*ets research institute

JULY
2026
RR–26-06

TOEFL® RESEARCH SERIES TOEFL–RR-109

Second Language Writing Processes in the TOEFL iBT® Test: Examining the Write for an Academic Discussion Task Using Eye-Tracking, Keystroke Logging, and Stimulated Recall

AUTHORS

Ching-Ni Hsieh and Renka Ohta

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

TOEFL® Research Series

The TOEFL® test is the world's most widely respected English language assessment, primarily used for admissions purposes in more than 130 countries. Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, recognizing that rigorous research is essential to support a forward-looking assessment program and to substantiate claims about what test takers know or can do based on their test scores.

In addition to TOEFL-related research projects conducted by ETS staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language learning and testing experts from the academic community, funds an annual program of research carried out by external researchers around the world. To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books, in addition to the more than 300 peer-reviewed reports about TOEFL research published by ETS.

To obtain more information about the TOEFL programs and services, visit www.ets.org/toefl.

ETS Research Report Series ISSN 2330-8516

Second Language Writing Processes in the TOEFL iBT® Test: Examining the Write for an Academic Discussion Task Using Eye Tracking, Keystroke Logging, and Stimulated Recall

Ching-Ni Hsieh & Renka Ohta
ETS Research Institute, ETS, Princeton, New Jersey, United States

Abstract

The present study investigates the construct validity of the TOEFL iBT® writing test and focuses on the writing processes elicited by the Write for an Academic Discussion (WAD) task. Nineteen adult second-language (L2) English users completed the WAD task while their eye movements and keystrokes were recorded, followed by stimulated recall interviews to capture their writing strategies and thought processes. Eye-tracking data revealed sustained attention to the writing window, with frequent reference to the task question and example posts. Keystroke logging indicated substantial initial pauses and language-related revision behaviors. Qualitative analysis of stimulated recalls identified five core cognitive processes: prompt (re)reading, planning, formulation, monitoring, and revision. The integration of eye tracking, keystroke logging, and stimulated recall demonstrates that the WAD task engages L2 writers in processes consistent with theoretical models of academic writing, supplying backing for the construct validity of the TOEFL iBT Writing test. The findings also reveal that task design features may influence L2 writing processes and shape writers' strategic behaviors and attention allocation.

Keywords: writing processes, eye tracking, keystroke logging, stimulated recall, writing assessment, TOEFL iBT®, write for an academic discussion

Corresponding author: Ching-Ni Hsieh, E-mail: chsieh@ets.org

Acknowledgments

A generative AI tool was used to edit the texts to improve readability.

Introduction

According to the *Standards for Educational and Psychological Testing* (AERA et al., 2014), test developers are responsible for determining whether modifications to a test affect the interpretation of its scores. ETS Standards for Quality and Fairness (ETS, 2014) also states that empirical evidence needs to be gathered to demonstrate the technical quality, utility, and fairness for major revisions of existing ETS products and services. In 2023, the TOEFL iBT® test underwent several changes in its content, structure, and delivery in order to improve the overall test-taking experience, reduce administration time, and modernize the test content. A key change in the modernized TOEFL iBT test content was the replacement of the Independent Writing (IND) task with the Write for an Academic Discussion (WAD) task. Similar to the retired IND task, the WAD task measures test takers' ability to give and support an opinion in academic writing and is expected to engage test takers in critical academic writing processes including planning, formulating, monitoring, and revising (Cumming et al., 2021; Kellogg, 1996).

Unlike the IND task, which presents a general statement or question, the WAD task simulates an online class discussion. Test takers read a professor's question and two short student responses then write their own post to contribute to the discussion within 10 minutes. Prior research shows that task requirements (e.g., audience, purpose) and design features (e.g., time limits, prompt format) can influence the processes and strategies second language (L2) writers employ (Barkaoui, 2016; Naismith et al., 2024; Yu et al., 2017). The WAD task's innovative design and authentic context make it well suited for investigating academic writing processes that mirror real-world online discussions.

This study aimed to examine the validity evidence of the TOEFL iBT Writing test and focused on evidence related to construct validity. A central validity claim about the TOEFL iBT test is that test scores reflect the theoretical construct of academic writing proficiency (Chapelle et al., 2008; Cumming et al., 2021). This claim assumes that the response processes required to complete the test tasks align with the theoretical definition of the test construct. Backing for this claim is essential if the test construct is to serve as the explanation for the meaning of the TOEFL iBT test scores. As Chapelle et al. (2008) noted, such backing can be obtained through cognitive processing studies and by examining the relationship between task characteristics and the processes test takers employ. Evidence showing that the WAD task engages test takers in construct-relevant writing processes would support construct representation of the TOEFL iBT

Writing test and confirm that test scores are influenced by relevant psycholinguistic knowledge, processes, and strategies.

To address the study aims, we examined the writing processes L2 writers engage in when they respond to the WAD task through a combined use of eye tracking, keystroke logging, and stimulated recalls. The findings will help determine whether L2 writers employ the underlying academic writing ability the test is intended to measure, thereby providing empirical evidence for construct validity and score interpretation. Additionally, the results will offer implications for assessment task design and contribute to the triangulated use of eye tracking, keystroke logging, and verbal reports in test validation research.

Literature Review

Second Language Writing Processes

Writing in an L2 is a cognitively demanding activity that requires the coordination of multiple processes, including planning, formulating, monitoring, and revising (Cumming et al., 2021; Révész et al., 2019). For L2 writers, these demands are further intensified by the need to manage the interplay of cognitive, linguistic, and cultural expectations.

Exploring writing processes offers insights into the real-time cognitive activities and behavioral strategies that L2 writers employ during composition (Chukharev-Hudilainen et al., 2019; Révész et al., 2022). Unlike product-based writing research, which focuses primarily on final written texts (e.g., Ortega, 2015; Yoon & Polio, 2017), process-oriented studies (a) describe observable writing behaviors such as pausing and revision (Barkaoui, 2019; Révész et al., 2017; Wengelin et al., 2024) and (b) explore cognitive activities such as idea generation, attention allocation, and linguistic decision-making (Michel et al., 2020).

Information about writing processes helps researchers and test developers determine whether an assessment task activates the writing behaviors and mental processes considered essential in theoretical models of L2 writing proficiency (Cohen, 2012; Cumming et al., 2021). It also clarifies the extent to which a writing assessment task engages test takers in the same cognitive processes and subprocesses (e.g., planning content, retrieving lexis) involved in real-life writing contexts (Cohen, 2012; Shaw & Weir, 2007). Understanding how L2 writers navigate composition processes further informs models of L2 writing development, enhances instructional practices, and supports the interpretation and use of writing test scores.

Theoretical Framework of the Study

This study adopted Kellogg's (1996) cognitive model of writing as its guiding theoretical framework. Similar to most writing models (e.g., Bereiter & Scardamalia, 1987; Flower & Hayes, 1981; Hayes, 2012), Kellogg's model was originally developed to explain first language (L1) writing processes. We selected this model because, unlike other theoretical models of writing (e.g., Hayes, 2012), it places stronger emphasis on linguistic encoding—the processes involved in converting intended ideas into written text. These processes are particularly demanding for L2 writers, as tasks such as retrieving vocabulary, constructing syntax, and ensuring cohesion require greater effort and are less automatic compared to L1 writers, whose linguistic encoding skills are more fully automatized (Kormos, 2012; Roca de Larios et al., 2008). Furthermore, Kellogg's model has been successfully applied in empirical studies investigating L2 writing processes in TOEFL iBT independent and integrated writing tasks (e.g., Michel et al., 2020; Révész et al., 2022), making it well suited for the present study.

Kellogg's (1996) model identifies three key components of writing processes: (a) formulation, (b) execution, and (c) monitoring. Formulation encompasses higher order processes involved in generating ideas and planning content and lower order processes of translating planned ideas into written text such as the consideration of vocabulary, grammar, and cohesion. Execution refers to the processes of using motor skills related to handwriting or typing for transcription. Monitoring comprises the processes of reading and editing, where writers check for clarity and correctness and assess whether the text meets the intended purpose. If a discrepancy arises between the text and the writer's intention, editing may be initiated, which can involve local edits (e.g., word choice) or global edits (e.g., paragraph organization). Kellogg's model emphasizes the recursive nature of these processes and highlights the importance of managing cognitive load, as each component requires mental resources that can impact writing performance.

Research Methodologies in L2 Writing Processes

Empirical research in L2 writing processes has primarily employed three research methodologies: (a) verbal reports (e.g., think-aloud protocols, stimulated recalls), (b) eye tracking, and (c) keystroke logging. Each method has its own merits and limitations, and with advances in technology, L2 writing researchers increasingly adopt multimethod approaches to triangulate data from multiple sources and examine digital writing behaviors alongside

corresponding cognitive activities (e.g., Chukharev-Hudilainen et al., 2019; Michel et al., 2020; Révész et al., 2017, 2022). The following sections briefly review each methodology.

Verbal Reports

Early process-oriented studies in L2 writing have relied heavily on verbal report techniques such as think-aloud (e.g., Plakans, 2008; Yu et al., 2011) and stimulated recall procedures (e.g., Barkaoui, 2015; Sasaki, 2000). For example, Plakans (2008) asked 10 L2 writers to complete one integrated and one independent academic writing task from a university English placement exam while verbalizing their thought processes. Qualitative analysis of the think-aloud data revealed notable differences in planning behaviors. Participants devoted more effort to prewriting planning for the independent task because writers had to develop their own ideas before composing. In contrast, online planning (planning during writing) was more frequent for the integrated task because the source materials gave writers some information and ideas for writing.

In light of the results of Plakans (2008), in our study, we expected that L2 writers would engage more in online planning, generating and organizing ideas while composing, than exhibit extensive prewriting planning when responding to the WAD task that provides example posts as content support.

In another study utilizing verbal reports, Barkaoui (2015) used stimulated recalls to examine the writing processes of 22 L2 learners responding to the TOEFL iBT independent and integrated writing tasks. Findings indicated that participants engaged in a wide range of construct-relevant activities—such as planning, generating, evaluating, and revising—across both tasks. Interestingly, while task type significantly influenced writing behaviors and cognitive activities, differences between high- and low-proficiency learners were less pronounced. We expected to observe similar patterns in our study, using stimulated recalls and eye tracking. Compared to the sole use of stimulated recalls, eye-tracking data add an objective, real-time measure of writers' attention and cognitive effort to verbal reports, which may suffer from writers' memory biases or an inability to articulate all thought processes.

Eye Tracking

Recent L2 writing research has increasingly employed eye-tracking technology to investigate writing behaviors and cognitive processes during task execution (Chukharev-

Hudilainen et al., 2019; Hu & Aryadoust, 2024; Révész et al., 2022). This process-tracing technique records writers' moment-to-moment eye gaze and fixations in real time. The technique assumes a connection (i.e., the eye–mind link) between what an individual is thinking and how one's eyes move (Reichle, 2006). The focus, sequence, and length of eye fixations are thought to reflect writers' attentional allocation in processing visual information or written texts.

Several studies have utilized eye tracking to examine cognitive processes involved in standardized writing assessment tasks (e.g., Lu et al., 2025; Wang, 2022; Yu et al., 2017). Wang (2022) analyzed how 16 Chinese college students engaged with source materials while completing an integrated reading-to-write task from the Test of Business English Majors-Band 8. Participants spent, on average, 26.4% (6.9 minutes) reading task instructions and source materials and 73.6% (27.0 minutes) writing. Eye-tracking data revealed frequent shifts of attention among source materials as participants searched for lexical support and ideas during composition.

Yu et al. (2017) investigated cognitive processes elicited by IELTS Academic Writing Task 1, which required interpreting graphical data (e.g., line graphs, pie charts) and producing a response in 20 minutes. Findings showed substantial individual differences in eye-movement patterns and navigation across task instructions, graphs, and the text box. On average, participants spent 10% of task time reading instructions, 20% viewing graphs, and 70% focusing on the text box.

Given these findings, we predicted that L2 writers would repeatedly shift their gaze between the writing window and the prompt materials throughout the composing process to integrate prompt information and monitor for alignment with task requirements when responding to the WAD task.

Keystroke Logging

Keystroke logging typically consists of a logging system that is activated on a computer and allows researchers to record every keystroke and mouse click or movement during text production without interrupting the writing process (Chukharev-Hudilainen et al., 2019). The resulting logging data are time-stamped and provide fine-grained information that enables in-depth analyses of various writing behaviors such as pausing and revision.

Pausing behaviors, calculated from interkey intervals in keystroke logging, are considered indicators of cognitive activities (Barkaoui, 2019). Analyzing pausing patterns helps

researchers understand how writers allocate time to processes such as planning, monitoring, or evaluating, which are otherwise invisible in the final written text. Revision behaviors, typically operationalized as deletions and insertions in keystroke logs, represent another key cognitive component in writing models (Kellogg, 1996). Information regarding deletions and insertions in keystroke logs allows researchers to distinguish revisions made at different levels (e.g., content versus language) and text positions (e.g., at the point of inscription versus after careful rereading of previously produced text) and to estimate the time and efforts devoted to revision (Barkaoui, 2019; Chukharev-Hudilainen et al., 2019).

Many researchers have used keystroke logging to reconstruct and analyze writing processes in terms of time and cognitive effort (Leijten et al., 2019; Wengelin, 2006). Applications include examining L2 writers' revision behaviors (Barkaoui, 2016) and writing fluency (Barkaoui, 2019; Xu & Qi, 2017) as well as the effects of language proficiency (Barkaoui, 2016), task type (Barkaoui, 2019; Michel et al., 2020), and task complexity (Kim, 2022; Révész et al., 2017) on L2 writing processes. For example, Barkaoui (2019) investigated 68 L2 writers' pausing behaviors when they responded to the TOEFL iBT independent and integrated writing tasks. Analysis of keystroke logging data revealed that participants paused longer at higher level transition points (e.g., between paragraphs) and shorter at lower order transition points (e.g., between words). Moreover, participants paused longer, on average, with the integrated task, especially during the early stage of writing, and exhibited a higher proportion of revision pauses with the independent task, indicating significant task effects on the writing processes.

Based on these findings, we expected that L2 writers would exhibit distinct pausing patterns throughout the process of responding to the WAD task. Specifically, we anticipated the longest pauses to occur at the beginning of the task when writers are processing the prompt materials. During text production, we anticipated minimal in-word pauses because typing fluency should not be a bottleneck in the TOEFL iBT Writing test (Gong et al., 2022). We also expected to see short between-word pauses as writers engage in cognitive activities such as lexical retrieval and syntactic encoding.

Combined Use of Methodologies

Eye tracking and keystroke logging provide rich data on writers' eye-gaze patterns and keyboarding behaviors, information that can aid our understanding of an individual's attentional

processing during language assessment tasks (Godfroid et al., 2020). Nonetheless, L2 writing researchers have noted challenges in connecting these behavioral measures to the underlying cognitive operations of L2 writers. To address this limitation, several studies adopted mixed-method approaches, combining fine-grained eye tracking and keystroke data with verbal reports to triangulate findings (e.g., Chukharev-Hudilainen et al., 2019; Michel et al., 2020). Through L2 writers' conscious verbalization, researchers can interpret the cognitive processes reflected in eye movements and keystroke logs and identify what writers attend to, the problems they encounter, and the strategies they employ during composition.

For example, Michel et al. (2020) investigated Chinese college students' writing behaviors and cognitive processes when completing two TOEFL iBT independent and integrated writing tasks. Analyses of eye movements, keystroke logs, and stimulated recalls revealed similar writing behaviors and cognitive processes across tasks, except for source use in the integrated task. Additionally, Révész et al. (2022) found that high-proficiency L2 writers engaged in more varied activities across different stages of composition—such as more planning in the beginning stages and longer pauses for monitoring toward the end—compared to low-proficiency writers when participants completed two TOEFL iBT independent writing tasks.

Effects of Task Design

L2 writing research consistently shows that task type influences writers' strategic behaviors and cognitive processes (Barkaoui, 2016; Plakans, 2008). Consequently, empirical studies have examined writing processes across diverse task types, including argumentative essays (e.g., Gánem-Gutiérrez & Gilmore, 2018; Ranallia et al., 2018; Roca de Larios et al., 2008; Sasaki, 2000), emails (Conijn et al., 2019), summaries (e.g., Conijn et al., 2019), integrated writing tasks (e.g., Plakans, 2008; Wang, 2022), and online discussion board tasks (Oh, 2022). Overall, findings suggest that L2 writers engage in major composing processes—planning, formulating, evaluating, and revision—outlined in cognitive writing models (e.g., Kellogg, 1996). However, these processes are dynamic, recursive, and vary across individuals in complex, dynamic, and nonlinear ways (Plakans, 2008).

Although the potential impact of task design features on L2 writing processes is well recognized (Rong & Révész, 2025), systematic examination of the impact of specific task variables on L2 writing processes remains limited (e.g., Jung, 2020; Révész et al., 2017; Rong & Révész, 2025; Yu et al., 2011; Yu et al., 2017). For example, Révész et al. (2017) investigated the

effect of task complexity, operationalized as the absence versus presence of content support, on 73 advanced L2 writers' writing behaviors and underlying cognitive activities during argumentative essay writing. Using keystroke logging, the study found that content support led to fewer pauses, more revisions, and texts with greater lexical and syntactic complexity, suggesting that content support reduces planning burden and facilitates attention to linguistic encoding.

Yu et al. (2017) investigated the cognitive processes of 27 test takers completing four IELTS Academic Writing tasks that incorporated different types of graphs (e.g., bar, line, pie chart) in the prompts. Using eye tracking, the study examined how graph features influenced participants' eye movements and cognitive processing. Results indicated that both the features of a graph and the amount of information contained in the graph significantly impacted eye-gaze patterns and writing processes.

Given that academic writing is cognitively demanding, the amount of time allotted for composing may shape how writers engage with task components (Naismith et al., 2024). For example, Sasaki (2000) compared the writing processes of more skilled and less skilled Japanese EFL college students completing argumentative essays under a 30-minute time limit with flexibility to extend up to 1 hour. Findings showed that more skilled writers devoted significantly more time to planning before starting to write (M = 7.5 minutes; 18.3% of total writing time) and produced detailed organizational plans, whereas less skilled writers spent less time planning (M = 1.7 minutes; 4.6% of total writing time) and created less detailed outlines. In another study, Barkaoui (2016) investigated L2 writers' revision behaviors in one TOEFL iBT independent (30 minutes) and one integrated (20 minutes) writing task. Findings show that, under time constraints, participants prioritized language-related revisions over content revisions. Similar findings were shown in Révész et al. (2019) that explored revision behaviors involved in IELTS Academic Writing Task 2 and found that participants also made overall more language-focused than content-focused revisions. In this study, we expected that L2 writers would make both language-related and content-related revisions.

Research Questions

Insights into L2 writing processes can help test developers determine whether assessment tasks engage the cognitive mechanisms considered essential in theories of L2 knowledge and writing performance (Cumming et al., 2021) and the extent to which these tasks elicit processes

similar to those used in real-world academic writing contexts (Barkaoui, 2015; Cumming et al., 2021). The following research questions guided this study:

1. What writing processes do L2 writers engage in when completing the TOEFL iBT Write for an Academic Discussion (WAD) task?
2. How do these processes compare to those elicited by other writing tasks as reported in the literature?

Drawing on previous research findings, we expected L2 writers to engage in core cognitive processes—planning, formulating, monitoring, and revising—identified as fundamental to academic writing performance and the construct measured by the TOEFL iBT Writing test. If the WAD task fails to elicit these processes, the validity claim underlying score interpretation is compromised.

Method

Participants

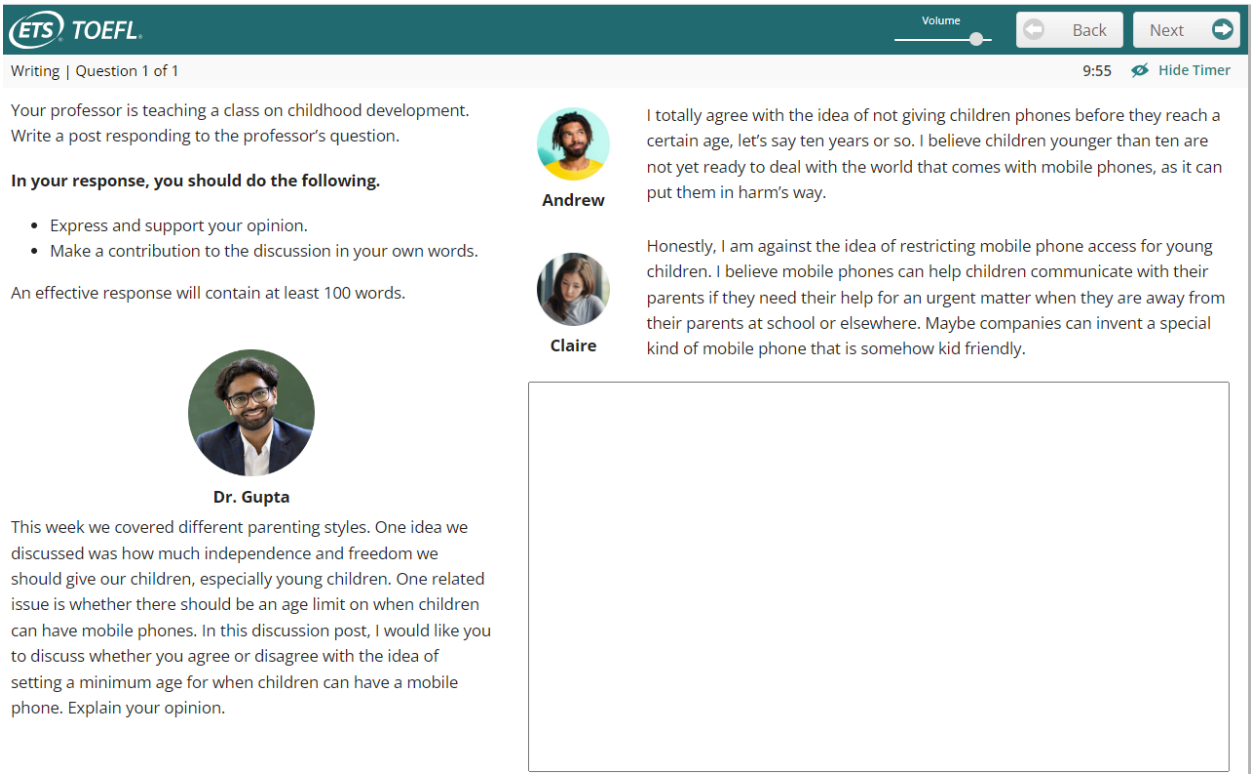
Participants included 19 adult L2 users of English, recruited through convenience sampling. Thirteen were female and six were male with ages between 18 and 45. Their L1s included Arabic (n = 2), Bengali (n = 1), Chinese (n = 5), Korean (n = 1), Hindi (n = 1), Japanese (n = 7), Romanian (n = 1), and Ukrainian (n = 1). On average, they had learned English as a foreign language for 13.74 years (SD = 8.64). Each participant received a gift card valued at $100 for their participation.

Write for an Academic Discussion Task

The WAD task is situated within an online discussion forum for a college class. In this forum, two students have posted opposing views on a given topic, each supported by their own arguments. Test takers are asked by the professor of the class to provide their opinion on the topic and contribute to the discussion. Test takers are required to write a 100-word essay in 10 minutes. Two prompts were used in this study and counterbalanced among the participants. The first prompt asked participants to write about the best way to determine whether online information is accurate and truthful. The second prompt asked participants to discuss whether they agree or disagree with the idea of setting a minimum age when children can have a mobile phone. An online platform designed to mirror the operational TOEFL iBT test delivery platform

was developed by the ETS Research Technology Group to administer the WAD task during eye-tracking experiments (see the Data Collection section). As shown in Figure 1, the task instruction and the professor's question were presented on the left-hand side of the screen, and the two students' example posts and the writing window were presented on the right part of the screen, similar to the operational version of the task.

Figure 1. Screenshot of Write for an Academic Discussion Task (Prompt 2)



Source: TOEFL iBT, screenshot

Data Collection

All participants took part in an individual session with one of the authors who served as the moderator at the eye-tracking lab on ETS campus in Princeton, New Jersey. After providing informed consent, participants were seated about 65 centimeters from the center of a ThinkVision computer monitor set to a screen resolution of 1920 X 1080 pixels with a refresh rate of 60 Hz. An eye tracker, a mobile Tobii Pro X3-120, was mounted on a standard tripod below the monitor screen. The eye tracker uses a combination of dark and bright pupil tracking techniques to detect eye movements. It collects data at a sampling rate of 120 Hz, or 120 data points (or gaze positions), per second (this frequency is high and allows for a relatively precise

estimation of fixation duration). To optimize the setup, the moderator made sure that participants were facing the front of the eye tracker and the monitor and could comfortably read the texts on the screen and reach the keyboard and mouse while seated at the recommended distance from the screen. After participants were properly positioned and seated, they were taken through a calibration procedure, which involved participants looking at a moving target displayed on the screen (a standard 9-point calibration pattern was used). Participants were instructed not to move their heads too much throughout calibration and during the writing task to ensure the accuracy of the data collected.

Following successful calibration, participants completed the WAD task on the computer while their on-screen typing behaviors were recorded and eye movements were tracked by the Tobii eye tracker. Since each part of the task was fixed on the screen, no scrolling was required, which allowed the eye tracker to capture eye movement data within each individual area on the screen. The moderator used a separate laptop computer, which was attached to the eye tracker and whose screen showed participants' live eye movements, to monitor participants' composing processes and deal with any technical issues that arose during the experiments.

Once participants completed the writing task, they took a short break and then participated in a stimulated recall session with the moderator to reflect on the thought processes they engaged in when carrying out the WAD task. The recall was prompted by a video recording of the screen during task performance with participants' eye movements overlaid. The moderator paused the video at different moments, such as during long pauses, rereading of prompt, or editing, and prompted participants to articulate their thoughts at those times. Participants were also encouraged to pause the recording at any point they wished to share the thoughts they had during the writing task. The interviews with the Japanese and Chinese speakers were conducted in Japanese and Mandarin Chinese, respectively. The remaining interviews were conducted in English. Most participants were comfortable speaking English; care was taken to ensure that participants had ample time for reflection and responding. The stimulated recalls were audio-recorded and translated and/or transcribed by the researchers for subsequent coding exercises.

ETS's keystroke logging pipeline engine was integrated into the online task delivery platform. The engine captures every keystroke event, cursor location, and corresponding time stamp in milliseconds. The collected log files were extracted from the system and processed for analysis.

Essay Scoring

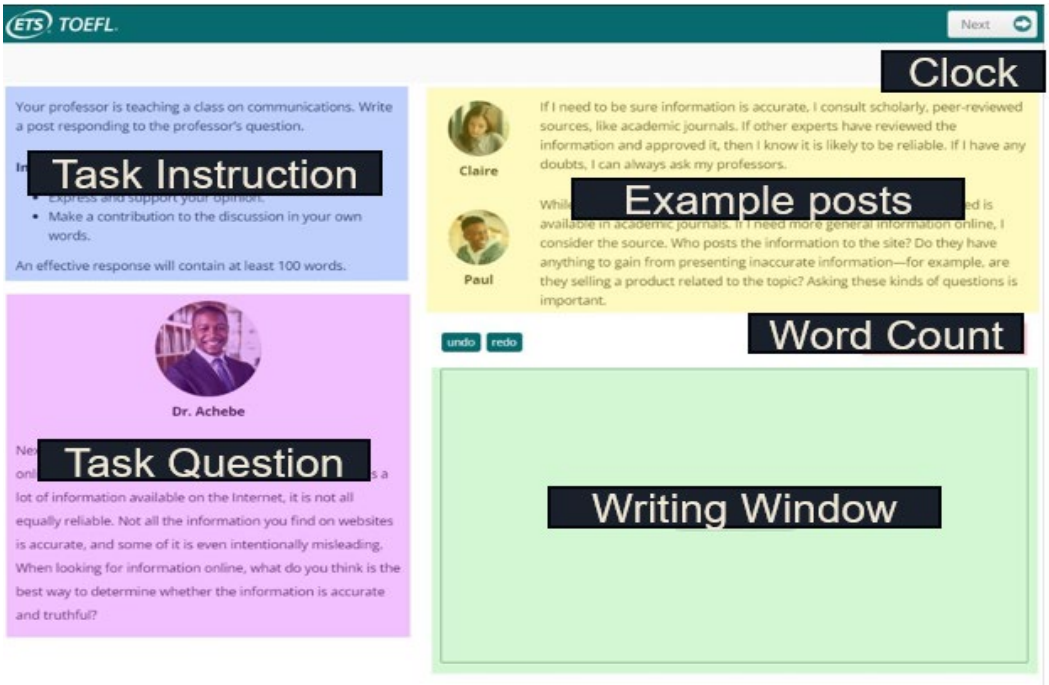
Each written response was scored by one experienced TOEFL iBT essay rater using the WAD task scoring rubric (see Appendix A). These raters have to go through rigorous training and calibration sessions to become certified. The TOEFL® program regularly evaluates raters' performance to ensure score consistency and reliability (ETS, 2025). Those who participated in the study were recruited and recommended by the operational scoring team; they all met certification requirements and regularly scored TOEFL iBT written responses at the time of data collection. The holistic 5-point scoring rubric the raters used included rating criteria of language use, relevance and elaboration of content, and contribution to the online discussion. The mean essay score was 3.53 (*SD* = 0.77).

Data Analysis

Analysis of Eye-Tracking Data

Participants' eye movements were measured in designated areas of interest (AOIs) on the screen using the Tobii Pro Lab software. We identified six AOIs for the WAD task: task instruction, task question, example posts, writing window, clock, and word count (see Figure 2).

Figure 2. Areas of Interest (AOIs) Defined for the WAD Task



Source: TOEFL iBT, adapted screenshot

Using the output generated from the Tobii Pro Lab, we report eight commonly used eye-tracking metrics to examine participants' viewing behaviors while completing the WAD task. These metrics include (a) total fixation duration, (b) average fixation duration, (c) fixation count, (d) time to first fixation, (e) first fixation duration, (f) total visit duration, (g) average visit duration, and (h) visit count. Table 1 provides the descriptions and interpretations of these metrics.

Table 1. Definitions and Interpretations of Eye-Tracking Metrics Analyzed

Metric	Definition	Interpretation
Total fixation duration	The aggregated duration of all fixations in an AOI	Indicates how much attention is paid to an AOI overall
Average fixation duration	The mean duration of individual fixations within an AOI	Reflects the depth of cognitive processing per fixation
Fixation count	The number of fixations made in an AOI	Higher counts may suggest greater interest or difficulty in processing
Time to first fixation	The time period from entering an AOI until the first fixation is made	Measures how quickly attention is drawn to an AOI
First fixation duration	The duration of the first fixation on an AOI	Indicates initial interest or immediate cognitive engagement with an AOI
Total visit duration	The total time spent across all visits to an AOI, including all fixations and saccades within that AOI	Reflects overall attention or engagement with an AOI; broader than total fixation duration
Average visit duration	The mean duration across all visits to an AOI	Reflects engagement or complexity of an AOI
Visit count	The total number of visits to an AOI	Indicates how often the AOI is revisited, which may reflect interest, confusion, or importance

Note. AOI = area of interest. A visit refers to when a participant's gaze first enters an AOI until exiting that AOI.
Source: See Conklin et al. (2019) and Godfroid (2019) for more information.

Analysis of Keystroke Logging Data

We analyzed the keystroke log files with a focus on pausing and revision behaviors. Our decision was informed by previous research that reports relationships between text quality and writing processes with keystroke logging (Choi & Deane, 2021; Conijn et al., 2022; Xu & Qi, 2017) and the keystroke features made available by ETS's keystroke pipeline engine. For pausing patterns, we examined participants' initial pause, in-word pause, and between-word pause durations. The initial pause refers to the time from the start of the writing task until the

first key press and may involve the activities of reading the prompt, planning content, or generating ideas. The in-word and between-word pauses are measured by all interkey intervals of keystrokes within- and between-words. These pause durations are thought to be associated with cognitive activities of planning the linguistic units that follow and could reflect processes related to lexical retrieval, syntactic encoding, or translation (Vandermeulen et al., 2020; Wengelin, 2006). Such information offers insights into cognitive efforts L2 writers experience when responding to a writing task with longer pauses, potentially suggesting more complex processing or difficulty.

For revision, we focused on three keystroke features that may provide insights into writers' revision strategies and self-monitoring during text production (Conijn et al., 2019). These keystroke features, including the time spent inserting or deleting characters and backspacing, have been used as indicators of writing revision processes (Baaijen & Galbraith, 2018). Extended periods of time spent on these keystroke events have been linked to editing behaviors, where words or phrases are spontaneously replaced or removed when planning, text production, or revising occur simultaneously (Deane, 2014). Of note is that deletion removes the character immediately following the cursor's current position or selected texts whereas backspacing removes the character immediately preceding the cursor's current position. The keystroke features analyzed are presented in Table 2. All durations were measured in seconds.

Table 2. Descriptions of Keystroke Features Analyzed

Phenomenon	Keystroke feature	Description
Pausing	Initial pause	Total time spent before the first alphanumeric keystroke
	In-word pause	Mean in-word pause duration
	Between-word pause	Mean between-word pause duration
Revision	Insertion time	Total time spent inserting characters between words
	Deletion time	Total time spent deleting characters between words
	Backspacing time	Total time spent on continued backspace runs

Note. All keystroke features are measured in seconds.

Analysis of Stimulated Recall Data

We conducted qualitative analysis of the stimulated recalls to identify the cognitive processes and subprocesses that the participants reported. We started with a preliminary list of cognitive activities associated with writing processes according to previous models of writing (Hayes, 1996; Kellogg, 1996). We then read and reread the transcripts of the stimulated recalls to refine the list of cognitive processes based on patterns emerging from the data. The final coding

scheme focused on five main processes: (a) prompt (re)reading, (b) planning, (c) formulation, (d) monitoring, and (e) revision and their associated subprocesses (see Table 3). Of note is that the prompt materials in the WAD task include the task instruction, the professor's question, and the students' example posts.

We used the coding scheme to systematically analyze the stimulated recall comments. To establish coder agreement, we independently coded data from two randomly selected participants. The intercoder percentage agreement for category identification was 80%; any discrepancies were resolved through discussion until we reached 100% agreement. We then independently coded the remaining data into the five main categories. For each participant, we summed up the comments by main cognitive processes and subprocesses to form frequency counts. We calculated the percentage of the stimulated recall comments falling into each process for each participant with the total number of stimulated recall comments reported by the participant serving as the denominator. We report the resulting frequency and proportions of comments.

Table 3. Coding Scheme for Stimulated Recalls

Process	Subprocess
Prompt (re)reading	Understand professor's question and requirements
	Check content alignment with the example posts
	Borrow text(s) from example posts
Planning	Brainstorm idea
	Manage time allocation
	Produce texts while planning online
Formulation	Retrieve lexis from mental lexicon (e.g., word choice, spelling, synonyms)
	Construct sentences or structures
	Create cohesive and coherent texts
Monitoring	Review texts to identify linguistic errors
	Review texts to look for disjointed sentences or ideas
	Evaluate texts to determine if task requirements were met
Revision	Fix spelling errors
	Review texts to improve clarity or organization
	Edit language to improve accuracy

Results

To answer the research questions, we first analyzed the eye-tracking measurements and keystroke logging features to understand the observable writing behaviors and then summarized

the stimulated recall comments to explore the various cognitive processes and subprocesses the participants reported.

Eye-Tracking Results

We investigated participants' gaze behaviors through eight eye-tracking metrics. Table 4 presents the descriptive statistics of these metrics for each AOI. The data showed that the Writing Window (AOI 4) consistently received sustained and frequent attention based on the four duration-based metrics (Metrics 1, 2, 6, 7) and two frequency counts (Metrics 3 and 8). This result implied that the writing window was the most heavily engaged AOI. Unsurprisingly, Task Instruction (AOI 1) had the quickest time to first fixation (Metric 4), meaning that it was accessed early in the response process. Task Question (AOI 2) and Example Posts (AOI 3) received intermittent and meaningful attention, with average visit counts of 27.84 and 23.65 times respectively, signaling that participants referred to or revisited the prompt materials throughout the task potentially for reference or inspiration. Clock (AOI 5) and Word Count (AOI 6) were minimally attended, with consistently low fixation and visit frequency.

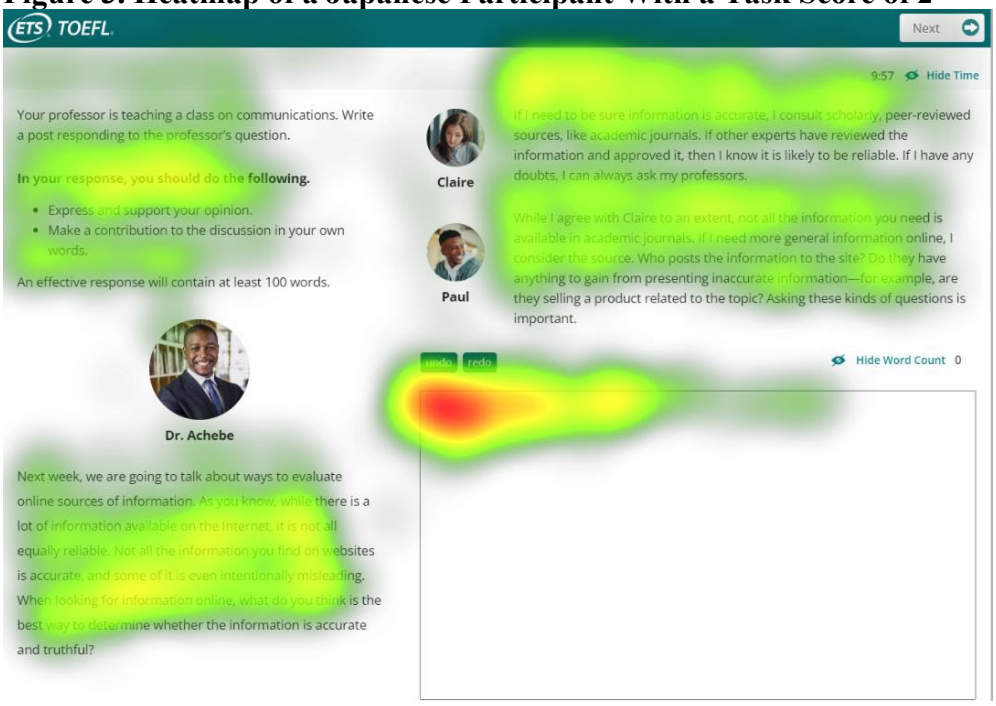
The eye-tracking software generates heatmaps that provide visual representations of eye fixations at different parts of the task (see example heatmaps in Figures 3 and 4). The colors in the eye-tracking heatmaps represent areas of high and low attention and engagement. The red and orange colors signal the hottest areas that receive the most attention and where participants spend the largest amount of time viewing. The yellow and green colors show areas that receive moderate attention where participants look but not as frequently nor for as much time as the red and orange areas. The blue or no color are the coldest areas where participants don't spend much time looking or completely overlook (Conklin & Pellicer-Sanchez, 2016). As can be seen in Figure 3, the participant, who received a task score of 2, focused heavily on reading the task instruction and example posts with limited attention spent on viewing the writing window. In contrast, Figure 4 shows a Korean speaker who exhibited a focused, sustained attention and engagement with the writing window, which aligned with a better writing performance (a task score of 5).

Table 4. Descriptive Statistics for Eye-Tracking Metrics by Area of Interest (AOI)

AOI	Total fixation duration		Average fixation duration		Fixation count		Time to first fixation		First fixation duration		Total visit duration		Average visit duration		Visit count	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
AOI 1	17.72	15.92	0.18	0.05	87.79	65.79	6.62	8.08	0.16	0.06	30.81	22.43	3.38	2.43	11.00	5.71
AOI 2	32.82	23.81	0.19	0.04	158.58	97.49	26.15	31.32	0.14	0.08	50.34	31.48	2.32	1.97	27.84	14.43
AOI 3	31.04	37.67	0.20	0.04	145.74	159.80	25.78	85.84	0.19	0.11	54.68	61.74	2.03	1.26	23.68	21.94
AOI 4	155.93	75.71	0.26	0.06	591.58	253.00	85.98	84.41	0.27	0.31	319.74	94.94	6.97	3.79	54.95	24.50
AOI 5	1.00	1.04	0.19	0.07	5.11	3.79	123.28	155.01	0.18	0.08	1.07	1.15	0.25	0.17	4.06	2.58
AOI 6	1.77	1.51	0.17	0.07	9.05	6.75	106.63	134.86	0.15	0.09	1.93	1.68	0.26	0.15	6.42	4.53

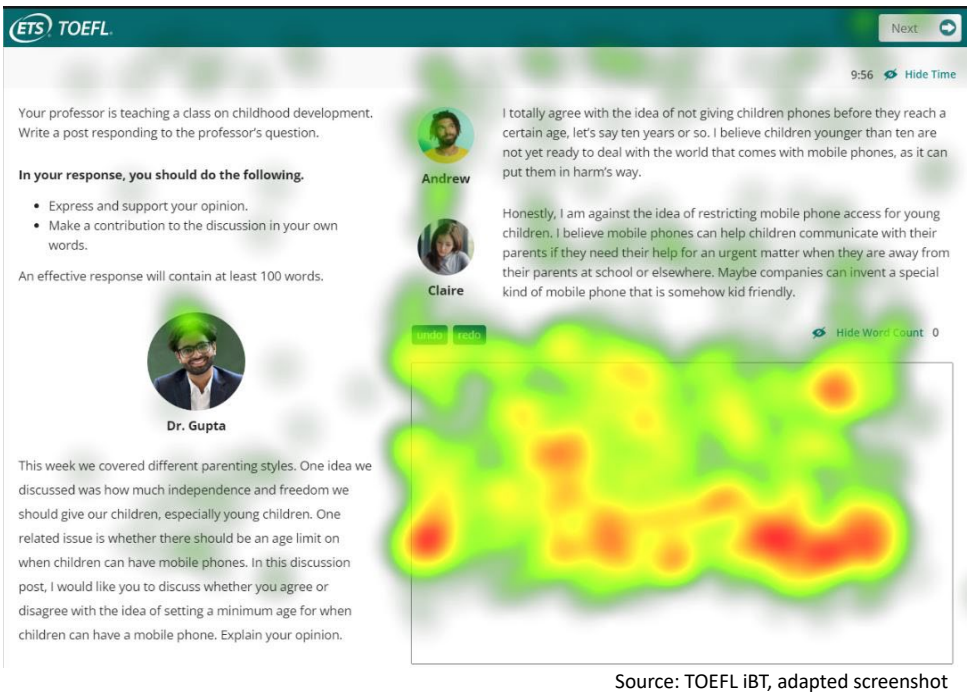
Note. AOI 1 = Task Instruction; AOI 2 = Task Question; AOI 3 = Example Posts; AOI 4 = Writing Window; AOI 5 = Clock; AOI 6 = Word Count

Figure 3. Heatmap of a Japanese Participant With a Task Score of 2



Source: TOEFL iBT, adapted screenshot

Figure 4. Heatmap of a Korean Participant with a Task Score of 5



Source: TOEFL iBT, adapted screenshot

Keystroke Logging Results

The keystroke logging data provide insights into participants' pausing and revision behaviors when they respond to the WAD task. As shown in Table 5, participants spent a

substantial period of time in initial pause (*M* = 129.54 seconds). Writers may have engaged in various cognitive activities during this time, such as reading the task instruction, the task question, or the example posts; planning; or formulating language before initiating text production. The wide variability in this measure (*SD* = 90.80) indicates huge individual differences in the writers' strategic behaviors—some writers may have quickly transitioned to writing while others took more time to process the task materials or plan their responses.

The mean duration of in-word pause was minimal, indicating that most participants typed with relative ease or may have reached automaticity in spelling and word formation. This finding aligns with previous research suggesting that shorter in-word pauses are indicative of lower cognitive load at lexical level (e.g., Gong et al., 2022; Wengelin, 2006). In contrast, between-word pauses were longer and more variable, potentially reflecting higher level cognitive processes such as syntactic planning or idea generation.

Turning to revision, we observed short average durations and large variabilities in the time spent on insertion, deletion, and backspacing. These variabilities reflect substantial individual differences in editing strategies. Deletion time was the longest among the three revision features, suggesting that deletion may involve more deliberate and cognitively demanding decisions, such as evaluating content relevance, clarity, or coherence. Unlike insertion or backspacing, deletion often requires writers to reconsider larger textual units. It is possible that participants' deletion behaviors reflect their engagement with more global revision strategies. As Table 5 shows, insertion and backspacing times were relatively brief, which may indicate more spontaneous or surface-level corrections that contrast with the more deliberate nature of deletions.

Table 5. Descriptive Statistics of Keystroke Features

Phenomenon	Keystroke feature	*M*	*SD*
Pausing	Initial pause	129.54	90.80
	In-word pause	0.18	0.058
	Between-word pause	0.60	0.319
Revision	Insertion time	20.55	12.98
	Deletion time	46.29	29.93
	Backspacing time	22.14	13.39

Note. All the keystroke features are measured in seconds.

Stimulated Recall Results

The stimulated recall comments offer insights into the cognitive processes that underlie

the writers' writing behaviors. The analysis of the stimulated recalls yielded a total of 91 comments associated with the five main cognitive processes that are reflected in theoretical models of writing processes (Kellogg, 1996) and found in empirical studies of L2 writing processes (Kim, 2022; Révész et al., 2017). Table 6 summarizes the comments from the 19 participants. We report both the frequency of mentions and the percentage relative to the total number of comments.

Table 6. Frequency and Percentages of Reported Cognitive Processes and Subprocesses

Process	Subprocess	*n*	%
Prompt (re)reading	Understand professor's question and requirements	14	15.4
	Check content alignment with the example posts	15	16.5
	Borrow texts from example posts	7	7.7
Planning	Brainstorm idea	5	5.5
	Manage time allocation	4	4.4
	Plan online	11	12.1
Formulation	Retrieve lexis from mental lexicon	8	8.8
	Construct sentences or structures	3	3.3
	Create cohesive and coherent texts	3	3.3
Monitoring	Review texts to identify linguistic errors	5	5.5
	Review texts to look for disjointed sentences or ideas	2	2.2
	Evaluate texts to determine if task requirements were met	3	3.3
Revision	Fix spelling errors	2	2.2
	Review texts to improve clarity or organization	5	5.5
	Edit language to improve accuracy	4	4.4

Note. $N = 91$. % = total percent of N.

Overall, participants most frequently commented on their prompt reading and reading behaviors, likely due to the repeated viewing observed during the task. Checking content relevance to the example posts (16.5%) and understanding the professor's question and requirements (15.4%) were the most frequently mentioned subprocesses overall. These results highlight the importance of ensuring that the ideas developed were aligned with the online discussion in participants' writing processes; additionally, the example posts appeared to serve not only as a source of task understanding but also as a linguistic and content scaffold.

Beyond the high frequency of prompt reading, we also noted strategic differences between higher scoring and lower scoring participants. When reflecting on how they approached the prompt, seven lower scoring writers (task scores 2 or 3) reported reading and rereading the prompt materials multiple times. Their reasons varied and included trying to understand the

professor's question, checking whether their own writing met the task requirements or aligned with the online discussion posts, or borrowing language from the example posts. Two comments illustrate this recurring pattern:

> I was looking at the professor's part here to borrow some words because I didn't know how to express my opinions in English. (ID 5, Japanese, Female, Task Score 2)

> I read the question two times here because I want to understand what they are saying. Like the first time, I just read. The second time, I read and my brain is trying to get some idea from what the teacher say and from what the students say. (ID10, Arabic, Male, Task Score 3)

In contrast, higher scoring writers (task scores of 4 or 5) described using more efficient strategies, such as skimming or scanning, when processing the prompt materials. For example, one participant explained:

> After I read the professor's question here, I moved on to the right-hand side to read the students' answers. I didn't read through all the details. I just got the gist of it. (ID8, Chinese, Female, Task Score 4)

Planning process was present in roughly one fifth of the comments (22.0%). The multiple mentions of *online planning* suggest a tendency to plan while writing, rather than engaging in comprehensive prewriting strategies. Many participants appeared to adopt a more fluid, on-the-go planning process, such as generating texts or organizing ideas during the actual writing process. For example, one said:

> I had a vague idea but while writing I was thinking about how I would say my opinion. (ID 2, Japanese, Female, Task Score 4).

Under time pressure, writers had to prioritize getting words on the screen quickly over fully developing their ideas beforehand. This adaptive behavior is likely strategic, particularly for more proficient writers, so that they could reserve more time for drafting and potentially revising. One participant commented:

> After I quickly read through the two sample responses, I just started writing. (Q: Did you plan ahead what you wanted to write first?) No, I just keep writing and

> thinking about what to write at the same time. I didn't come up with some concrete ideas first before I started typing. (Q: Why not?) Well, because this is a timed writing task, I don't have time to plan ahead. I need to start writing right from the beginning and then plan on what to write while you are also writing. (ID 13, Chinese, Female, Task Score 4)

Formulation process was modestly reported (15.4%). Participants referred to core linguistic production processes, including retrieving vocabulary, constructing sentences, and creating cohesion and coherence.

> I was looking at here (task question) to get some expressions. I didn't know how to say "I have never examined the information" so I was looking here to take the expressions. (ID 5, Japanese, Female, Task Score 2)

> Here I am thinking about how to put sentences together. (ID 7, Arabic, Male, Task Score 3)

The relatively low frequency of comments on sentence construction and discourse-level cohesion may reflect a few interacting factors. First, given the time constraint, participants appeared to strategically limit their syntactic and rhetorical sophistication in favor of more concise, idea-driven writing. Second, the nature of the WAD task may have shaped the kinds of formulation strategies required. Because the task is designed to simulate an online discussion forum rather than a formal argumentative essay such as the independent writing task, the discussion format might have influenced how participants conceptualized their language production. The WAD task format might allow or encourage more concise, direct, and conversational language over the more elaborate written discourse. As two noted:

> Once I made up my mind that I was going to talk about attention and distraction of cell phone, it was just a matter of putting words on the screen. That part was fairly easy and perhaps I should have thought about more arguments about not having a cell phone, but I thought that as long as I can emphasize just one bullet point here, that should be enough to carry me around 100 words or so. (ID 1, Hindi, Male, Task Score 4)

> I wanted to be able to put down all my ideas, like write them down, in ten

> minutes. You know, it's a short time limit.… I think this task is more about generating my ideas, instead of the structure or the format of academic writing. To me, 100 words, you can't really have a full academic writing anyway. Also the format is a discussion. In other words, it's a bit less formal, less like academic writing. I think it's more focused on my ideas. (ID 16, Chinese, Female, Task Score 4)

These two quotes reflect a shared perception that the task design features—particularly the 100-word guideline and the online discussion format—shape their process of linguistic and idea formulation. Rather than aiming for linguistic complexity or extended discourse structures, they focus on clarity, efficiency, and idea expression.

Only a small proportion of the comments involved monitoring (11.0%), which included reviewing linguistic errors, checking coherence of ideas, and evaluating task fulfilment. For example,

> I just double check to see if I have any like grammar errors. (ID 7, Arabic, Male, Task Score 3)

> After I finished writing, I quickly reviewed what I've written here. I just wanted to make sure that it reads smoothly and I expressed everything I wanted to say here. (ID 15, Chinese, Female, Task Score 4)

> I'm trying to evaluate the structure here and see where I can strengthen my discussion. I wanted to tie it back to the topic, make it become a more coherent discussion. (ID19, Chinese, Female, Task Score 4)

Participants made occasional comments related to revision (12.1%). Revisions were made in order to improve linguistic accuracy, clarity, or organization. Depending on the evaluation of the texts produced, participants commented that they generally concentrated on editing or addressing local linguistic problems while few focused on the broader content or organizational issues. For example,

> I found the spelling error, "the" was misspelled as "ten," so I fixed it, and the time was up. (ID 4, Japanese, Male, Task Score 3)

> I deleted the texts here because I didn't have concrete idea. (ID 11, Japanese,

Female, Task Score 2)

I wanted to make sure that the intro paragraph is clear, but I didn't notice it while I was typing it. Oh, here I don't like this, I change the wording. I don't like this. So let me change it. (ID 17, Romanian, Male, Task Score 4)

Overall, the stimulated recalls reflect a range of strategic revision behaviors, with many participants prioritizing quick, surface-level edits while few engaged in more substantive content revisions. These revision patterns suggest that participants varied in their ability, or willingness, to allocate their cognitive resources toward revision because they often had to balance the urgency of finishing the task within 10 minutes against their goals for improving either the linguistic form or content message.

Discussion

Construct Validity

We addressed the first research question by exploring the writing processes L2 writers engaged in when they responded to the WAD task through the analyses of three data sources. Overall, our findings suggest that the WAD task activates core academic writing processes—planning, formulation, monitoring, and revision—as outlined in Kellogg's (1996) cognitive model of writing. This alignment supports the construct validity of the modernized TOEFL iBT Writing test and confirms that test scores reflect the intended theoretical construct of academic writing proficiency.

It is worthwhile pointing out that our analysis revealed that prompt processing is a special case of planning in the context of the WAD Task. The stimulated recalls suggest that prompt engagement is not only recurrent but also mediates other cognitive processes, such as linguistic formulation or monitoring, especially in timed writing tasks with complex prompts. Future research should further investigate this dynamic and recursive process, examining how prompt processing is shaped by task characteristics and L2 writing proficiency.

Literature and Study Insights on L2 Writing Processes

The second research question focused on how the writing processes elicited by the WAD task compare to those observed in other writing tasks, as documented in the literature. Drawing on previous empirical studies of integrated writing tasks (e.g., Plakans, 2008), we anticipated that

L2 writers responding to the WAD task would engage more in online planning and allocate more of their planning time during the act of writing rather than before it. Consistent with Plakans (2008), our findings showed that many participants adapted their planning strategies to meet the specific requirements of the WAD task, often planning and organizing ideas while composing rather than engaging in extensive prewriting planning.

A notable result from our keystroke logging data was the substantial variability in initial pause durations among participants, indicating prominent individual differences in how L2 writers balance planning and immediate text production. The stimulated recall interviews further revealed that higher scoring writers tended to process the prompt efficiently and begin writing sooner, whereas lower scoring writers spent more time processing the prompt or searching for appropriate linguistic resources before starting to type. This pattern diverges from findings by Barkaoui (2019), who reported that more proficient L2 writers typically paused longer at the outset of independent and integrated TOEFL iBT writing tasks than less proficient writers. One possible explanation for this discrepancy is the influence of task characteristics, such as topics, source materials, and time constraints, which may place different demands on L2 writers' cognitive resources and attention allocation.

Informed by previous empirical research utilizing eye tracking to investigate L2 writing processes (e.g., Wang, 2022; Yu et al., 2017), we anticipated that L2 writers would exhibit recursive viewing patterns, frequently shifting their gaze between the writing window and various prompt materials, in response to the changing WAD task environment—a distance between the text produced so far and the objectives the writer sets for themselves. This expectation was confirmed by our eye-tracking data, which showed high visit counts and extended fixation durations for both the task question and example posts, indicating writers' strategic attention allocation throughout the writing process.

Based on theoretical work and empirical studies that use keystroke logging to reconstruct and analyze writing processes across different writing task types (Barkaoui, 2019; Chukharev-Hudilainen et al., 2019; Leijten et al., 2019; Wengelin, 2006), we expected that L2 writers would pause for different durations at different textual levels. In line with our predictions, we found that the longest pauses occurred at the beginning of the WAD task when participants were engaged in prompt reading or initial planning. As anticipated, in-word pause durations were minimal, indicating a high degree of typing fluency and automaticity in spelling and word formation

among participants. In contrast, we observed short but variable between-word pauses, likely corresponding to the cognitive activities of lexical retrieval, syntactic encoding, and linguistic formulation. These dynamic pausing patterns and stage-specific writing behaviors observed in this study corroborate those reported in the literature (e.g., Barkaoui, 2019; Plakans, 2008; Révész et al., 2022).

In examining revision behaviors, our analyses confirmed that L2 writers made both language-related and content-related revisions. Consistent with previous research, our data showed that L2 writers tended to focus on surface-level, language-focused revisions, as evidenced by the relatively brief durations spent on deletion, insertion, and backspacing during the WAD task. The stimulated recall data further revealed that participants made a variety of strategic revision choices, with a clear emphasis on lower level edits such as correcting linguistic accuracy and spelling, while only a minority attempted more global revisions related to organization or coherence. These findings align with previous research (Barkaoui, 2016; Révész et al., 2019), which demonstrated that L2 writers, particularly under timed academic writing conditions, tend to prioritize language-related revisions over content-related changes.

Theoretical and Practical Implications

Our findings demonstrate that the WAD task elicits a range of cognitive processes that align with theoretical models of academic writing (Cumming et al., 2021; Kellogg, 1996). The distinctive features of the WAD task, including its brief time limit, online discussion format, and multicomponent structure, place unique cognitive demands on L2 writers. These demands not only shape the nature of the writing processes engaged, but they also accentuate individual differences in attention allocation, strategic behaviors, and composing efficiency. Importantly, these results highlight the critical role of task design in shaping L2 writing processes, particularly within the context of timed, standardized writing assessments. Consideration of these factors is essential for both test developers and language educators seeking to interpret writing performance and support effective assessment practices.

Our findings highlight the importance of considering prompt processing as part of the planning process for L2 writers. The recurring and strategic engagement with prompts suggests that it mediates L2 writing processes and performance. From a practical perspective, our results point out the need for designing prompts that are both accessible and sufficiently challenging to elicit construct-relevant cognitive writing activity. We recommend that future research further

explore how task design, cognitive activity, writing proficiency, and writing processes interact, and that prompt processing demands be carefully considered in the development of writing assessment tasks.

Limitations

Several limitations should be acknowledged. First, the small sample size limits the generalizability of our findings to other contexts. Second, some participants were interviewed in their L2, which may have affected their ability to fully express their thoughts. Third, we recruited the participants through convenience sampling due to practical challenges in data collection. More representative samples with varying proficiency levels should be included in future research. Additionally, an external measure of language proficiency should be used to gauge the relationship between writing processes and language ability. Despite these limitations, the findings deepen our understanding of L2 writing processes involved in a timed, contextualized writing assessment task and illuminate how task design may influence writers' approaches to strategic behaviors and decision-making.

Conclusion

The combined use of eye tracking, keystroke logging, and stimulated recall offers a deep understanding of how L2 writers engage with the WAD task and enables meaningful comparisons with prior research. Findings of the study provide backing for the construct validity of the WAD task and supply new empirical evidence to inform score interpretation of the modernized TOEFL iBT Writing test. The WAD task is shown to elicit construct-relevant cognitive processes that are aligned with established theoretical models of writing (Cumming et al., 2021; Kellogg, 1996). The findings support Kellogg's (1996) cognitive model of writing in that writers need to balance attentional resources across formulation, execution, and monitoring. These findings deepen our understanding of L2 writers' strategic behaviors and cognitive processes when composing under time constraints. Ultimately, the insights gained from this study can guide future refinements of the WAD task and inform the design of similar writing assessment tasks that require test takers to process prompt materials and respond within standardized, timed conditions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Psychological Association.

Baaijen, V. M., & Galbraith, D. (2018). Discovery through writing: Relationships with writing processes and text quality. *Cognition and Instruction*, *36*(3), 199–223. https://doi.org/10.1080/07370008.2018.1456431

Barkaoui, K. (2015). *Test takers' writing activities during the TOEFL iBT® writing tasks: A stimulated recall study* (Research Report No. RR-15-04). ETS. https://doi.org/10.1002/ets2.12050

Barkaoui, K. (2016). What and when second-language learners revise when responding to timed writing tasks on the computer: The roles of task type, second language proficiency, and keyboarding skills. *The Modern Language Journal*, *100*(1), 320–340. https://doi.org/10.1111/modl.12316

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

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

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Choi, I., & Deane, P. (2021). Evaluating writing process features in an adult EFL writing assessment context: A keystroke logging study. *Language Assessment Quarterly*, *18*(2), 107–132. https://doi.org/10.1080/15434303.2020.1804913

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H.-H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, *41*(3), 583–604. https://doi.org/10.1017/S027226311900007X

Cohen, A. D. (2012). Test-taking strategies. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to assessment* (pp. 96–104). Cambridge University Press.

Conijn, R., Cook, C., van Zaanen, M., & Van Waes, L. (2022). Early prediction of writing quality using keystroke logging. *International Journal of Artificial Intelligence in Education*, *32*, 835–866. https://doi.org/10.1007/s40593-021-00268-w

Conijn, R., Roeser, J., & van Zaanen, M. (2019). Understanding the keystroke log: The effect of writing task on keystroke features. *Reading and Writing*, *32*, 2353–2374. https://doi.org/10.1007/s11145-019-09953-8

Conklin, K., & Pellicer-Sanchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research*, *32*(3), 435–467. https://doi.org/10.1177/0267658316637401

Conklin, K., Pellicer-Sanchez, A., & Carrol, G. (2019). *Eye-tracking: A guide for applied linguistics research*. Cambridge University Press. https://doi.org/10.1017/9781108233279

Cumming, A., Cho, Y., Burstein, J., Everson, P., & Kantor, R. (2021). Assessing academic writing. In X. Xi & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 107–151). Routledge. https://doi.org/10.4324/9781351142403-4

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (Research Report No. RR-14-03). ETS. https://doi.org/10.1002/ets2.12002

ETS. (2014). *ETS standards for quality and fairness*. ETS. https://www.ets.org/pdfs/about/standards-quality-fairness.pdf

ETS. (2025). *Reliability and comparability of TOEFL iBT® scores*. (TOEFL Research Insight Series, Vol. 3). ETS. https://www.ets.org/pdfs/toefl/toefl-ibt-insight-s1v3.pdf

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*(4), 365–387. https://doi.org/10.2307/356600

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

Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge. https://doi.org/10.4324/9781315775616

Godfroid, A., Winke, P., & Conklin, K. (2020). Exploring the depths of second language processing with eye tracking: An introduction. *Second Language Research*, *36*(3), 243–

255. https://doi.org/10.1177/0267658320922578

Gong, T., Zhang, M., & Li, C. (2022). Association of keyboarding fluency and writing performance in online-delivered assessment. *Assessing Writing*, *51*, 100575. https://doi.org/10.1016/j.asw.2021.100575

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 1–27). Routledge.

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

Hu, X., & Aryadoust, V. (2024). A systematic review of eye-tracking technology of second language research. *Languages*, *9*(4), 141. https://doi.org/10.3390/languages9040141

Jung, J. (2020). Effects of content support on integrated reading-writing task performance and incidental vocabulary learning. *System*, *96*, 102288. https://doi.org/10.1016/j.system.2020.102288

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

Kim, Y. (2022). Effects of task complexity on EFL learners' writing behaviors and performance. *English Teaching*, *77*(4), 49–69. https://doi.org/10.15858/engtea.77.4.202212.49

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

Leijten, M., Van Waes, L., Schrijver, I., Bernolet, S., & Vangehuchten, L. (2019). Mapping Master's students' use of external sources in source-based writing in L1 and L2. *Studies in Second Language Acquisition*, *41*(3), 555–582. https://doi.org/10.1017/S0272263119000251

Lu, X., Révész, A., Michel, M., Kourtali, N., Lee, M., & Borges, L. (2025). Investigating the relationships of writing behaviours to linguistic complexity and accuracy in independent and integrated writing task performance. *Applied Linguistics Review*, *16*(4), 1551–1579. https://doi.org/10.1515/applirev-2024-0004

Michel, M., Révész, A., Lu, X., Kourtali, N.-E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study.

Second Language Research, *36*(3), 307–334. https://doi.org/10.1177/0267658320915501

Naismith, B., Attali, Y., & LaFlair, G. T. (2024). The impact of task duration on the scoring of independent writing responses of adult L2-English writers. *Assessing Writing*, *62*, 100895. https://doi.org/10.1016/j.asw.2024.100895

Oh, S. (2022). The use of spelling and reference tools in second language writing: Their impact on students' writing performance and process. *Journal of Second Language Writing*, *57*, 100916. https://doi.org/10.1016/j.jslw.2022.100916

Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82–94. https://doi.org/10.1016/j.jslw.2015.06.008

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, *13*, 111–129. https://doi.org/10.1016/j.asw.2008.07.001

Ranallia, J., Feng, H.-H., & Chukharev-Hudilainena, E. (2018). Exploring the potential of process-tracing technologies to support assessment for learning of L2 writing. *Assessing Writing*, *36*, 77–89. https://doi.org/10.1016/j.asw.2018.03.007

Rayner, K., & Clifton, C. (2009). Language processing in reading and speech perception is fast and incremental: Implications for event-related potential research. *Biological Psychology*, *80*(1), 4–9. https://doi.org/10.1016/j.biopsycho.2008.05.002

Reichle, E. D. (2006). Computational models of eye-movement control during reading: Theories of the "eye-mind" link. *Cognitive Systems Research*, *7*(1), 2–3. https://doi.org/10.1016/j.cogsys.2005.07.001

Révész, A., Kourtali, N.-E., & Mazgutovab, D. (2017). Effects of task complexity on L2 writing behaviors and linguistic complexity. *Language Learning*, *67*(1), 208–241. https://doi.org/10.1111/lang.12205

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed methods study. *Studies in Second Language Acquisition*, *41*(3), 605–631. https://doi.org/10.1017/S027226311900024X

Révész, A., Michel, M., Lu, X., Kourtali, N., Lee, M., & Borges, L. (2022). The relationship of proficiency to speed fluency, pausing, and eye-gaze behaviours in L2 writing. *Journal of Second Language Writing*, *58*, 100927. https://doi.org/10.1016/j.jslw.2022.100927

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

Language Writing, *17*(1), 30–47. https://doi.org/10.1016/j.jslw.2007.08.005

Rong, X., & Révész, A. (2025). The effects of task complexity on second language writing processes and products in computer-assisted collaborative writing. *Journal of Second Language Writing*, *69,* 101216. https://doi.org/10.1016/j.jslw.2025.101216

Sasaki, M. (2000). Toward an empirical model of EFL writing processes: An exploratory study. *Journal of Second Language Writing*, *9*(3), 259–291. https://doi.org/10.1016/S1060-3743(00)00028-X

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

Vandermeulen, N., Leijten, M., & Van Waes, L. (2020). Reporting writing process feedback in the classroom using keystroke logging data to reflect on writing processes. *Journal of Writing Research*, *12*(1), 109–140. https://doi.org/10.17239/jowr-2020.12.01.05

Wang, P. (2022). Investigating EFL test-takers' engagement with source materials in reading-to-write: Evidence from eye-tracking. *Open Journal of Social Sciences*, *10*, 442–457. https://doi.org/10.4236/jss.2022.104032

Wengelin, Å. (2006). Examining pauses in writing: Theory, methods and empirical data. In K. Sullivan & E. Lindgren (Eds.), *Computer key-stroke logging and writing: Methods and applications (Studies in Writing)* (pp. 107–130). Elsevier. https://doi.org/10.1163/9780080460932_008

Wengelin, Å., Johansson, R., Frid, J., & Johansson, V. (2024). Capturing writers' typing while visually attending the emerging text: A methodological approach. *Reading and Writing*, *37*, 265–289. https://doi.org/10.1007/s11145-022-10397-w

Xu, C., & Qi, Y. (2017). Analyzing pauses in computer-assisted EFL writing—A computer-keystroke-log perspective. *Educational Technology & Society*, *20*(4), 24–34. https://www.jstor.org/stable/26229202

Yoon, H. J., & Polio, C. (2017). The linguistic development of students of English as a second language in two written genres. *TESOL Quarterly*, *51*(2), 275–301. https://doi.org/10.1002/tesq.296

Yu, G., He, L., & Isaacs, T. (2017). The cognitive processes of taking IELTS Academic Writing Task 1: An eye-tracking study. *IELTS Research Reports Online Series*, *2017/2*. British Council, IDP IELTS, and Cambridge English Language Assessment.

Yu, G., Rea-Dickins, P. M. & Kiely, R. (2011). The cognitive processes of taking IELTS Academic Writing Task 1, *IELTS Research Reports*, *11,* 373–449. British Council, IDP IELTS, and Cambridge English Language Assessment.

Appendix A. Write for an Academic Discussion Task Scoring Rubric

toefl ibt.

Writing Scoring Guide

Write for an Academic Discussion

SCORE	DESCRIPTION
5	**A fully successful response** The response is a relevant and very clearly expressed contribution to the online discussion, and it demonstrates consistent facility in the use of language. A typical response displays the following: • Relevant and well-elaborated explanations, exemplifications and/or details • Effective use of a variety of syntactic structures and precise, idiomatic word choice • Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there/their*)
4	**A generally successful response** The response is a relevant contribution to the online discussion, and facility in the use of language allows the writer's ideas to be easily understood. A typical response displays the following: • Relevant and adequately elaborated explanations, exemplifications and/or details • A variety of syntactic structures and appropriate word choice • Few lexical or grammatical errors
3	**A partially successful response** The response is a mostly relevant and mostly understandable contribution to the online discussion, and there is some facility in the use of language. A typical response displays the following: • Elaboration in which part of an explanation, example or detail may be missing, unclear or irrelevant • Some variety in syntactic structures and a range of vocabulary • Some noticeable lexical and grammatical errors in sentence structure, word form or use of idiomatic language
2	**A mostly unsuccessful response** The response reflects an attempt to contribute to the online discussion, but limitations in the use of language may make ideas hard to follow. A typical response displays the following: • Ideas that may be poorly elaborated or only partially relevant • A limited range of syntactic structures and vocabulary • An accumulation of errors in sentence structure, word forms or use
1	**An unsuccessful response** The response reflects an ineffective attempt to contribute to the online discussion, and limitations in the use of language may prevent the expression of ideas. A typical response displays the following: • Words and phrases that indicate an attempt to address the task, but with few or no coherent ideas • Severely limited range of syntactic structures and vocabulary • Serious and frequent errors in the use of language • Minimal original language; any coherent language is mostly borrowed from the stimulus
0	**The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt or consists of arbitrary keystrokes.**

Appendix B. Stimulated Recall Interview Protocol

Part 1. Participant's thought processes and strategy use.

[Show screenshot of final page with prompt and response to help participant recall]

1. Tell me the steps you took to respond to this writing task. (e.g., What did you do first, how did you decide what to do next, how did you organize your ideas, did you review or revise before your submitted)

[Play eye-tracking video, pause at points of interests based on observations]

2. I see that you stopped writing here (Show a long pause). Can you tell me what you were thinking at this point?
3. I see that you were looking at this (Point the area of interest on the screen). Can you tell me what you were thinking at this point?
4. I see that you changed (added, deleted, reordered, etc.) the texts here. Can you tell me what you were thinking at that point?

Part 2. Perceptions of time limit, task difficulty, and instructional clarity.

5. Did you have enough time to write?
6. How easy was it for you to think of what to write?
7. Was there anything about this task that was difficult or confusing to you?

Part 3. Perceptions of test construct and domain representation.

8. Does this writing task allow you to show how well you can write in English?
9. Is this writing task similar to what you would do in daily-life writing activities?

Part 4. Wrap Up

10. Do you have anything else that you would like to say about this writing task?

Suggested Citation

Hsieh, C.-N., & Ohta, R. (2026). *Second language writing processes in the TOEFL iBT® test: Examining the Write for an Academic Discussion task using eye-tracking, keystroke logging, and stimulated recall* (Research Report). ETS. https://doi.org/10.64634/tk1sd682

Action Editor: Zuowei Wang

Reviewers: Lin Gu and Saerhim Oh

Find other ETS-published reports by searching the ETS ReSEARCHER database.