interclause and intersentence locations. Contrary to expectations, neither writing fluency nor pausing was influenced by task differences.

To date, no empirical research has directly investigated the impact of task complexity on L2 revision behaviors. A study by Thorson (2000), however, looked into how revision patterns of L2 writers differ depending on task genre, utilizing keystroke logging software. Each of the 18 participants wrote both a newspaper article and a letter to a pen pal in their L1 English and L2 German. Participants were found to revise more when composing the article than the letter in their L1. However, against predictions, no task effects were observed for the amount and type of L2 revision.

As yet, no studies have looked into how task complexity may influence the cognitive processes that underlie pausing and revision behaviors, despite the fact that cognitive writing processes have received increasing attention in L2 writing research (e.g., Roca de Larios, Manchón, Murphy, & Marín, 2008; Schoonen, Snellings, Stevenson, & van Gelderen, 2009), with a few studies also looking into task effects. Ong (2014), for example, considered how planning time, topic, and availability of content and macrostructure (i.e., guidelines how to organize an essay) influenced writers' metacognitive processes, as reflected in retrospective questionnaire responses. The study revealed, in line with Kellogg's (1990) predictions, that those who received assistance with content and organization dedicated fewer of their

attentional resources to metacognitive processes, such as generating and organizing new ideas, than their counterparts who had no such support available.

L2 Writing Behaviors and Text Quality

The question as to whether L2 text quality relates to fluency, pausing, and revision behaviors also requires further study. Although these associations have been the subject of considerable research in the area of L1 writing (e.g., van den Bergh, Rijlaarsdam, & Breetvelt, 1994), only a few L2 studies have examined this link. Among them is Stevenson, Schoonen, and Glopper's (2006) study, which investigated whether text quality is predicted by type of revision behavior. The participants (22 secondary school students) wrote four essays, two in L1 Dutch and two in L2 English. It was anticipated that, in L2 writing, there would be a negative correlation between lower level revisions (word-level changes) and conceptual text quality, given that writers need to pay more attention to lower level writing processes, leaving less attentional resources available for higher level cognitive operations and, thus, revisions (changes above the word level). Contrary to this prediction, the researchers found no relationship between revision type and text quality. In a more recent study, Tillema (2012) investigated how the temporal distribution of cognitive writing processes, including revising, influenced the quality of texts produced by 14- to 15-year-old writers in their L1 Dutch and L2 English. While engaging in revision at certain points of the writing process had a positive impact on text quality in the case of L1 writing, no such link was found for composing in the L2.

Spelman Miller, Lindgren, and Sullivan (2008) additionally considered whether text quality is predicted by pausing and fluency. The participants were Swedish high school students studying L2 English. The research spanned three years, with each participant writing one essay each year. Writing speed was assessed in terms of fluency, burst (typed characters between pauses and/or revisions), and fluency during burst (writing time between pauses

and/or revisions). Number of revisions (deletions or insertions) were also counted. Two indices of fluency (burst and fluency during burst) were found to be strong predictors of text quality, expressed as a composite score of content, grammatical and lexical range, accuracy, and fluency. Neither pausing nor revision behaviors explained a significant amount of variation among text quality scores. In sum, existing research suggests that text quality may be influenced by fluency, but is not related to revision or pausing. Clearly, more research is needed to confirm these trends. Research is also needed to explore how task complexity may moderate the association between writing behaviors and text quality.

Task Complexity and Linguistic Complexity

An additional aim of this study was to contribute to previous research exploring the effects of task complexity on the linguistic complexity of written production, an area that has received more research attention. Several task complexity dimensions have been investigated in relation to linguistic complexity, including reasoning demands (Kuiken & Vedder, 2008; Ruiz-Funes, 2014), number of task elements (Kuiken & Vedder, 2008; Ruiz-Funes, 2014), planning time (Ellis & Yuan, 2004; Kellogg, 1988, 1990; Ong & Zhang, 2010), revising conditions (Ong & Zhang, 2010), provision of writing support (Ong & Zhang, 2010), storyline complexity (Tavakoli, 2014), narrating in the here-and-now versus there-and-then (Ishikawa, 2007), and telling a story with the content available versus without the content given (Kormos, 2011). The studies conducted by Ong and Zhang (2010) and Kormos (2011) are particularly relevant to the present research since, similar to this study, they focused on the impact of providing help with content on text quality.

Ong and Zhang (2010), as part of a larger study, examined whether writing support affects lexical complexity. The participants were 108 Chinese EFL university students, who were asked to write an argumentative essay under one of three conditions: topic, ideas, and macrostructure given; topic and ideas given; and topic given. Lexical complexity was

assessed by the ratio of word types squared to the total number of words in the final text. The researchers found that, after revising their texts, participants produced more lexically complex texts when ideas alone or ideas and macrostructure were given, as compared to when they only received the essay topic. Notably, this difference was not observed when comparing the initial drafts produced by participants.

Kormos (2011) also looked into the effects of content support, but focused on a wider range of linguistic characteristics. The participants were 44 upper-intermediate EFL learners in a Hungarian secondary school. They were asked to narrate a comic strip consisting of six ordered pictures, then tell a story based on six unrelated pictures. Kormos assumed that the story without a given plot would place greater cognitive load on the participants, requiring them to use their imagination to link the pictures and invent a story around them. Lexical complexity was captured in terms of lexical variability expressed as Malvern and Richard's (1997) D-value, Nation's vocabulary range (Heatley, Nation, & Coxhead, 2002), and the Coh-Metrix indices of frequency and concreteness of content words (McNamara, Louwerse, Cai, & Graesser, 2005). Syntactic complexity was operationalized as subordination complexity, clausal complexity, and phrasal complexity. Out of all the semantic and syntactic complexity measures, only concreteness of content words was influenced by task complexity. Participants showed more extensive use of abstract words when the content was predetermined. Taken together, these two studies suggest that support with content can promote certain aspects of lexical complexity, but that it has no impact on syntactic complexity. Further studies, however, are warranted to confirm these findings.

Research Questions

In light of previous research on L2 writing behaviors, text quality, as well as task and linguistic complexity, the following research questions were formulated:

1. What are the effects of task complexity on L2 writing behaviors and the cognitive processes underlying them?
2. What are the effects of task complexity on the linguistic complexity of L2 texts?
3. Are there links between between L2 writing behaviors and the linguistic complexity of L2 texts? If yes, does task complexity moderate such links?

Task complexity was operationalized as the presence or absence of content support in the form of ideas given to include in an argumentative essay. L2 writing behaviors were operationalized in terms of online measures of speed fluency, pausing, and revision. Cognitive writing processes were investigated by eliciting participants' stimulated recall comments on their internal composing processes. Linguistic complexity was defined using indices of lexical and syntactic complexity (described below).

Method

Design

The participants were 73 L2 speakers of English, all international students at the University of London. Of these participants, 35 students carried out a less complex version of an argumentative essay writing task (simple group), whereas 38 students completed a more complex version of the same task (complex group). Participants were randomly assigned to the simple and complex groups. The participants' online writing processes were recorded by the keystroke logging software *Inputlog 5.2* (Leijten & Van Waes, 2013) and screen capture technology. Eight randomly selected students (four each from the simple and complex groups) were additionally asked to describe their thought processes during task performance via stimulated recall, prompted by the playback of the recordings of their keystrokes and mouse clicks. The nonstimulated recall participants were asked to complete a brief perception questionnaire immediately after the writing task.

Participants

There were a total of 81 participants recruited for the study, but eight participants were excluded due to technical problems. The majority of the remaining 73 participants (65.8%) were enrolled in postgraduate programs, and approximately a third (34.2%) attended undergraduate courses. The participants came from a variety of L1 backgrounds. Approximately a third of participants were Chinese (31.5%), a smaller percentage of students had other Asian (30.1%), European (34.2%), or African (4%) language backgrounds; 59 students were female. The mean age was 30.09 years (*SD* = 7.30). The simple and complex groups had similar demographic characteristics, as shown in Appendix S1 in the Supporting Information online.

The simple and complex groups also had similar writing proficiency. All participants had IELTS writing and overall scores in the 7.0–7.5 range, which is equivalent to level C1 in the Common European Framework of Reference (CEFR). The two groups also achieved comparable scores on a version of the Trinity Integrated Skills in English (ISE) III Correspondence Task (calibrated to CEFR level C1). As summarized in Appendix S2 in the Supporting Information online, independent-samples *t* tests found no significant differences between the simple and complex groups using the rating criteria employed by Trinity examiners to assess the ISE III controlled written examination. Neither did the texts produced by the simple and complex groups on the ISE III Correspondence Task differ along any of the linguistic complexity measures used in this study.

Writing Tasks

In this study, a version of the Trinity Integrated Skills in English (ISE) III Argumentative Writing Task was used as the complex argumentative writing task.[1] According to the test specifications, the task was designed for CEFR level C1. The prompt for the task was:

> Following a discussion about History at school, you have been asked to write an essay giving your opinions on the topic: "When studying the past, it is more important to know about ordinary people than famous people. Do you agree?"

The simple task version additionally provided participants with ideas to include in the essay. Students were asked to consider two issues: (a) what kind of information we can learn about ordinary or famous people of another historical era, and (b) what some of the benefits are of learning about ordinary or famous people. For both issues, participants were presented with subtopics, from which they were encouraged to select and expand on those selected (for the full prompts of the writing tasks, see Appendix S3 in the Supporting Information online). As discussed above, this simple task version was expected to place lower cognitive demands on planning processes than the original Trinity prompt, given that participants needed to exert less mental effort to conceptualize the content of the essay (Kellogg, 1996) due to the availability of ideas. In line with the Trinity exam specifications, the word limit for both essay versions was 200–250 words. Participants had 45 minutes to complete the tasks.

Although originally developed for assessment purposes, this task satisfied oft-cited criteria for defining tasks (e.g., Ellis & Shintani, 2013). The task was likely to generate a primary focus on meaning (participants were only given broad ideas even in the simple condition, so they had to generate specific content on their own), there was an opinion gap between the writer and their expected audience, the writers largely had to resort to their own linguistic resources, and the task had a nonlinguistic outcome in that the assessment criteria, defined according to the Trinity ISE III rating scale, included both nonlinguistic and linguistic categories.

Perception Questionnaire

The aim of the perception questionnaire was to test the validity of the task complexity manipulation, that is, to confirm that the task version designed to be more cognitively

demanding indeed required greater mental effort (Norris & Ortega, 2003; Révész, 2014). The perception questionnaire included five statements that participants judged on a 9-point scale. Of the five statements, three are relevant to the present study. These statements assessed perceptions of (a) overall mental effort exerted, (b) overall task difficulty, and (c) difficulty in planning the essay content. Higher values on the scale indicated greater effort and difficulty.

Stimulated Recall

Stimulated recall was aimed at eliciting participants' thoughts while completing the writing task, in order to determine whether the task version, designed to be differentially complex, generated quantitatively and/or qualitatively differential cognitive processes. Participants watched the screen recording of their own writing performance, and were encouraged to pause the recording at any time when they wished to share the thoughts they had while completing the task. The researcher additionally stopped the screen recording and elicited participants' thoughts when they paused or revised their production. Stimulated recall sessions were conducted in English. Participants did not express difficulty in verbalizing their thoughts given their advanced proficiency.

Data Collection

All participants attended one session during the project. This took about one hour for the nonstimulated recall group and two hours for the stimulated recall group, including a short break. The stimulated recall participants met with a researcher individually, but prior to the stimulated recall, followed exactly the same procedures as the nonstimulated recall participants. All participants completed the writing task in a quiet computer room; their online writing behaviors were captured by the keystroke logging software *Inputlog 5.2* (Leijten & Van Waes, 2013) and screen capture technology. The perception questionnaire was administered immediately after participants finished their essays.

Data Analysis

Online Writing Behaviors

In the analyses of online writing behaviors, we focused on students' production of initial drafts, targeting linear events constituting forward progression as well as nonlinear events, such as revision at the leading edge and in other parts of the text. We excluded texts produced as part of titles, explicit planning episodes (i.e., when writers stop producing full text in order to plan on the screen), revision drafts (i.e., when individuals go back to the beginning of the text and systematically go through, edit, and revise their initial drafts), and end revisions (i.e., when writers revise text outside the final paragraph while working on the final paragraph). Our rationale for excluding these stages was that they involve processes different from those entailed in initial draft production (Baaijen, Galbraith, & de Glopper, 2012). Once we isolated the texts produced as part of the initial drafts, we analyzed the keystroke log files in terms of speed fluency, pausing, and revision.

We operationalized speed fluency narrowly, adopting a process oriented perspective (Abdel Latif 2013; Van Waes & Leijten, 2015). According to Abdel-Latif, valid measurement of speed fluency involves calculating indices of the length of the writer's production units, that is, bursts occurring between pauses (P-burst). In line with this, we utilized four measures to capture speed fluency: 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). In terms of Van Waes and Leijten's multidimensional model of fluency, this operationalization falls into the larger category of production fluency. The threshold for pausing behavior was set at 2 seconds, following conventions in writing research (e.g., Spelman Miller et al., 2008; Wengelin, 2006).

Pausing behavior was expressed in terms of pause frequency and pause length. Specifically, we calculated the number of pauses per 100 words and the mean length of

pauses. Pauses were also classified according to whether they occurred within words, between words, between clauses, or between sentences. These are boundaries typically considered in writing research (e.g., Wengelin, 2006). We treated between-word pauses as one pause, since between-word pauses often include one pause before the spacebar is pressed and one pause before the start of the next word.

Revision behaviors, such as deletions and substitutions, were measured in terms of quantity and type. Revision quantity was assessed by comparing the number of words/characters in the final text and the number of words/characters produced during the entire writing process. Following Stevenson et al. (2006), revisions were additionally coded according to location, whether they occurred below the word level, below the clause level, or at the clause level or above. Ten percent of the data, randomly selected, were analyzed by a second coder for revision location. Intercoder agreement was high (93%).

Stimulated Recall Comments

The analysis of the stimulated recall protocols involved four phases. First, the stimulated recall comments were transcribed. Second, one of the researchers reviewed the comments and, following Kellogg's (1996) model, grouped them into the categories of planning, translation, and monitoring. Comments on planning were further subcategorized into content- and organization-related comments, and comments on translation (when possible) were additionally classified according to whether they included reference to lexical retrieval, syntactic encoding, or use of cohesive devices (see Table 1 for examples of each coding category). Sixteen percent of the time, participants were not able to recall why they paused while writing. In the next step, the researcher double-checked all of the annotations. Finally, the comments falling into a specific category were added up to form a frequency count for each participant and task version. For pausing- and revision-related comments, frequency counts were also calculated by pause location and type of revision. For two participants, who

were randomly selected, the data were double-coded by another researcher. Intercoder agreement was high (94%).

TABLE 1

Written Texts

The written texts produced by the participants were also analyzed for a range of lexical diversity and syntactic complexity measures. The complexity indices were obtained using computer-based text analysis tools. Prior to submitting the texts for machine coding, they were corrected for misspellings and punctuation errors to ensure that the software functioned as intended.

Jarvis (2013) recently suggested that, in order to capture lexical diversity, at least six subconstructs need to be considered, including 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 study examining links among these facets, Jarvis found that volume, evenness, and dispersion are highly correlated. In light of this, we decided to operationalize lexical diversity in terms of rarity, variability, and disparity, given that the participants in this study had to produce texts of the same length (see also Mazgutova & Kormos, 2015).

Using Cobb's (2016) online Vocabprofiler, rarity was measured in terms of proportion of K1 and K2 words (K1 and K2 standing for the first thousand and the second thousand most frequently used words in the English language, respectively), proportion of academic words (AWL; Coxhead, 2000), and proportion of off-list words. Lexical variability was assessed by Malvern and Richards's (1997) D formula and the measure of textual lexical diversity (MTLD; McCarthy & Jarvis, 2010). The estimation of D was performed based on a probabilistic mathematical model that utilizes a series of randomly sampled tokens to create a

type-token ratio curve against increasing token size. MTLD was defined as the mean length of sequential word strings that maintain a given threshold of type-token ratio in a text. Values of D and MTLD were obtained through Coh-Metrix 3.0 (McNamara et al., 2005). Following Jarvis' (2013) suggestion, disparity was assessed by a latent semantic analysis (LSA) index, which was also produced by Coh-Metrix 3.0. This LSA index captured the conceptual similarity of each sentence to every other sentence in the text by considering the semantic overlap between the words in the sentences.

Syntactic complexity was assessed in terms of four types of indices, drawing on recent work by Bulté and Housen (2012) and Norris and Ortega (2009): overall complexity, subordination complexity, phrasal complexity, and syntactic sophistication. Following previous task complexity studies in L2 writing research, t-unit was adopted as the principle unit of analysis (e.g., Kuiken & Vedder, 2008). Overall complexity was expressed as the ratio of words to t-units. Subordination complexity was operationalized as the proportion of clauses in relation to t-units. These indices were calculated by utilizing the text analysis software SynLex (Lu, 2010). As a measure of phrasal complexity, the mean number of modifiers per noun phrase was calculated by Coh-Metrix 3.0. The level of syntactic sophistication was assessed through Coh-Metrix 3.0 using a syntactic structure similarity index. This measure estimates the extent to which syntactic structures are consistent in a text, that is, a lower syntactic structure similarity index reflects more varied use of structures.

Statistical Analyses

First, we analyzed the perception questionnaire data to check the validity of our task complexity manipulation. Independent-samples *t* tests were employed to compare participants' responses across the conditions with and without content support. Next, the data for all measures of writing behaviors and linguistic complexity were inspected for outliers. The outliers were trimmed to values of two standard deviations from the mean for each

measure per group. For pause length, outliers were also identified and trimmed within participants using the same threshold. Then, a series of independent-samples *t* tests were used to compare the effects of task complexity on the indices of writing behaviors (fluency, pausing, and revision) and linguistic complexity (lexical and syntactic complexity). The alpha level was set at .05 for all tests. Cohen's *d* was calculated as a measure of effect size. Following Plonsky and Oswald (2014), values larger than .40, .70, and 1.00 were considered as small, medium, and large, respectively. Pearson correlations were computed to examine the relationships between the writing behavior and linguistic complexity measures. Given the large number of correlations, a more conservative alpha level of .01 was adopted to decrease the chance of Type 1 error. We considered correlation coefficients of .25 as small, .40 as medium, and .60 as large (Plonsky & Oswald, 2014). Standard diagnostic procedures were used to ensure the appropriateness of using *t* tests and parametric correlations.

Results

Validity Evidence for Task Complexity Manipulation

Table 2 provides the descriptive statistics for participants' perceptions of mental effort, task difficulty, and difficulty involved in planning, as reported by the nonstimulated recall participants in the perception questionnaire. These data indicate that, in line with our intended task complexity manipulation, participants rated the simple task version where content support was available as requiring less mental effort, being less difficult, and posing less difficulty in planning the essay content, compared to the complex task version. Independent-samples *t* tests confirmed these differences in ratings to be significant for all three scales: mental effort, $t(63) = -3.21$, $p = .002$, 95% CI = [–1.96, –.45], $d = .79$; task difficulty, $t(63) = -2.42$, $p = .018$, 95% CI = [–1.74, –.17], $d = .60$; difficulty in planning content, $t(63) = -2.03$, $p = .047$, 95% CI = [–1.81, –.01], $d = .51$. The effect size for mental effort was large, whereas the effect sizes were medium for overall task difficulty and difficulty of planning.

TABLE 2

Task Complexity and L2 Writing Behaviors

Table 3 provides the descriptive statistics for the measures of fluency, pausing, and revision behaviors for the simple and complex task versions, as well as the results of the independent-samples *t* tests that compared participants' behaviors under the simple and complex task conditions. Task complexity was found to have a significant, medium-size effect on only two of the indices: number of pauses between sentences and revisions below the word level. When participants were provided with ideas to include in the essay, they were found to pause significantly less frequently between sentences, and they made significantly more revisions below the clause level. The *t* tests yielded no significant results for fluency and the remaining measures of pausing and revision.

TABLE 3

Task Complexity and Cognitive Processes Underlying L2 Writing Behaviors

Table 4 summarizes the stimulated recall comments, which were elicited to shed light on the cognitive processes underlying participants' pausing behavior across the simple and complex task versions (for participant-level breakdown of the data, see Appendix S4 in the Supporting Information online). Under the simple condition when content support was available, the largest percentage of stimulated recall comments referred to translation processes (59%), followed by comments describing planning operations (27%). In contrast, under the complex condition where participants were not provided help with content, considerably more stimulated recall comments concerned planning (48%) than translation (33%). Participants made reference to monitoring with similar frequency on the simple (7%) and complex (6%) task versions. The distribution of subprocesses associated with planning and translation was also similar regardless of task complexity, with the majority of planning comments referring to planning content (simple: 83%, complex: 91%) and most formulation comments focusing

on lexical encoding mechanisms (simple: 51%, complex: 61%). Finally, Table 4 indicates that, while the majority of pauses between sentences were associated with planning (simple: 67%, complex: 81%), pauses between words reflected either planning or translation. In line with the trend observed for the total number of pauses, a larger percentage of planning-related pausing occurred between words when no ideas were made available (22%), compared to when content support was provided (13%). To sum up, the stimulated recall comments suggested that lack of content support led to greater pressure on planning processes, and this was most apparent in the proportionately higher number of planning-related pauses between words when no ideas were made available.

TABLE 4

Table 5 presents the summary of the stimulated recall comments elicited to describe participants' thoughts during revision (for participant-level breakdown of the data, see Appendix S5 in the Supporting Information online). Contrary to what was found for pausing, the distribution of revision-related comments was more evenly spread across the two task complexity conditions. Irrespective of whether content support was provided, participants referred to translation mechanisms far more frequently (simple: 78%, complex: 72%) than to planning processes (simple: 19%, complex: 26%). Under the simple condition, however, there was a slightly higher overall percentage of translation-related comments. The patterns by level of revision (within word, below word, below clause, below sentence) were also largely comparable for the simple and complex essay.

TABLE 5

Task Complexity and Linguistic Complexity

Table 6 gives the descriptive statistics and results of independent-samples *t* tests for lexical diversity and syntactic complexity across the complex and simple task versions. Only three tests yielded significant results. Under the simple condition where content support was

supplied, participants used significantly smaller proportions of K1 words, but larger proportions of K2 words, and produced higher number of words per t-unit. That is, decreased task complexity led to more extensive use of less frequent words and greater overall complexity. The effect sizes for these lexical and syntactic complexity indices were of large and medium size, respectively. Task complexity, however, did not have a significant impact on the remaining nine linguistic complexity measures.

TABLE 6

Task Complexity, Revision Behaviors, and Linguistic Complexity

Table 7 provides a summary of the significant correlations between the writing behavior indices and linguistic complexity measures. For the simple essay, one significant correlation was identified. The essays of those participants who paused longer between clauses included less diverse syntactic structures. For the complex essay, two significant correlations were detected. Participants who paused longer between sentences overall produced essays with less sophisticated lexis, as indicated by the lower number of off-list, rare words included in the texts. More revision at the clause level and above was also associated with less sophisticated lexical choices, reflected in the smaller proportion of academic words in the essays.

TABLE 7

Discussion

Task Complexity, Writing Behaviours, and Cognitive Processes Underlying Writing Behaviors

Our first research question was concerned with the effects of manipulating the cognitive complexity of a writing task on the L2 writing process, in particular, whether providing content support in the form of ideas would influence fluency, pausing, and revision behaviors and the cognitive processes underlying these behaviors. According to a series of independent-samples *t* tests, task complexity did not have a significant impact on fluency, but had a

significant effect on one pausing and one revision index. Under the high task complexity condition, when ideas were not made available, participants paused more frequently between sentences and revised less below the word level. The stimulated recall data revealed that there was also variation in the processes underlying pausing and revision behaviors, depending on whether content support was available. Absence of content support led to more planning- than translation-related pauses and revisions, but this difference was substantially more pronounced in the stimulated recall comments elicited in response to pausing behaviors.

The findings for speed fluency ran counter to our predictions, but confirmed the findings of Spelman Miller's (2000) small-scale study. A possible explanation for the lack of effects for speed fluency might lie in the relative resistance of this construct to task differences in writing, a suggestion also put forward by Spelman Miller. In line with this possibility, De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2013) found that, in L2 speaking, linguistic knowledge and skills had the strongest association with speed fluency among measures of speed, breakdown, and repair fluency. Assuming that the same applies in L2 writing, it could be speculated that the impact of task complexity on speed fluency is negligible in magnitude relative to the influence of proficiency, a variable that was controlled for in this study.

The results for pause frequency were largely aligned with our expectations. Drawing on Kellogg's (1996) model, it was anticipated that lack of content support would increase pressure on planning, and this would lead to more extensive pausing at higher level discourse units, given that pauses at higher level constituents often reflect higher order writing processes (Schilperoord, 1996), such as creating content and considering organization. Indeed, in the complex group, the keystroke logs showed evidence of significantly more pauses between sentences, and the stimulated recall comments confirmed that pauses at sentence boundaries were associated with planning operations in the majority of cases. The

increased effort required in planning under the complex condition was further manifest in the overall proportionately higher planning- than translation-related stimulated recall comments produced by the complex group. These patterns are in line with Ong's (2014) observation that, when participants were not provided assistance with content and organization, they were more likely to engage in metacognitive processes, such as generating and organizing ideas. Our findings, however, contradict those of Spelman Miller (2000), who found no effects for task complexity in examining pause frequency.

Our results for pause length, on the other hand, ran parallel to the patterns detected by Spelman Miller (2000). In neither study did task complexity have a significant relationship with pause duration. Interestingly, in Spelman Miller et al.'s (2008) three-year longitudinal investigation, duration of pausing also remained stable, while pause frequency decreased over time. Similarly, De Jong et al. (2013) identified no association between duration of silent pauses and linguistic skills in L2 speaking, but found significant effects for pause frequency. It is possibly the case that the length of time writers pause, on average, might be resistant to factors such as task complexity and proficiency; length of pausing might instead be determined by personal writing style or personality characteristics, as was speculated by De Jong et al. in the context of research on L2 speaking fluency.

Finally, it is worth highlighting a finding in relation to pausing at sentence boundaries. Participants, when prompted to recall their thoughts during pauses between sentences, tended to refer to planning-related processes regardless of task complexity. In contrast, at word boundaries, the distribution of planning- versus translation-related stimulated recall comments differed across the two task complexity conditions. Under the simple condition, pauses between words generated more comments describing translation than planning operations. Taken together, this suggests that in L2 writing, similar to what was suggested for L1 writing (Schilperoord, 1996), pausing at higher discourse units is more likely to be

associated with higher level writing processes, even when there are sufficient cognitive resources available to allow for both translation and planning. Clearly, more research is warranted to ascertain this finding, especially for writers at lower proficiency.

Turning to revision behaviors, our results were partly consistent with what was anticipated based on Kellogg's model. We speculated that in the absence of content support, the enhanced demands posed on planning processes would leave participants with fewer attentional resources to allocate to translation and monitoring. This, in turn, was expected to lead to a decrease in language revisions. We found that under the simple condition, significantly more below-word revisions were recorded in the keystroke logs and, in both the simple and complex groups, a considerably larger percentage of stimulated recall comments were recorded at the level of word-described translation than planning processes. These results together indicate that, as expected, participants made more language-related revisions below the word level under the low task complexity condition. Our prediction received some further confirmation from the stimulated recall comments, since a slightly higher percentage of the revision comments referred to translation- than planning-related mechanisms. It is worth pointing out, however, that although the trends for revision below the clause level and clause and above were also in the expected direction across the two task conditions, these differences did not reach significance. Neither were any task effects detected for total amount of revision, a finding consistent with that of Thorson (2000).

Task Complexity and Linguistic Complexity

Our second research question asked whether task complexity affected the linguistic complexity of L2 texts, as measured by indices of lexical diversity and syntactic complexity. We hypothesized that the availability of ideas would facilitate increased linguistic complexity, drawing on insights derived from Kellogg's (1996) model. Of the 12 independent-samples *t* tests, only three yielded significant results, but each was consistent

with our prediction. When content support was provided to participants, they produced texts with substantially more sophisticated vocabulary and superior overall syntactic complexity. Our findings for lexical complexity were well aligned with those of Kormos (2011). Ong and Zhang (2010) also found similar trends for revision drafts, although these patterns were not observed for initial drafts as in this research. Unlike the present study, however, Kormos (2011) found no effects for syntactic complexity.

Task Complexity, L2 Writing Behaviours, and Linguistic Complexity

The final research question of this study examined whether there were links between L2 writing behaviors and linguistic complexity, and if yes, how these relationships were moderated by task complexity. Our prediction was that pausing and revision behaviors would probably not be positively linked to linguistic complexity in the high task complexity group, given that participants in the absence of content support would be more likely to pause to engage in planning and be less able to make language-related revisions due to decreased attentional resources. In contrast, under the simple condition, we expected that linguistic complexity would be positively linked to the amount of pausing and revision made, given the decreased demands on planning operations. Indeed, no positive correlations were detected between linguistic complexity and the measures of pausing and revision when no ideas were made available. In fact, greater amounts of pausing between sentences and revision at the clause level and above were associated with the use of less sophisticated lexis. This finding corresponds well to the fact that the complex group participants, as compared to their counterparts in the simple group, more often reported to be engaged in planning- than translation-related processes when asked to recall their thoughts during pauses between sentences and revisions at the clause level and above. It would appear, then, that more time devoted to planning left fewer resources available for lexical encoding processes. At odds with our expectation, however, less varied syntax was utilized by those who paused more

between clauses under the simple condition. A possible explanation for this finding is that greater length of pausing between sentences was a manifestation of less advanced syntactic knowledge. Finally, it is worth noting that our results were different from those of existing research in that text quality was found to be related to pausing and revision behaviors (cf. Spelman Miller et al., 2008; Stevenson et al., 2006), but not to fluency (Spelman Miller et al., 2008).

Implications

The findings of this study are of theoretical, methodological, and pedagogical significance. At the theoretical level, the fact that our results confirmed predictions derived from Kellogg's (1996) cognitive model of L1 writing, suggests that this model (and possibly other L1 models of writing) can offer a suitable starting point for conceptualizing research on task-generated L2 pausing and revision behaviors and the processes underlying them. This is an important outcome, given that the appropriateness of using task-based models of L2 speech production as a theoretical basis for investigating task-based writing has been questioned (e.g., Kormos, 2011; Kormos & Trebits, 2012; Manchón, 2014; Tavakoli, 2014). On the methodological front, our study has revealed that the combination of stimulated recall and keystroke logging has the capacity to yield more valid and accurate interpretations about task-generated processes than relying on either of the data sources alone. Turning to pedagogy, a potential implication of this research is that the provision of content support, albeit compromising authenticity, may allow learners to dedicate more attention to linguistic encoding processes, which in turn might be beneficial in terms of "stretching" their interlanguage.

Limitations and Future Research

In evaluating these findings, it is also important to acknowledge the limitations of the study. One methodological shortcoming lies in the fact that we applied a pause threshold of 2 seconds. Although 2 seconds has been the typical threshold utilized in writing research to

date, which facilitates the comparability of our research to previous studies, it has been argued (e.g., Baaijen et al., 2012) that this operationalization constrains the analysis to longer pauses which are likely to reflect higher level writing processes and excludes shorter pauses that are more often associated with lower level processing. There are also weaknesses associated with the between-subjects design utilized in this research. Although we established that the two groups were comparable, future studies in this area could employ within-subjects designs by requiring writers to produce several texts in order to further increase the generalizability of the findings.

Another limitation concerns the use of the stimulated recall methodology. While tackling issues with reactivity, an inherent limitation in this procedure is that it can only provide information about conscious operations and, due to memory decay, only a subset of the conscious processes during writing are likely to be recalled by participants. In further research, this problem could be mitigated by employing introspective protocols together with data from eye tracking. This combination would enable researchers to obtain a more complete picture of the processes occurring during L2 text production, providing insights not only on conscious but also some of the unconscious operations involved (see Brunfaut & McCray, 2015). Another interesting direction for future research would be to analyze revision drafts in addition to initial text production to more fully capture the writing process. Other important avenues for future research would include extending the research questions here to other task complexity manipulations, task types, and populations. This study involved a single task type, only one task complexity operationalization, and advanced L2 writers. It would be interesting to examine whether our findings would transfer to other task types, different task complexity manipulations, and lower proficiency writers.

Conclusion

The primary aim of this study was to initiate investigation into how task complexity manipulations may affect the online behaviors of L2 writers and the associated underlying cognitive processes, in an attempt to address the gap in current task-based research on explanatory processes mediating the relationship between task complexity manipulations and the linguistic product of the writing process. In addition, we intended to launch research into whether task complexity may moderate the relationships between writing behaviors and linguistic complexity, and extend existing research by investigating the effects of task complexity on linguistic complexity. In the context of task-based research, the methodological innovation of the current study was in the triangulation of data obtained from keystroke logging, stimulated recall, and computer-based textual analysis.

Largely in line with predictions derived from Kellogg's (1996) model of writing, we found that the availability of content support led to less frequent pausing and greater amount of revision, and resulted in increased lexical complexity. When content support was present, more frequent pauses were also associated with the production of more lexically complex language. These results, combined with insights which emerged from the stimulated recalls, were interpreted as suggesting that the availability of ideas, as anticipated based on Kellogg's model, allowed participants to dedicate more attention to linguistic encoding processes, which in turn led to the observed effects on writing behaviors and text quality.

Final revised version accepted 6 June 2016

Note

1 This task was provided by Trinity College London, the testing board who funded our research.

References

Abdel Latif, M. M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics, 34,* 99–105. doi:10.1093/applin/ams073

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

Brown, J. D., Hudson, T. D., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessment*. Honolulu, HI: University of Hawai'i Press.

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

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.

Byrnes, H., & Manchón, R. (2014) (Eds.). *Task-based language learning: Insights from and for L2 writing*. Amsterdam: John Benjamins. doi:10.1075/tblt.7

Cobb, T. (2016). VocabProfiler [Computer software]. Retrieved from http://www.lextutor.ca/vp/eng

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*, 213–238. doi:10.2307/3587951

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, *34*, 893–916. doi:10.1017/S0142716412000069

Ellis, R, & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. New York: Routledge.

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26,* 59–84. doi:10.1017/S0272263104026130

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range and Frequency programs [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/resources/range.aspx

Ishikawa, T. (2007). The effect of manipulating task complexity along the [+/–Here-and-Now] dimension on L2 written narrative discourse. In C. M. Garcia-Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 136–156). Clevedon, UK: Multilingual Matters.

Jarvis, S. (2013). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 13–45). Amsterdam: John Benjamins. doi:10.1075/sibil.47

Kellogg, R. T. (1988). Attentional overload and writing performance: Effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 355–365. doi:10.1037/0278-7393.14.2.355

Kellogg, R. T. (1990). Effectiveness of prewriting strategies as a function of task demands. *American Journal of Psychology, 103,* 327–342. doi:10.2307/1423213

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

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing, 20,* 148–161. doi:10.1016/j.jslw.2011.02.001

Kormos, J., & Trebits, A. (2012). The role of task complexity, modality and aptitude in narrative task performance. *Language Learning*, *62*, 439–472. doi:10.1111/j.1467-9922.2012.00695.x

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *45*, 261–284. doi:10.1515/iral.2007.012

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, *17*, 48–60. doi:10.1016/j.jslw.2007.08.003

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

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*, 474–496. doi:10.1075/ijcl.15.4.02lu

Macaro, E. (2014). Reframing task performance: The relationship between tasks, strategic behaviour, and linguistic knowledge in writing. In H. Byrnes & R. Manchón, R. (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 53–77). Amsterdam: John Benjamins. doi:10.1075/tblt.7

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

Manchón, R. (2014). The internal dimension of tasks: The interaction between task factors and learner factors in bringing about learning through writing. In H. Byrnes & R. Manchón, R. (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 27–52). Amsterdam: John Benjamins. doi:10.1075/tblt.7

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3–15. doi:10.1016/j.jslw.2015.06.004

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*, 381–392. doi:10.3758/BRM.42.2.381

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2005, January 1). Coh-Metrix version 1.4 [Computer software]. Retrieved from http//:cohmetrix.memphis.edu

Norris, J., & Ortega, L. (2003). Defining and measuring L2 acquisition. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Malden, MA: Blackwell.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*, 555–578. doi:10.1093/applin/amp044

Ong, J. (2014). How do planning time and task conditions affect metacognitive process of L2 writers? *Journal of Second Language Writing*, *23*, 17–30. doi:10.1016/j.jslw.2013.10.002

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing*, *19*, 218–233. doi:10.1016/j.jslw.2010.10.003

Plonsky, L., & Oswald, F. L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. doi:10.1111/lang.12079

Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, *35*, 87–92. doi:10.1093/applin/amt039

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for investigating task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). New York: Cambridge University Press.

Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance*. Amsterdam: John Benjamins.

Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing*, *17*, 30–47. doi:10.1016/j.jslw.2007.08.005

Ruiz-Funes, M. (2014). Task complexity and linguistic performance in advanced college-level foreign language writing. In H. Byrnes & R. Manchón, R. (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163–191). Amsterdam: John Benjamins. doi:10.1075/tblt.7

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

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen, A. (2009). Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language writing. In R. M. Manchón (Ed.), *Learning, teaching, and researching writing in foreign language contexts.* Bristol, UK: Multilingual Matters.

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

Skehan, P. (Ed.) (2014). Processing perspectives on task performance. Amsterdam: John Benjamins.

Spelman Miller, K. (2000). Academic writers on-line: Investigating pausing in the production of text. *Language Teaching Research*, *4*, 123–148. doi:10.1177/136216880000400203

Spelman Miller K., Lindgren E., & Sullivan K. P. H. (2008). The psycholinguistic dimension in second language writing: Opportunities for research and pedagogy using computer keystroke logging. *TESOL Quarterly*, *42*, 433–453. doi:10.1002/j.1545-7249.2008.tb00140.x

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*, *15*, 201–233. doi:10.1016/j.jslw.2006.06.002

Tavakoli, P. (2014). Storyline complexity and syntactic complexity in writing and speaking tasks. In H. Byrnes & R. Manchón, R. (Eds.), *Task-based language learning: Insights from and for L2 writing* (pp. 163–191). Amsterdam: John Benjamins. doi:10.1075/tblt.7

Thorson, H. (2000). Using the computer to compare foreign- and native-language writing processes: A statistical and case study approach. *Modern Language Journal*, *84*, 55–70. doi:10.1111/0026-7902.00059. 2000

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

van den Bergh, H., Rijlaarsdam, G., & Breetvelt, I. (1994). Revision process and text quality: An empirical study. In G. Eigler & T. Jechle (Eds.), *Writing: Current trends in European research* (pp. 133–148). Freiburg, Germany: Hochschul Verlag.

Van Waes, L., & Leijten, M. (2015). Fluency in writing: A multidimensional perspective on writing fluency applied to L1 and L2. *Computers and Composition, 38*, 79–95. doi:10.1016/j.compcom.2015.09.012 8755-4615

Wengelin, A. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. H. Sullivan & E. Lindgren (Eds.), *Computer keystroke logging and writing: Methods and applications* (pp.107–130). Oxford, UK: Elsevier Science.

Supporting Information

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

Appendix S1. Background Information per Group.

Appendix S2. Descriptive and Inferential Statistics for Lexical Diversity and Syntactic Complexity on the Trinity ISE III Correspondence Task.

Appendix S3. Prompts for Simple and Complex Conditions.

Appendix S4. Reasons for Pausing: Summary of Stimulated Recall Comments.

Appendix S5. Reasons for Revision: Summary of Stimulated Recall Comments.

Table 1 Examples for stimulated recall comments

Process/subprocess	Example
Planning	
Content	Here I'm thinking again about the statement honestly I asked myself what's the most important thing to know about ordinary people or about the famous ones and it was like a dilemma. (pausing)
	I wanted to say something why we should be aware about famous people like they offered many things but then I wanted to say I changed my mind because I decided not to write it this way as there are there were personalities that didn't offer to the society. (revision)
Organization	Yeah I was thinking whether I should I was seeing whether the word count was enough and if I could finish the essay or if I had to write another paragraph. (pausing)
	I decided to change the structure of the text, of the essay. (revision)
Translation	
Lexical retrieval	Fundamental what? Fundamental theories? Fundamental scientific concepts? So, I was looking for a word and probably the next one is theories. I put scientific theories. (pausing)
	Yeah I was thinking early human what? Human species, human remains. No, species is the wrong word scientifically. So, I needed to put something that indicates human bodies or human remains. I was looking for the best word. (revision)
Syntactic encoding	I was debating whether to use present tense or past tense, I was confused there. (pausing)
	I realized that I didn't need the article. (revision)
Cohesion	I was not sure whether I was supposed to use *and* or *or*. (pausing)

	I think I'm gonna use a linking device instead of repeating using first person. (revision)
Unspecified	I just remember that in this sentence I didn't know how to express myself. (pausing) Yeah I wanted to know cause now I wanted to say that most of what historians assume of how life was like in Rome and in other cities of the Roman empire it's due to Pompei. So, I knew that's what I wanted to say but I was trying to find the right words. (revision)
Monitoring	OK and I think this is where I finished, so I'm just gonna read, I wanted to read once again. (pausing) I was reading the whole sentence again to see whether it made sense. (pausing)

Table 2 Descriptive statistics for perceptions of mental effort and task difficulty

	Simple (n = 31)			Complex (n = 34)		
Rated item	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Mental effort	4.68	1.62	[4.13, 5.24]	5.88	1.41	[5.42, 6.32]
Task difficulty	3.87	1.45	[3.37, 4.38]	4.82	1.70	[4.22, 5.42]
Difficulty in planning content	4.35	1.60	[3.82, 4.95]	5.26	1.97	[4.63, 5.90]

Table 3 Descriptive and inferential statistics for fluency, pausing, and revision behaviors

	Simple (*n* = 35)				Complex (*n* = 38)				Comparison (*t* test)			
Measure	*N*	*M*	*SD*	95% CI	*N*	*M*	*SD*	95% CI	*t*	*p*	95% CI	*d*
Fluency												
Minutes/word	35	0.04	0.01	[.04, .04]	38	0.05	0.01	[.04, .05]	–1.47	.15	[–.009, .001]	.36
Minutes/character	35	0.01	0.00	[.01, .01]	38	0.01	0.00	[.01, .01]	–0.96	.34	[–.001, .000]	.23
Words/P-burst	35	3.69	1.75	[3.13, 4.35]	38	3.56	1.90	[3.05, 4.28]	0.31	.76	[–.69, .92]	.07
Characters/P-burst	35	22.76	10.89	[19.32, 26.91]	38	23.05	14.80	[19.26, 28.70]	–0.09	.93	[–6.47, 4.92]	.02
Pause length in milliseconds (log)												
Total	35	8.44	0.14	[8.39, 8.49]	38	8.45	0.17	[8.39, 8.51]	0.04	.97	[–.07, .08]	.06
Within words	34	8.18	0.33	[8.08, 8.30]	37	8.17	0.24	[8.08, 8.23]	0.21	.84	[–.12, .15]	.03
Between words	35	8.40	0.16	[8.34, 8.46]	38	8.39	0.18	[8.33, 8.44]	0.37	.71	[–.06, .09]	.16
Between clauses	31	8.43	0.33	[8.29, 8.52]	34	8.47	0.21	[8.38, 8.52]	–0.56	.58	[–.18, .10]	.14
Between sentences	33	8.67	0.28	[8.57, 8.78]	37	8.67	0.32	[8.56, 8.78]	–0.20	.84	[–.01, .07]	<.01
Number of pauses per 100 words (log)												
Total	35	3.29	0.39	[3.13, 3.39]	38	3.34	0.57	[3.11, 3.51]	–0.41	.69	[–.28, .18]	.10
Within words	34	0.67	1.00	[.34, 1.02]	37	0.88	0.87	[.61, 1.15]	–0.95	.34	[–.65, .23]	.22

Between words	35	2.79	0.71	[2.55, 3.03]	38	2.73	0.91	[2.43, 3.02]	0.34	.74	[–.32, .45]	.07
Between clauses	31	1.17	0.55	[.99, 1.41]	34	1.42	0.76	[1.14, 1.67]	–1.50	.14	[–.58, .08]	.38
Between sentences	35	1.25	0.50	[1.09, 1.51]	38	1.56	0.69	[1.33, 1.78]	–2.19	.03	[–.59, –.03]	.51
Revision overall: Number of words/characters in product per number of words/characters written during process												
Words		0.78	0.13	[.74, .82]		0.77	0.12	[.73, .81]	0.36	.72	[–.05, .07]	.08
Characters		0.74	0.12	[.70, .78]		0.73	0.12	[.69, .76]	0.69	.50	[–.04, .08]	.08
Number of revisions per 100 words (log)												
Below word	35	3.22	0.62	[3.01, 3.42]	38	2.91	0.69	[2.69, 3.14]	1.97	.05	[<–.01, .61]	.47
Below clause	35	1.26	0.39	[1.13, 1.39]	38	1.12	0.35	[1.00, 1.23]	1.53	.13	[–.04, .31]	.38
Clause and above	27	–.05	0.41	[–.19, .10]	25	–0.11	0.32	[.–.23, .02]	0.62	.54	[–.14, .27]	.16

Table 4 Reasons for pausing (number of comments) from stimulated recalls

Pause location	Planning			Translation				Monitoring	No recall	Total[b]
	Content	Organization	All[a]	Lexical retrieval	Syntactic encoding	Cohesion	All[a]			
Simple										
Within words	1	0	1 (1%)	2	1	0	4 (2%)	0 (0%)	7 (4%)	12 (7%)
Between words	25	1	26 (13%)	55	4	3	93 (47%)	1 (1%)	7 (4%)	127 (65%)
Between clauses	2	0	2 (1%)	3	3	0	12 (6%)	1 (1%)	1 (1%)	16 (8%)
Between sentences	16	8	24 (13%)	0	1	1	8 (5%)	12 (7%)	1 (1%)	45 (26%)
Total	44	9	53 (27%)	60	9	4	117 (59%)	14 (7%)	16 (8%)	200 (100%)
Complex										
Within words	2	0	2 (1%)	3	1	0	6 (4%)	0 (0%)	8 (6%)	16 (11%)
Between words	30	1	31 (22%)	24	2	0	32 (22%)	0 (0%)	10 (7%)	73 (51%)
Between clauses	9	0	9 (6%)	0	0	0	4 (3%)	0 (0%)	1 (1%)	14 (10%)
Between sentences	22	5	27 (18%)	2	0	1	5 (3%)	8 (5%)	1 (1%)	41 (27%)

Total	63	6	69 (48%)	29	3	1	47 (33%)	8 (6%)	20 (14%)	144 (100%)

Notes. [a]Values for subcategories do not necessarily add up to the total, given that some comments were not specific enough to allow for further subcategorization. [b]Due to rounding some totals do not add up to 100.

Table 5 Reasons for revision (number of comments) from stimulated recalls

Pause location	Planning			Translation				No recall	Total[b]
	Content	Organization	All[a]	Lexical retrieval	Syntactic encoding	Cohesion	All[a]		
Simple									
Below word	5	0	5 (2%)	2	1	2	13 (5%)	2 (1%)	20 (8%)
Single word[c]	11	0	11 (4%)	43	5	11	106 (43%)	5 (2%)	122 (49%)
Below clause	21	0	21 (9%)	15	5	7	51 (21%)	2 (1%)	74 (31%)
Clause and above	8	1	9 (4%)	4	1	2	20 (8%)	0 (0%)	29 (12%)
Total	45	1	46 (19%)	64	12	22	190 (78%)	9 (4%)	245 (100%)
Complex									
Below word	4	0	4 (2%)	6	5	0	24 (10%)	2 (1%)	30 (13%)
Single word[c]	16	0	16 (7%)	20	13	0	72 (29%)	2 (1%)	90 (37%)
Below clause	19	1	20 (8%)	10	7	11	63 (26%)	3 (1%)	86 (35%)
Clause and above	19	4	23 (9%)	2	2	0	17 (7%)	0 (0%)	40 (16%)

Total	58	0	63 (26%)	38	27	11	176 (72%)	7 (3%)	246 (100%)

Notes. [a]Values for subcategories do not necessarily add up to the total, given that some comments were not specific enough to allow for further subcategorization. [b]Due to rounding some totals do not add up to 100. [c]One full word added, deleted, or substituted.

Table 6 Descriptive and inferential statistics for lexical diversity and syntactic complexity

	Simple (n = 35)			Complex (n = 38)			Comparison (t test)			
Measure	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*t*	*p*	95% CI	*d*
Lexical diversity										
K1 words	87.31	4.15	[85.92, 88.60]	90.74	4.09	[89.46, 91.95]	–3.56	.001	[–5.35, –1.51]	.84
K2 words	3.37	1.48	[2.88, 3.88]	2.06	1.21	[1.69, 2.41]	4.16	<.001	[.68, 1.94]	.99
Academic words	5.14	2.25	[4.45, 5.87]	4.24	2.01	[3.61, 4.88]	1.81	.08	[–.09, 1.89]	.43
Off-list words	3.92	2.31	[3.21, 4.71]	3.03	2.24	[2.32, 3.73]	1.68	.10	[–.17, 1.96]	.40
MTLD	67.43	13.83	[63.16, 71.98]	66.81	13.95	[62.55, 71.28]	0.19	.85	[–5.87, 7.12]	.05
D value	68.74	15.32	[64.05, 73.70]	67.97	16.48	[62.68, 73.21]	0.20	.84	[–6.68, 8.21]	.05
LSA	0.20	0.05	[.18, .22]	0.21	0.07	[.18,.23]	–.42	.68	[–.03, .02]	.10
Syntactic complexity										
Words/t-unit	22.27	4.19	[20.97, 23.69]	19.56	3.96	[18.32, 20.85]	2.84	.006	[.81, 4.61]	.67
Clause/t-unit	2.12	0.42	[1.98, 2.27]	1.98	0.38	[1.87, 2.11]	1.51	.14	[–.04, .33]	.36
Modifiers per NP	0.91	0.13	[.86, .95]	0.85	0.14	[.81, .90]	1.69	.09	[–.01, .12]	.40
Structural Similarity	0.07	0.02	[.07, .08]	0.08	0.02	[.07, .09]	–1.34	.18	[–.02, .003]	.32

Notes. MTLD = index of textual lexical diversity; LSA = latent semantic analysis; D value = measure of lexical variability based on Malvern and Richards (1997).

Table 7 Significant correlations (Pearson *r*) between writing behavior and linguistic complexity measures

Writing behavior	Linguistic complexity	*r*	95% CI	*p*
Simple				
Pause length between clauses (log)	Structural similarity	.46	[.15, .69]	.010
Complex				
Pause length between sentences (log)	Off-list words	–.47	[–.67, –.22]	.003
Revisions clause level and above	Academic words	–.50	[–.74, –.22]	.005