



Research Memorandum

ETS RM-23-06

A Comparison of Two TOEFL® Writing Tasks

Larry Davis
John M. Norris

August 2023



ETS Research Memorandum Series

EIGNOR EXECUTIVE EDITOR

Daniel F. McCaffrey
Lord Chair in Measurement and Statistics

ASSOCIATE EDITORS

Usama Ali
Senior Measurement Scientist

Beata Beigman Klebanov
Principal Research Scientist

Heather Buzick
Senior Research Scientist

Tim Davey
Director Research

Larry Davis
Director Research

Jamie Mikeska
Senior Research Scientist

Gautam Puhan
Principal Psychometrician

Jonathan Schmidgall
Senior Research Scientist

Jesse Sparks
Senior Research Scientist

Michael Walker
Distinguished Presidential Appointee

Klaus Zechner
Senior Research Scientist

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

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 Memorandum 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 Memorandum 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.

A Comparison of Two TOEFL® Writing Tasks

Larry Davis

ETS, Princeton, New Jersey, United States

John M. Norris

ETS Japan, Tokyo, Japan

August 2023

Corresponding author: L. Davis, Email: ldavis@ets.org

Suggested citation: Davis, L., & Norris, J. M. (2023). *A comparison of two TOEFL® writing tasks* (Research Memorandum No. RM-23-06). ETS.

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

To obtain a copy of an ETS research report, please visit
<https://www.ets.org/research/researcher.html>

Action Editor: Beata Beigman Klebanov

Reviewers: Alexis Lopez and Rick Tannebaum

Copyright © 2023 by Educational Testing Service. All rights reserved.

E-RATER, ETS, the ETS logo, TEXTEVALUATOR, TOEFL, and TOEFL IBT are registered trademarks of Educational Testing Service (ETS). TOEFL ESSENTIALS is a trademark of ETS. All other trademarks are the property of their respective owners.

Abstract

One component of the 2023 update to the TOEFL iBT® test was the replacement of the Independent Writing (IND) task with the Writing for an Academic Discussion (WAD) task. Though both tasks assess academic English writing ability by eliciting extended writing performances, there are certain differences in their design. Most apparently, the WAD task allows for up to 10 minutes of writing time, compared with 30 minutes for the IND task, hence the amount of writing elicited may differ considerably between the two tasks. Nevertheless, both tasks have the same communicative goal (express and support an opinion) and are scored according to a similar rating scale, and the new task is intended to contribute to the calculation of a total writing section score in the same way as the previous task. To justify this update, we compared performances on the two tasks by a sample of test takers ($N = 242$) who completed both tasks on operational tests at similar points in time. We found that human scores on the two tasks distributed test takers into similarly broad ranges of ability and that scores on the two tasks were identical or within one score point 93% of the time. We also found that direct measures of writing performances using automated natural language processing tools revealed substantial similarities in the quality of texts produced by test takers on the two tasks, in terms of the syntactic complexity, grammatical accuracy, lexical variety, discourse cohesion and elaboration, and fluency of their writing. Only slight differences were identified in a few measures of lexical, fluency, and discourse dimensions of the tasks. Overall, findings provide initial support for using the new task to support interpretations about English writing ability. We discuss these findings in light of the validity of interpretations and uses of the TOEFL iBT test, and we highlight implications for the standardized assessment of academic English proficiency.

Keywords: second language writing, performance assessment, task design, construct validity, score comparisons, linguistic measures

Acknowledgments

We appreciate Shuhong Li's assistance in compiling this data set and in computing simulated writing section scores, as well as the suggestions for improvement by several reviewers.

Table of Contents

Assessing Second Language Academic Writing Ability	2
Writing Task Design and the Measurement of L2 Performance	5
Opinion Writing in the TOEFL iBT® and TOEFL® Essentials™ Tests	9
The Current Study	14
Research Questions	14
Methods.....	15
Participant Sample	15
Test Administration	19
Scoring and Linguistic Measures	21
Statistical Analyses	22
Results.....	23
RQ1: To what extent do both tasks produce similar scores?	23
RQ2: To what extent are the characteristics of written responses similar or different across task type?	29
Discussion	43
Limitations	46
Conclusion.....	48
References	49
Appendix. Scoring Rubrics for the TOEFL iBT® Independent Writing Task and the TOEFL® Essentials™ Write for an Academic Discussion Task	58
Notes.....	60

In 2023, several revisions were made to the content, format, and delivery of the TOEFL iBT® test, with the intent of enhancing the test-taking experience, decreasing test administration time, and updating test content. The most substantial revision to test content involved the replacement of the Independent Writing (IND) task with a new Writing for an Academic Discussion (WAD) task to be administered along with the remaining Integrated Writing task as the writing section of the test. Both the IND task and the WAD task elicit extended opinion writing, a typical academic genre that features in all major academic English assessments (Cumming et al., 2021). However, the tasks differ in a variety of ways (described subsequently) and most noticeably in the length of time allowed to respond, with 30 minutes for the IND task and only 10 minutes for the WAD task. Nevertheless, the intent was that the new task would be used to measure the same construct of academic English writing proficiency as the prior task, that it would contribute to the calculation of a total writing section score in the same way as the previous task, and that interpretations about English writing ability would be equally warranted if not somewhat enhanced with the new task.

According to the test revision standards in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014, p. 93), “It is the test developer’s responsibility to determine whether revisions to a test would influence test score interpretations.” In this study, we report on an investigation of the potential impact of the current test revision on score interpretation. Specifically, we compare the two writing tasks in terms of the characteristics of second language (L2) writing ability elicited by each and the rating-scale scores awarded to these writing performances. In the following, we first briefly introduce contemporary approaches to academic English writing assessment, and we consider the role played by task design in the measurement of various dimensions of L2 writing ability. We then sketch out the nature of writing tasks designed for the TOEFL iBT test, from the

perspectives of construct representation, task design, scoring rubrics, and performance descriptors. With this background in mind, we then report on our initial investigation of task comparability in terms of test-taker performance. Ultimately, we consider the findings in light of implications for test score interpretation.

Assessing Second Language Academic Writing Ability

There are numerous approaches to assessing L2 academic writing ability or proficiency, reflecting both the diversity of communicative demands associated with writing in academic settings and the lack of a unifying theory of L2 academic writing proficiency (Cumming et al., 2021). Approaches to writing assessment also depend on the purposes to which tests are put, ranging from standardized proficiency assessments for higher education admissions purposes to placement testing into university writing programs to assessing the development of writing abilities as a result of classroom instruction (Weigle, 2002). The design of academic writing tests and tasks, then, varies considerably. For example, a university writing placement exam might consist of extended assessment activities, including reading input, then essay planning and drafting sessions, followed by revising and editing sessions, or variations on these processes (e.g., Harrington, 2005). Classroom-based assessments, by contrast, might focus on the process of writing specific genres or tasks, including the extended development and testing of research report writing skills (e.g., Romova & Andrew, 2011), or the treatment of more contained tasks like email writing (e.g., Yasuda, 2011).

In standardized assessments of academic writing proficiency, used primarily for making admissions decisions in higher education settings, issues of test length, fairness, and generalizability have constrained the design of writing tasks somewhat (Cumming et al., 2021). Particularly for large-scale English language assessments, writing sections tend to feature

extemporaneous tasks that can be completed within a relatively short amount of time by test takers from a variety of educational and first-language backgrounds around the world and with the goal of eliciting aspects of English writing ability that underlie effective writing across multiple genres. As a result, the most prototypical task type is of the expressive/argumentative variety, asking test takers to express an opinion or to agree/disagree with a position in a multiparagraph essay completed within 20 to 40 minutes.

A second task type that has gained popularity is of the explanatory variety, where test takers are provided with some kind of input and must write an explanation that demonstrates their more-or-less sophisticated understanding of the input. A version of this task type is the TOEFL iBT Integrated Writing task, introduced with the advent of the test in 2006 (Chapelle et al., 2008; see also Plakans, 2015), where learners both read and listen to academic source materials and then write an explanatory response that addresses a related question about the content presented. Although other, more discrete task types (e.g., write a sentence, fill-in-the-blank to complete a sentence or paragraph) may also feature in large-scale academic writing assessments, the extent to which they are capable of tapping into complex, holistic processes (e.g., composing, editing) that reflect actual academic writing by adult writers is dubious (Wagner, 2020).

In large-scale standardized admissions testing, writing performances on extemporaneous task types are scored on the basis of polytomous rating scales that describe characteristics of writing quality at different levels of ability (Shaw & Weir, 2007). Generally speaking, these characteristics are expected to underlie most genres of academic writing, hence the corresponding scores provide a basis for generalizing from test task performance to interpretations of writing ability in the academic domain. Rating rubrics typically feature dimensions such as task completion or accomplishment, discourse cohesion and coherence,

grammatical accuracy and complexity, vocabulary diversity and sophistication, fluency, and mechanics, these features often being combined into a holistic judgment of an overall level on the rating scale. Scoring is undertaken by trained human raters, automated scoring systems, or a combination of both, with the intent that resulting scores represent a generalizable level of academic writing ability (Weigle, 2013).

Despite the constraints on large-scale standardized assessments, recent recommendations for enhancing the validity of interpretations about English L2 academic writing have called for adjustments to the representation of writing in assessment tasks. In light of the challenges of time and test-taker fatigue, Cumming et al. (2021) advocated that “any tasks added to current tests be brief and manageable while also realizing interactionist principles of composing and assessing” (p. 134). Regarding interactionist principles, in particular, they advocated enhancing the descriptions and stimuli used to prompt writing tasks such that the audiences, contexts, and purposes for writing, as well as other important criteria for task accomplishment, are made explicit. In terms of task types, among other possibilities, they recommend that the representation of the academic writing domain be expanded to include new tasks reflective of different purposes and media for writing, attending to distinct genres such as transactional and diversified explanatory writing, as well as expanding the use of integrated writing. These recommendations were considered in the current revisions to the TOEFL iBT writing section. Prior to describing the new writing task, we summarize approaches to measuring different aspects of L2 performances elicited by different kinds of academic writing tasks.

Writing Task Design and the Measurement of L2 Performance

The relationship between different types of writing tasks, as well as the genres or registers they are designed to represent, and various dimensions of language performance has been the subject of substantial research over the past several decades. From early work that attempted to identify direct measures of L2 writing phenomena that correlated with global measures of proficiency (e.g., Ferris, 1994; Ortega, 2003; Wolfe-Quintero et al., 1998) or distinguished among different task types (e.g., Cumming et al., 2005) to initial efforts at automated tagging and analysis of large samples of distinct textual registers (e.g., Biber, 1985; Biber et al., 2002; Grant & Ginther, 2000) to increasingly sophisticated automated measurement of numerous performance features based on natural language processing (NLP) models (e.g., Crossley, 2013; Graesser et al., 2004), this body of research boils down to the basic technical question of which aspects of L2 written performance can be directly and objectively measured (i.e., as opposed to which aspects of writing quality can be subjectively interpreted by a trained human rater). The primary follow-up questions ask to what extent these direct measures, individually or combined into predictive models, can distinguish between specific writing tasks, or between distinct genres of writing, or between L2 learners deemed to be at differing global proficiency levels. Of interest for the current study is how we might utilize direct measures of writing performance as one way of comparing and evaluating the interpretations about L2 writing ability that are attributable to two distinct writing tasks.

Generally speaking, several categories of phenomena have been operationalized through direct measurement of L2 writing performance within this research domain. Perhaps most typically, researchers have attempted to identify and measure the occurrence and co-occurrence of specific aspects of the language system, such as parts of speech (prepositions, pronouns, verbs, nouns, etc.) or grammatical rules displayed in morphology and syntax

(subject-verb agreement, past tense, passive voice, etc.). These phenomena are then counted within selected corpora of texts (e.g., representing different genres of academic writing, such as expressive versus explanatory texts) and differences are used to interpret the effect of task type on writing performance (e.g., Biber & Gray, 2013). Another version of this approach has to do with the measurement of “N-grams” within a corpus of texts that reflect a particular task type or genre (e.g., Garner et al., 2020; O’Donnell et al., 2013; Zhang & Li, 2021). N-grams are groups of words that have a high probability of co-occurring, and they are typically measured in sets of two (bi-grams) or three words (tri-grams). The probability that certain words will co-occur can be estimated for a given text type based on analyses of large corpora, and different texts (e.g., representing different task types or different levels of writer proficiency) can then be compared according to the relative presence or frequency of these high-probability N-grams.

In a distinct approach to direct writing measures, researchers have investigated different ways of capturing complex phenomena that are hypothesized to represent important L2 writing ability constructs, such as syntactic complexity, composing fluency, grammatical accuracy, and lexical diversity or sophistication (e.g., Kuiken et al., 2005; Kuiken & Vedder, 2009; Malvern & Richards, 2012; Norris & Ortega, 2009; Polio, 1997; Polio & Shea, 2014). These measures consist of combinations of various linguistic features calculated in the form of ratios and averaged across texts to produce a summary index of the selected construct. For example, syntactic complexity is often represented in the form of counts of words or clauses divided by the number of sentences or T-units (a main verb and its constituent clauses), and the resulting average (e.g., words per sentence) for a given text is interpreted as an overall indication of how complex the syntactic structures are that the learner produced in that text. For another example, composing fluency is most characteristically calculated as the simple ratio of number of words produced divided by the total amount of time spent composing, but writing fluency

measures may also include editing behaviors observed in process data (e.g., Abdel Latif, 2013; Deane, 2013) or other features (e.g., Tian et al., 2021).

A final example of direct measures involves the identification of features associated with higher order discourse phenomena related to the coherence and cohesion of texts, the development of multiparagraph writing, and the representation of relevant content within texts. Discourse measures generally focus on the relative presence of specific word types or uses, such as transition words, ellipsis, and word repetition, as well as on the overall structure of texts, including the use of topic sentences, main ideas, supporting ideas, conclusions, and related (e.g., Burstein, Tetreault, Chodorow, et al., 2013; Connor, 1990; Crossley et al., 2016a; Crossley & McNamara, 2009). Measures of content may include the representation of ideas in a writing performance as compared with the number of ideas addressed in source materials, as well as the topic relevance of the text as reflected in expected use of terminology (e.g., Attali, 2011; Plakans, 2015).

The academic writing of L2 learners has been widely investigated using direct measures of the kinds outlined above. Of specific interest for the current study is the comparison of different kinds of academic writing on the basis of these measures—as a way of determining the expectations of different task designs in terms of L2 performance—and the potential of different task designs to elicit consistent writing performance differences from learners at differing L2 proficiency levels. Summarizing considerably here, on the one hand, it is clear that different academic writing tasks are frequently associated with different qualities in the writing performances they elicit. For example, Biber and Gray (2013; see also Biber et al., 2014) found clear differences in the frequency of grammatical complexity features (e.g., nominalizations, noun-complement clauses, prepositional phrases, finite relative clauses) used by L2 writers in TOEFL iBT integrated writing tasks versus independent writing tasks. In another example, Yang

et al. (2015) found clear differences in the syntactic complexity produced by L2 writers depending on which of two different essay topics they responded to.

It is also the case that various measures of writing performance are consistently associated with differences in learners' L2 proficiency levels and/or the holistic writing assessment scores assigned by trained raters (e.g., Barrot & Agdeppa, 2021; Byrnes et al., 2010; Crossley, 2020; Crossley & McNamara, 2012; Crossley et al., 2016a; Garner et al., 2019; Jung et al., 2019; Plakans et al., 2019). Thus, in general, the writing performances of learners at higher proficiency levels (and/or higher holistically rated essays) are (a) longer, in terms of the number of words produced in a given time period; (b) more syntactically complex, in terms of the length and combination of different syntactic structures; (c) less prone to grammatical or lexical (word choice) errors; (d) more lexically diverse and sophisticated; and (e) discursively more elaborated, in terms of coherence, cohesion, and textual logic. Of course, the relative values and combinations of these phenomena may vary in important ways across different writing task types, as noted prior.

Of interest for the current study, then, is the nature of writing performances on two extemporaneous writing tasks designed to elicit expressive types of writing. Holistic ratings by trained raters provided one point of comparison between the tasks. In addition, we were interested in the extent to which a handful of representative direct measures of key writing phenomena would reveal similarities or differences in the writing performances of L2 learners on these tasks. We were also interested in the extent to which these measures would distinguish among learners at differing English proficiency levels in similar ways on the two tasks. We turn next to a close consideration of the two TOEFL writing tasks under investigation.

Opinion Writing in the TOEFL iBT® and TOEFL® Essentials™ Tests

The two writing tasks under comparison in the current study were originally designed for two different English proficiency assessments, the TOEFL iBT test and the TOEFL® Essentials™ test. While these two assessments differ in several ways, including how intensively they focus on academic English skills as well as the range of proficiency levels covered, they also share certain design parameters associated with specific construct interpretations. Namely, they both include constructed-response tasks (in both speaking and writing) designed to elicit extended communication in English in order to support interpretations about test-takers' abilities to use language for academic (and other) purposes. These extended communication tasks are also scored on both tests according to rubrics that describe various qualities of language use associated with different levels of task accomplishment. This combination of tasks and rating scales, designed to elicit and evaluate extended communicative performance on real-life academic tasks, provides a basis for extrapolating to test takers' ability to accomplish similar tasks in academic settings.

Of specific interest for the current study, both the TOEFL iBT and TOEFL Essentials tests include a single writing task (in addition to other writing tasks that complete the writing section) that elicits an opinion from the test taker, a type of expressive writing that is very common in higher education. In the case of the TOEFL iBT test, this item is referred to as the Independent Writing (IND) task, while in the TOEFL Essentials test the opinion task is labeled Write for an Academic Discussion (WAD).¹ These tasks both measure the test-taker's ability to create a short piece of writing in English that expresses their ideas in a clear and coherent way (Chapelle et al., 2008; Papageorgiou et al., 2021). In both tasks, the test taker is presented with an opinion question, such as whether they agree or disagree with a statement like "Television advertising directed toward young children (aged two to five) should not be allowed" (ETS,

2023). The test taker must then express an opinion on the given topic and support their view with relevant reasons and examples, drawing from sources such as observation, personal experience, or general knowledge. Beyond these basic similarities, the two tasks also differ in several ways.

In the TOEFL iBT IND task, an example of which is depicted in Figure 1, the test taker sees the question prompt along with a few sentences of instruction regarding the time allowed to write (30 minutes), the expected length of an “effective” response (300 words), and the content of the response (i.e., “use specific reasons and examples to support your answer”). The design intent for this task was that the input materials would provide the minimum information needed for the test taker to generate ideas on the topic and provide relevant elaboration (Cumming et al., 2000). This approach minimizes the time spent understanding input materials while maximizing the time for writing. The strategy of minimizing the input also supplements the other writing task used in the TOEFL iBT writing section, the Integrated Writing task, where test takers must summarize and integrate information presented in both a reading passage and a brief recorded lecture. However, a criticism of this approach to the IND task is that no explicit context is provided for writing; that is, neither the situation, the audience, nor the purpose for writing is provided. Such decontextualized writing may create challenges in making a principled evaluation of whether the test taker’s response is appropriate to the task, and anecdotal reports from test takers have suggested that some individuals find it challenging to quickly form an opinion, especially if the test taker has never given much thought to the specific issue.

Figure 1. TOEFL iBT Independent (IND) Writing Task

29:10

Directions:
Read the question below. You have 30 minutes to plan, write, and revise your essay. Typically an effective response will contain a minimum of 300 words.

Question:
Do you agree or disagree with the following statement?
It is more important for students to understand ideas and concepts than it is for them to learn facts.
Use specific reasons and examples to support your answer.

Cut Paste Undo

While the communicative goal of the TOEFL Essentials WAD task is equivalent to the IND task—to express and support an opinion—the WAD task is framed in a substantially different way. As depicted in Figure 2, the task is contextualized within an online discussion for a university course, where the test taker assumes the role of a student in the course. The general domain of the course is provided (e.g., “economics”), and an instructor figure presents a topic in a few sentences and then poses an opinion question for the class to discuss (Papageorgiou et al, 2021). Previous responses to the question from two fellow students are shown, where each student briefly (40–60 words) expresses and supports diverging views on the issue. The test taker then provides their own views, supporting these with relevant knowledge, experience, or reasoning and responding to the previous posts if desired. Like the TOEFL iBT IND task, the input materials for the WAD task also specify the time to write (10 minutes), the expected length of an “effective” response (100 words), and the content of the response (express and

support an opinion and contribute to the discussion). The context for writing provided in the WAD task is intended to support interpretations regarding what sort of responses are appropriate, and the responses from the other students help the test taker to generate ideas. Compared to the IND task, more time is needed to read and understand the input, but the discussion board scenario creates a situation where a context, purpose, and audience for writing can be established quickly. The WAD task also simulates a type of writing that has become increasingly common in academic coursework since the time the TOEFL iBT test was developed in the early 2000s (Fehrman & Watson, 2021). This is particularly the case in online contexts where such asynchronous discussion may serve as “the primary area where students learn and teachers teach” (Covelli, 2017, p. 140). Additionally, while the IND task provides more time to write (i.e., 30 minutes, versus 10 minutes for the new WAD task), it has been criticized both for the use of abstract decontextualized topics, where test takers may struggle to find something to say, and the potential for negative washback by encouraging mechanistic practice of the traditional “five paragraph essay” (Kim, 2017).

Scoring criteria are similar for both the IND and the WAD tasks (see the appendix), with the scoring rubrics addressing features of

- relevance of content and quality of elaboration;
- clarity of organization and coherence in expression; and
- facility in language use, including range and accuracy of grammar and vocabulary.


Figure 2. TOEFL Essentials Write for an Academic Discussion (WAD) Writing Task

Instructions

Your professor is teaching a class on economics. Write a post responding to the professor's question. In your response you should:


- express and support your personal opinion
- make a contribution to the discussion in your own words

An effective response will contain at least 100 words.




Professor Henson

When people are asked about the most important discoveries or inventions made in the last two hundred years, they usually mention something very obvious, like the computer or the cell phone. But there are thousands of other discoveries or inventions that have had a huge impact on how we live today. What scientific discovery or technological invention from the last 200 years—other than computers and cell phones—would you choose as being important? Why?



Paul N

I mean, we're so used to science and technology that we are not even aware of all the things we use in our daily lives. I would probably choose space satellites. This technology happened in the last hundred years, and it has become important for so many things. Just think about navigation, or telecommunications, or even the military.



Lena A

I am thinking about medical progress. Like, for example, when scientists discovered things about healthy nutrition. I am thinking of identifying all the vitamins we need to stay healthy. I am not sure exactly when the vitamin discoveries happened, but I know they are very important. Our health is much better than it was 200 years ago.

cut

paste

undo

Word Count: 0

hide

9:56

Both tasks are scored on a 0–5 scale, with a score of zero awarded for responses that do not contain any writing, are not in English or are not intelligible, or do not plausibly address the topic. Fully successful responses on both tasks are equated with writing that fulfills the task expectations by providing sufficient elaboration to support clearly stated ideas, exhibits logical organization of ideas, has few lexico-grammatical errors, and effectively utilizes a range of grammatical and syntactic structures. The primary difference in rating criteria between the two tasks has to do with a focus on coherence and progression of ideas in the IND rubric, with this difference attributable to the expectation of longer, typically multiparagraph, writing performances. Otherwise, expectations for other aspects of writing quality are represented in very similar ways at the different score points for both rating rubrics, and in the updated TOEFL iBT test, the scoring rubric for the WAD test is identical to the rubric used for TOEFL Essentials, created during the development of the TOEFL Essentials test (see appendix).²

The Current Study

The current study was motivated by plans to replace the TOEFL iBT IND task with the TOEFL Essentials WAD task in an updated version of the TOEFL iBT test. One requirement for the update was that interpretation of scores should not change substantially; that is, the updated test should be of equivalent difficulty, and scores should support similar interpretations regarding test takers' abilities. Accordingly, the goal of the study was to compare the two tasks in terms of scores produced, as well as the nature of the evidence upon which scores are based (i.e., the characteristics writing produced in response to each task). The comparison of scores supports claims that the level of ability associated with a given test score has not changed, which has important practical considerations for score users. Comparison of the writing elicited by each task supports claims that the abilities measured by the test have not been substantially altered.

The purpose of the study was to provide evidence to evaluate claims of equivalence between the current and updated writing section of the TOEFL iBT test. The study was carried out prior to the release of the updated test, and so made use of existing data collected from individuals who had taken both the TOEFL iBT and TOEFL Essentials tests. Additionally, the study made use of automated text evaluation tools to investigate various writing phenomena, an approach that allowed for the study to be carried out relatively quickly.

Research Questions

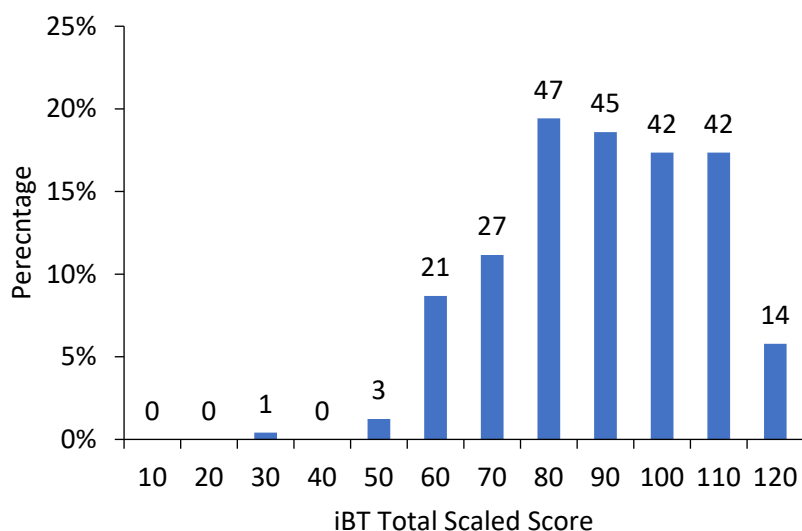
The goal of the study was to provide evidence regarding the equivalence of score interpretations across current and updated versions of the writing section of TOEFL iBT tests, and more specifically, the comparability of the current IND task versus the updated WAD task. Accordingly, the following research questions were addressed in the study:

- To what extent do both writing tasks produce similar results in terms of task scores?
- To what extent are written responses to each task similar or different in terms of various phenomena associated with writing quality?

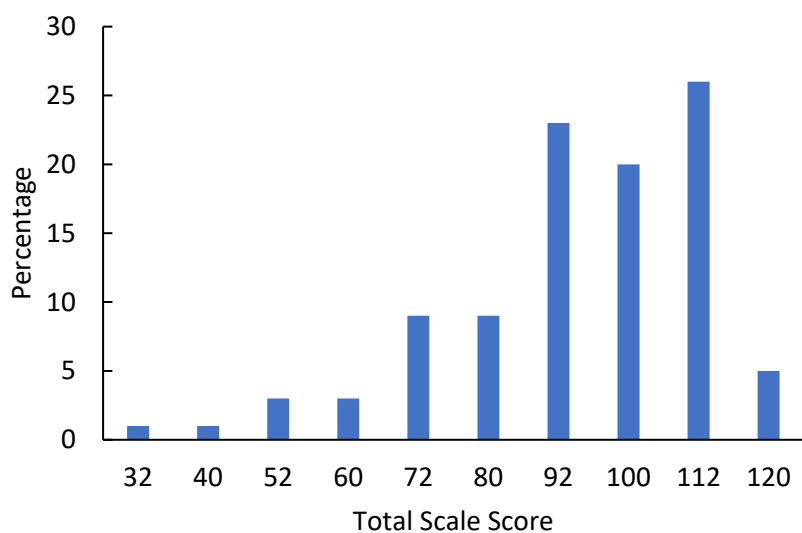
Methods

Participant Sample

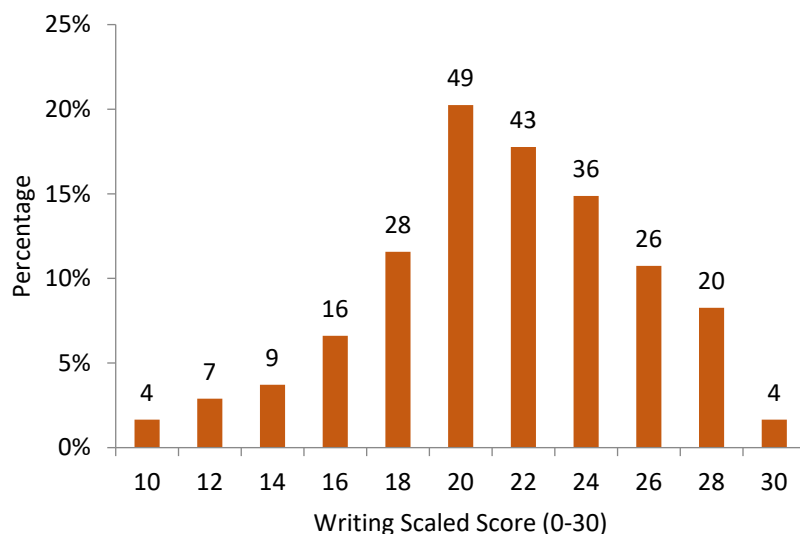
Performances on the IND task and the WAD task were compared for a group of 242 individuals who took operational versions of both the TOEFL iBT and TOEFL Essentials tests. This group was identified from test takers who tested over the period of May 2021 through January 2023. The individuals included in the study represented a range of global proficiency levels as indicated by total TOEFL iBT score (Figure 3); the distribution of proficiency levels among the sample was broadly similar to the test-taking population as a whole (Figure 4). The sample also showed a range of writing ability as indicated by scores on the writing section of the TOEFL iBT (Figure 5). The range of total observed scores extended primarily from low-intermediate to low-advanced levels; that is, levels B1 to C1 in terms of the Common European Framework of Reference (CEFR; Council of Europe, 2001). Using a previously established score mapping between TOEFL iBT and the CEFR (Papageorgiou et al., 2015; ETS, in press-a), the largest group of test takers would be classified into CEFR B2 (45%) with smaller and similar percentages of individuals being classified into levels B1 (22.3%) and C1 (30.2%; Table 1.) For the writing section scores, similar patterns were observed (Table 1), though a higher proportion of test takers scored at the B2 level (57.4%), with a relatively lower proportion at B1 (10.3%) and a handful falling below the B1 level.

Figure 3. TOEFL iBT Total Score for the Individuals Included in the Study

Note. $N = 242$. The number of individuals is shown above each column.

Figure 4. TOEFL iBT Total Score for All TOEFL Test Takers in 2022

Note. Data source: ETS, in press-b.

Figure 5. TOEFL iBT Writing Section Scaled Score for the Individuals Included in the Study

Note. $N = 242$. The number of individuals is shown above each column.

Table 1. CEFR Level of Test Takers as Indicated by TOEFL iBT Total Score and Writing**Section Score**

CEFR level	Total score			Writing section		
	Cut score	<i>n</i>	Percentage	Cut score	<i>n</i>	Percentage
Below B1	<42	1	0.4%	<13	11	4.5%
B1	42	54	22.3%	13	25	10.3%
B2	72	109	45.0%	17	139	57.4%
C1	95	73	30.2%	24	63	26.0%
C2	114	5	2.1%	29	4	1.7%

Note. $N = 242$. CEFR = Council of Europe Framework of Reference. Data source: Papageorgiou et al., 2015; ETS, in press-a.

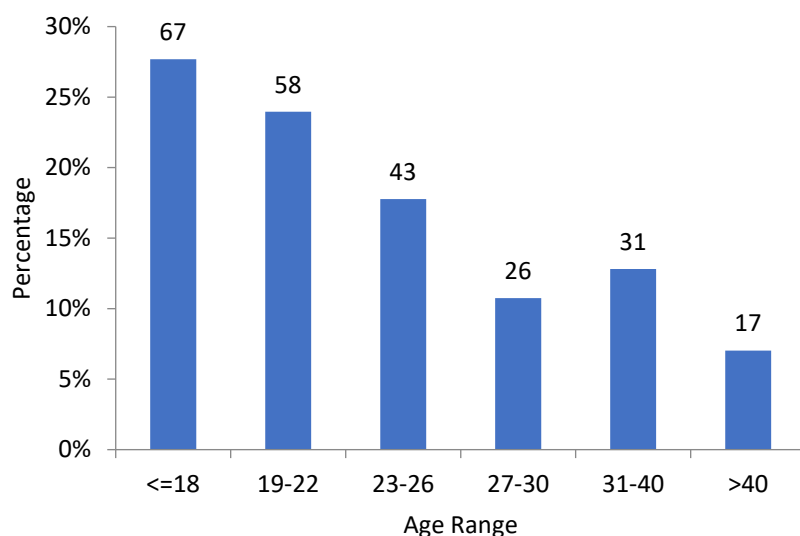
Test takers were relatively evenly split among those reporting as male ($n = 116$) and female ($n = 126$). Individuals came from a wide variety of language backgrounds, with 44 different languages reported as a first language (Table 2). However, the two most frequently reported languages were Chinese (15.7% of test takers) and Spanish (15.3%). Other languages reported by 5% or more of individuals included French, Japanese, English, German, Korean, and

Portuguese. On the whole, the sample captured the major language groups that form the population of TOEFL iBT test takers, although relative proportions diverged somewhat as might be expected from a relatively small convenience sample of test takers. In terms of age, test takers were predominantly in their late teens and 20s, consistent with the broader TOEFL iBT population (Figure 6).

Table 2. Language Backgrounds of Test Takers

Language	Count	Percentage
CHI	38	15.7%
SPA	37	15.3%
FRE	19	7.9%
JPN	17	7.0%
ENG	14	5.8%
GER	14	5.8%
KOR	14	5.8%
POR	12	5.0%
HIN	10	4.1%
ARA	8	3.3%
BEN, RUS, FAS, GUJ, TAM, TEL, TGL, TUR	26	1–2% each
25 additional languages	33	<1% each

Note. $N = 242$. CHI = Chinese; SPA = Spanish; FRE = French; JPN = Japanese; ENG = English; GER = German; KOR = Korean; POR = Portuguese; HIN = Hindi; ARA = Arabic; BEN = Bengali; RUS = Russian; FAS = Farsi; GUJ = Gujarati; TAM = Tamil; TEL = Telugu; TGL = Tagalog; TUR = Turkish.

Figure 6. Age at Most Recent Test Administration

Note. $N = 242$. The number of individuals is shown above each column.

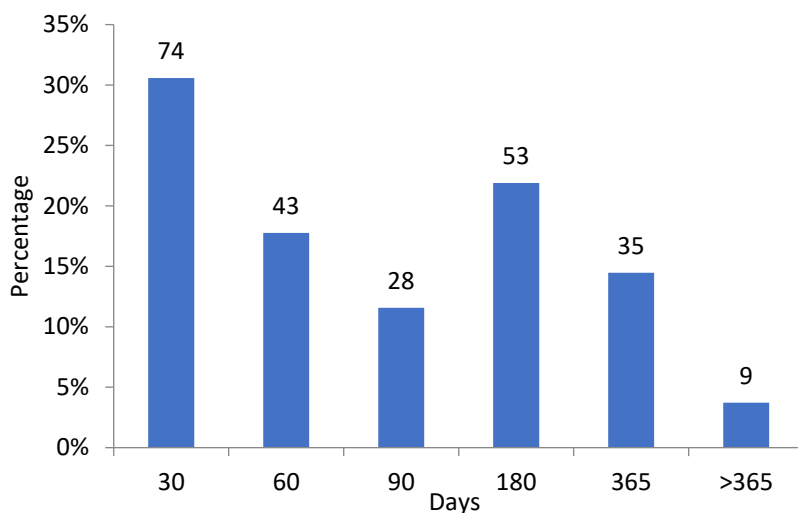
Test Administration

As mentioned previously, responses were drawn from individuals who completed operational versions of both the TOEFL iBT test and the TOEFL Essentials test. Data were drawn from administrations of TOEFL iBT from May 2021 to December 2022; for the TOEFL Essentials test, data were drawn from administrations ranging from the launch of the test in August 2021 to January 2023. Within the TOEFL iBT administrations, 106 IND writing prompts (tasks) were used, with a range of one to 11 responses collected for each prompt. Test-taker responses were obtained for a total of 34 TOEFL Essentials WAD prompts, with one to 22 responses collected for each prompt.

The data were limited to individuals who took each test either once or twice to limit any practice effect; if an individual took either (or both) tests twice, then the testing dates most proximate in time were used as the source of responses and other data. A majority (60%) of individuals took both tests within 90 days of each other, with all but a handful of individuals taking both tests within a 1-year period (Figure 7).

As data were obtained from operational administrations, individuals included in the study experienced the same procedures as all other test takers. The TOEFL Essentials test was completed in an online at-home administration format, where test takers log into their test at an assigned time and are supervised by a remote human proctor who monitors the test takers via webcam (Davis et al., 2023). Test takers use their own computer and are required to take the test in a room where they can be alone. For the TOEFL iBT test, administrations were conducted in a testing center.

During the period the data were collected the total duration of the TOEFL iBT test was approximately 3 hours; the writing section is the final portion of the test. Test takers first complete the Integrated Writing task (20 minutes response time), which requires summarizing academic content presented via a reading and a listening passage. This is followed by the IND task described earlier. Administration of TOEFL Essentials test requires roughly 1.5 hours; writing is the third section of the test, following listening and reading. An initial section of the writing test routes examinees into a second stage adapted for lower or higher proficient writers; those routed to the more advanced stage (a large majority of individuals) complete an email task (7 minutes response time) and the WAD task.

Figure 7. Interval Between TOEFL iBT and TOEFL Essentials Administrations

Notes. $n = 161$ took TOEFL iBT first; $n = 81$ took TOEFL Essentials first. The number of test takers is shown above each column.

Scoring and Linguistic Measures

Test-takers' written responses were officially scored in reference to the scoring rubrics for each task (see previous) and according to operational practices at the time of test administration. In the case of the TOEFL iBT test, responses were scored by both a human rater and a machine algorithm, with the final score being the average of the two. For TOEFL Essentials, each response was scored by two human raters with the average used as the final score. In addition to scores for the writing task, other data obtained included the text of each examinee's written response, information identifying the specific test administration and writing prompt used, total and section scaled scores for each test, and background information collected from test takers during test registration.

The texts produced by examinees on both tasks were subsequently analyzed using NLP technology to extract various direct measures of linguistic features associated with writing ability (as described previously). Automated text evaluation tools used for this purpose were

primarily systems developed at ETS, including TextEvaluator® (ETS, 2017; Sheehan et al., 2014) and e-rater® (Burstein, Tetreault, & Madnani, 2013; Quinlan et al., 2009). Additional measures were obtained from the analysis packages: TAALES 2.2 for lexical sophistication (Kyle & Crossley, 2015; Kyle et al., 2018), TAACO 2.0.4 for cohesion (Crossley et al., 2016b), and TAASC 1.3.8 for syntactic complexity (Kyle, 2016), with the latter also generating syntactic measures based on the Syntactic Complexity Analyzer (Lu, 2010, 2011). The linguistic evaluation tools used in the study generate a large number of measures related to various linguistic phenomena, from discrete frequency counts of different parts of speech to ratio-based construct measures and discourse phenomena. A small, representative subset of these measures was selected for analysis, with selection based on (a) ease of interpretation (i.e., the extent to which the measure was transparent in terms of commonly understood features of writing), (b) the relevance and representativeness of the measure to writing quality (i.e., as reflected in holistic writing rubrics used for rating the performances), and (c) consistency with the observed qualities of the written responses (i.e., measures did not appear to be producing erroneous results, as can sometimes happen when writing does not conform to expected conventions or to the norms on which the automated writing evaluation systems are trained). The selected linguistic features included direct measures of (a) text length and writing fluency, (b) syntactic complexity, (c) grammatical accuracy, (d) lexis, and (e) discourse cohesion and elaboration (see details on specific measures in the results).

Statistical Analyses

As an initial exploration of task performance by a small convenience sample of test takers, the study focused primarily on descriptive statistics and paired comparisons of scores and linguistic measures on the IND task and the WAD task. Linguistic features were compared

using a paired t-test for each feature; effect sizes were also calculated (Cohen's *d*). Calculation of an effect size provides an estimate of the magnitude of difference between tasks, in standard deviation units, and enables judgments of the meaningfulness of such differences. Although the study was considered exploratory rather than confirmatory, the use of repeated significance tests clearly increased the probability of falsely detecting a difference. Accordingly, in order to interpret inferential tests conservatively, the initial critical alpha level of $p < .05$ was reduced to $p < .01$.

Results

RQ1: To what extent do both tasks produce similar scores?

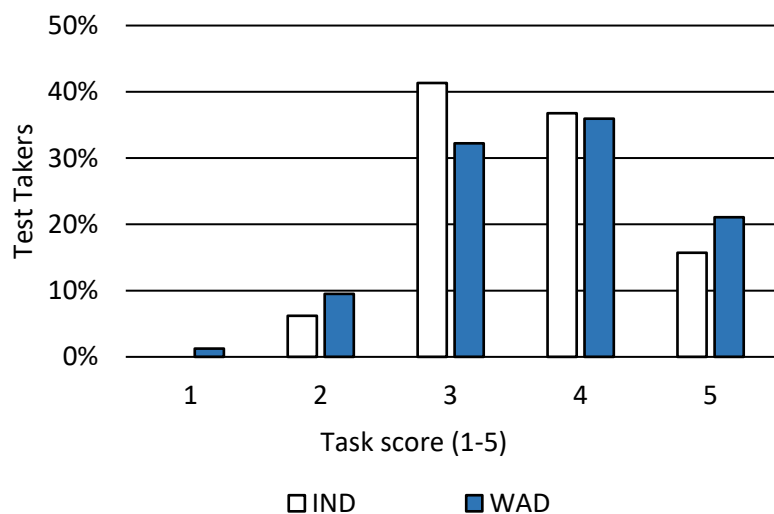
As mentioned earlier, both the IND task and the WAD task are scored on a 0–5 scale. For consistency in comparisons, analyses in this study were based on the score generated by a single human rater for each performance on each task; these scores were produced as part of normal operational scoring for both tests. Operational test scores are calculated as either a combination of a single human score and a machine score (IND task) or were the average of scores from two human raters (WAD task). Automated scoring was introduced for the TOEFL Essentials writing section in late 2022, using human plus machine contributory scoring, as is currently done for the IND task. This same contributory scoring approach will be used in the updated TOEFL iBT test. However, in the test-taker sample available for this study, 97% of individuals came from earlier test administrations of TOEFL Essentials, where automated scores were unavailable. Given the difference in how operational scores were produced for the IND and WAD tasks, and the fact that the machine-learning approach used for automated scoring (linear regression) may exhibit central tendency effects (Johnson et al., 2022), we feel that comparisons based on single human scores are easiest to interpret and provide the most

accurate basis for evaluating whether both tasks elicit similar evidence of writing ability. Accordingly, single human scores were primarily used in the analyses. We also provide a comparison of operational scores for the two tasks, although this comparison should be interpreted with caution given the difference in the operational scoring procedures. In those cases where two human scores were available for a response (i.e., all responses to the WAD task and roughly 10% of responses to the IND task), the score used for analyses was chosen randomly.

The distribution of scores on the two tasks is shown in Figure 8. There appeared to be a slight tendency for task-level scores to be higher for the WAD task, although there was no meaningful difference in overall mean score for the sample (TOEFL iBT mean score = 3.62, TOEFL Essentials mean score = 3.66, paired t-test $t(241) = -0.75$; p (two-tailed) = .45; Cohens $d = -0.05$). The WAD task also appeared to produce a slightly greater spread of scores, but overall both tasks produced similar score distributions. At an individual level of comparison, test takers received the same score on both tasks 50% of the time, and adjacent scores (± 1) in an additional 43% of cases (Figure 9 and Table 3). Only a handful of scores differed by more than one point. When scores differed, the direction of the difference was equally distributed across both tasks (Table 3). We also investigated one instance where an individual received a score of 5 on the IND task and a score of 2 on the WAD task. In this case, the result appeared to be a result of scoring variability, where in our evaluation the score of 2 for the IND task was somewhat lower than justified (and based on a single human score, where the response was apparently returned as unscorable by the automated system). Conversely, the score of 5 for the WAD task appeared rather higher than justified, and in fact the second human rater awarded a score of 4.

The Pearson correlation between individual human raters' scores on the two writing tasks was $r = .61$. This value may reflect a relative lack of measurement precision from using a single human score; note that the correlation between operational scores (which incorporated a machine score or a second human score) was $r = .65$. Additionally, the somewhat truncated nature of the distribution, with the large majority of scores occurring between points 3 and 5 on the scale (94% of IND scores and 89% of WAD scores), likely affected the magnitude of the correlation between scores on the tasks.

Figure 8. Distribution of Task Scores From One Human Rater, TOEFL iBT Task (IND) Versus TOEFL Essentials Task (WAD)



Note. $N = 242$. IND = Independent Writing task; WAD = Write for an Academic Discussion task.

Figure 9. Confusion Matrix of Task Scores From One Human Rater, TOEFL iBT Task (IND) Versus TOEFL Essentials Task (WAD)

		WAD – Write for an Academic Discussion task				
		1	2	3	4	5
IND – Independent task	1	<u>0</u>	0	0	0	0
	2	2	<u>2</u>	8	2	1
	3	1	13	<u>50</u>	31	5
	4	0	8	20	<u>42</u>	19
	5	0	0	0	12	<u>26</u>

Note. Green shaded cells with bold, underlined numbers indicate exact agreement; yellow shaded cells with italic numbers indicate adjacent agreement (+/- 1 point).

Table 3. Agreement Between Task Scores From One Human Rater, TOEFL iBT Independent Writing Task (IND) Versus TOEFL Essentials Write for an Academic Discussion (WAD) Task

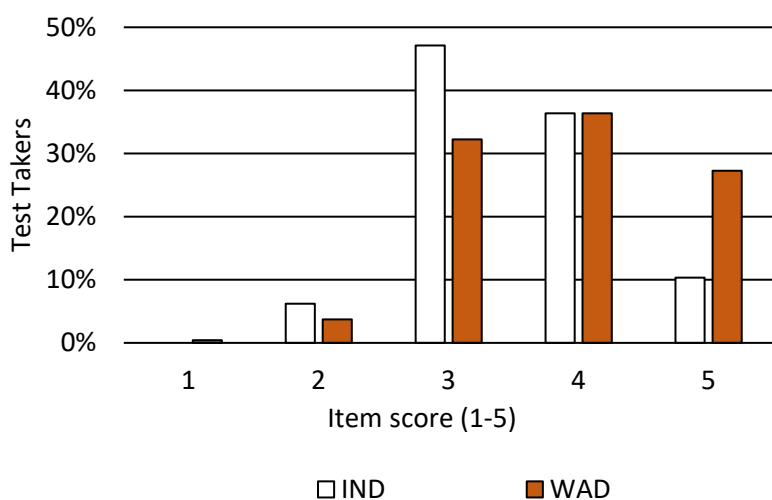
Agreement	All	IND task higher	WAD task higher
Exact	50%	---	---
(+/-) 1	43%	19%	24%
(+/-) 2	7%	4%	3%

Although we believe that human scores provide the most comparable basis for evaluating score agreement across the IND and WAD tasks, operational scores are of primary practical interest to test users. Comparison of operational scores, generated from a combination of human and machine scores (IND task) or two human raters (WAD task), showed similar results (Figure 10, Figure 11). We share these results for completeness; however, a degree of caution is necessary in extending these findings to interpretations regarding operational scores on the updated TOEFL iBT test. Expanding on the caveats mentioned earlier, the WAD task will be scored using human–machine contributory scoring in the TOEFL iBT test, a

different approach than the one used to produce the operational scores that we investigated, which were the only data available prior to release of the updated TOEFL iBT test. This complexity in the comparison of operational scores is a consequence of the use of a convenience sample and is a limitation of the study.

Overall, these data suggest that test takers were distributed in similar ways and received relatively similar scores on both tasks, particularly when considering that these observational data were largely uncontrolled for variables such as task prompt and time elapsed between test administrations, as well as the proficiency range represented in the participant sample.

Figure 10. Distribution of Operational Task Scores, TOEFL iBT Task (IND) Versus TOEFL Essentials Task (WAD)



Note. $N = 242$. IND = Independent Writing task; WAD = Write for an Academic Discussion task.

Figure 11. Confusion Matrix of Operational Task Scores, TOEFL iBT Task (IND) Versus TOEFL Essentials Task (WAD)

		WAD - Write for an Academic Discussion task				
		1	2	3	4	5
IND - Independent task	1	<u>0</u>	0	0	0	0
	2	1	<u>3</u>	8	2	1
	3	0	5	<u>57</u>	45	7
	4	0	1	13	<u>38</u>	36
	5	0	0	0	3	<u>22</u>

Note. Green shaded cells with bold, underlined numbers indicate exact agreement; yellow shaded cells with italic numbers indicate adjacent agreement (+/- 1 point).

The practical impact of replacing the IND task with the WAD task was also evaluated in terms of CEFR classifications for writing ability for a subsample of test takers who took both tests within 120 days ($N = 167$, Figure 12). Simulated scale scores for the updated writing section were computed by replacing the IND task score with the WAD task score for each individual. Using these simulated scores, test takers were classified into the same CEFR level 78.4% of the time. For CEFR level B2, defined as upper intermediate language proficiency and often used as a criterion for university admission, 93% of test takers classified as level B2 in the current TOEFL iBT test would be classified at B2 or higher in the updated test (90 out of 97 individuals). Once again, despite the uncontrolled variables present in the data, proficiency classifications in terms of CEFR levels were found to be similar across current and updated versions of the writing section.

Figure 12. Confusion Matrix of CEFR Levels Derived From Current and Simulated TOEFL iBT Writing Section Scale Scores

		Current Writing CEFR					
		<A2	A2	B1	B2	C1	C2
Simulated Writing CEFR	<A2	<u>0</u>	<i>1</i>	0	0	0	0
	A2	<i>0</i>	<u>5</u>	<i>6</i>	1	0	0
	B1	0	<i>2</i>	<u>11</u>	<i>6</i>	0	0
	B2	0	0	<i>1</i>	<u>78</u>	<i>3</i>	0
	C1	0	0	0	<i>12</i>	<u>34</u>	<i>0</i>
	C2	0	0	0	0	<i>4</i>	<u>3</u>

Note. $N = 167$. Current section scores were obtained from operational administrations of the TOEFL iBT test; simulated section scores were calculated by replacing the IND task score from the operational administration with the score on the TOEFL Essentials WAD task. Green shaded cells with bold, underlined numbers indicate exact agreement; yellow shaded cells with italic numbers indicate adjacent agreement (± 1 point).

RQ2: To what extent are the characteristics of written responses similar or different across task type?

Writing features in responses to the IND and WAD tasks were analyzed using NLP tools. Measures of various linguistic/textual phenomena were then compared to give an initial sense of the similarities or differences in writing produced in response to each task. In the following section, we describe results from these analyses, with linguistic measures addressing similar phenomena grouped together.

Response Length and Writing Fluency

We found that test takers produced more writing in response to the TOEFL iBT IND task compared to the TOEFL Essentials WAD task (Figure 11). This finding is no surprise given that the time allowed to respond to the IND task is 3 times as long as for the WAD task (30 minutes vs. 10 minutes). However, responses to the IND task were not fully 3 times as long in terms of words or sentences produced. When standardized for writing time, test takers produced

somewhat more text per unit of time on the Essentials WAD task (Table 4), including slightly more words per minute and nearly one sentence more per minute. This finding suggests there may have been slight differences in the writing process the test takers used for each task, although the nature of any change in strategy is unclear. One possibility is that in the WAD task test takers spent less time in the ideation and planning stages, which would be consistent with the design intent that additional context for writing enables faster generation of ideas. However, other differences in strategy could have produced a similar result, such as spending more time drafting versus editing. Closer examination of the writing strategies used in each task is an important area for future research.

Table 4. Measures of Response Length

Measure	N	ESS WAD		iBT IND		WAD-IND	
		Mean	SD	Mean	SD	<i>p</i> (2 tail)	Cohen's <i>d</i>
Number of words	242	126.98	38.31	341.43	67.00	0.00	-2.75
Number of sentences	242	6.45	2.38	16.59	5.24	0.00	-1.76
Number of paragraphs	242	1.83	1.25	4.80	1.94	0.00	-1.25
Words_Per10Min	242	126.98	38.31	113.81	22.33	0.00	0.29
Sentence_Per10Min	242	6.45	2.38	5.53	1.80	0.00	0.31

Note. ESS WAD = TOEFL Essentials Write for an Academic Discussion task; iBT IND = TOEFL iBT Independent Writing task; WAD-IND = Write for an Academic Discussion task – Independent Writing task.

Syntactic Complexity

Measures of syntactic complexity showed considerable variability, but the means were similar across tasks (Table 5). We note that variability in complexity measures is expected given the range of language proficiency levels among the test takers in the sample. Mean length of sentence and T-unit (an independent clause plus attached dependent clauses), as well as degree of clausal subordination, were very similar across tasks, with almost no magnitude of difference detected. However, two interrelated phenomena, mean length of clause and mean

length of noun-phrase, showed slightly higher values on the TOEFL iBT IND task. This pattern might be reflective of a somewhat stronger orientation to academic style for the IND task, given that nominalization is a feature typical of academic writing. However, effect sizes were small, so the differences should not be overinterpreted. We also note that syntactic complexity measures produced by automated text evaluation tools can be sensitive to errors in punctuation; accordingly, responses were spot checked as a quality control step to ensure that the measures reported in Table 4 were free of obvious inaccuracies. On the whole, then, measures of syntactic complexity indicated considerable similarity between the IND and WAD tasks.

Table 5. Measures of Syntactic Complexity

Measure	<i>N</i>	ESS WAD		iBT IND		WAD-IND	
		Mean	<i>SD</i>	Mean	<i>SD</i>	<i>p</i> (2 tail)	Cohen's <i>d</i>
Length of sentence ^a	242	23.07	13.03	23.30	9.25	0.71	-0.02
Length of T-unit ^a	242	19.40	9.64	20.18	6.42	0.21	-0.08
Length of clause ^a	242	9.43	2.01	10.02	1.74	0.00	-0.26
Clauses per T-unit	242	2.11	1.07	2.03	0.61	0.28	0.07
Dep. clauses per T-unit	242	1.01	0.83	0.96	0.50	0.33	0.06
Noun phrase length ^a	242	3.09	0.78	3.24	0.66	0.02	-0.15

Note. ESS WAD = TOEFL Essentials Write for an Academic Discussion task; iBT IND = TOEFL iBT Independent Writing task; WAD-IND = Write for an Academic Discussion task – Independent Writing task.

^aLength in number of words

Accuracy

Measures of accuracy were obtained from e-rater, which examines various aspects of accuracy in writing and combines measures of specific types of errors into a global representation. These global measures have been found useful in predicting human judgments of response quality (i.e., scores). No difference between the two tasks was seen in terms of overall grammaticality, grammar errors, or errors in mechanics (Table 6). Word usage errors

were slightly more common in the IND task, perhaps attributable to the fact that more words were used in that task, although the effect size was small as was the overall magnitude of word usage errors for both tasks.

Table 6. Measures of Accuracy

Measure	N	ESS WAD		iBT IND		WAD-IND	
		Mean	SD	Mean	SD	<i>p</i> (2 tail)	Cohen's <i>d</i>
Overall grammaticality	242	2.87	0.286	2.86	0.238	0.65	0.03
Grammar errors	242	-0.11	0.038	-0.11	0.032	0.43	0.05
Mechanics errors	242	-0.18	0.079	-0.17	0.069	0.30	-0.07
Word usage errors	242	-0.09	0.065	-0.10	0.047	0.004	0.19

Note. ESS WAD = TOEFL Essentials Write for an Academic Discussion task; iBT IND = TOEFL iBT Independent Writing task; WAD-IND = Write for an Academic Discussion task – Independent Writing task.

Lexis

In terms of various lexical phenomena, no differences were found across the two writing tasks for lexical diversity (moving-average type-token ratio) or use of collocations (prevalence of ill-formed collocations; Table 7). Responses to the IND task did show somewhat greater use of less-frequent words or academic vocabulary, although effect sizes again were small.

Responses to the two tasks were almost identical for lexical density, if slightly higher and statistically different for the WAD task. Note that measures of lexical density are sensitive to text length (Malvern & Richards, 2012), so this finding could be an artifact of the shorter responses produced for the WAD task.

Table 7. Measures of Lexis

Measure	N	ESS WAD		iBT IND		WAD-IND	
		Mean	SD	Mean	SD	<i>p</i> (2 tail)	Cohen's <i>d</i>
Moving-avg. type-token ratio (MATTR) ^a	242	0.79	0.077	0.79	0.063	0.949	0.00
Ill-formed collocations (per 100 words)	242	2.61	1.589	2.47	1.273	0.299	0.07
COCA acad. content words (log)	242	2.35	0.155	2.38	0.116	0.001	-0.22
Academic word list (normed)	242	0.04	0.027	0.05	0.025	0.017	-0.15
Lexical density (% content words)	242	0.46	0.048	0.45	0.041	0.004	0.19

Note. ESS WAD = TOEFL Essentials Write for an Academic Discussion task; iBT IND = TOEFL iBT Independent Writing task; WAD-IND = Write for an Academic Discussion task – Independent Writing task; COCA = Corpus of Contemporary American English.

^a50-word segments, content words only.

Cohesion

We evaluated cohesion in terms of overall cohesion, use of pronouns, use of connecting words, and word overlap across sentences. There was no meaningful difference across the two writing tasks for these measures (Table 8). It should be noted that these measures generally evaluate cohesion at a relatively local level, that is, within a single sentence or across neighboring sentences. Task-related differences in writing features that establish cohesion over longer stretches of text (e.g., between paragraphs), and which help to establish coherence in organization and elaboration of content, remain to be investigated.

Table 8. Measures of Cohesion

Measure	N	ESS WAD		iBT IND		WAD-IND	
		Mean	SD	Mean	SD	<i>p</i> (2 tail)	Cohen's <i>d</i>
TextEvaluator cohesion score	242	49.68	6.946	49.82	8.118	0.840	-0.01
Pronoun density (3rd person)	242	0.06	0.030	0.06	0.022	0.234	-0.08
Connectives (all types)	242	0.08	0.022	0.08	0.017	0.548	0.04
Content word overlap next 1 sentence	242	1.44	1.008	1.45	1.108	0.912	-0.01
Content word overlap next 2 sentences	242	2.23	1.439	2.29	1.612	0.634	-0.03

Note. ESS WAD = TOEFL Essentials Write for an Academic Discussion task; iBT IND = TOEFL iBT Independent Writing task; WAD-IND = Write for an Academic Discussion task – Independent Writing task.

Elaboration

Given differences in length of response reported earlier, it is reasonable to expect that there would be differences in the degree to which content was elaborated across the two writing tasks. However, the automated indices available for investigating elaboration of content are complex to interpret and the initial findings presented here should be considered tentative until confirmed by a more thorough analysis of a larger data set. No difference in the average length of discourse unit was observed, although more discourse units were detected in the IND task, as expected given increased length of response (Table 9). When evaluating the connectivity between ideas in the response (Somasundaran et al., 2016), average connectivity was very similar in both tasks, although the most-connected node (idea) in the IND task showed more extensive connections to other nodes, again as might be expected for a longer and more elaborated response.

Table 9. Measures of Elaboration

Measure	<i>N</i>	Mean	SD	Mean	SD	<i>p</i> (2 tail)	Cohen's <i>d</i>
Avg. discourse unit length (log)	242	3.88	0.424	3.89	0.389	0.833	-0.01
No. discourse units (log)	242	0.92	0.397	1.93	0.364	0.000	-1.73
Median degree of nodes ^a	242	10.55	4.647	11.22	3.598	0.076	-0.12
Degree of the most connected node ^b	242	27.84	11.543	47.73	17.602	0.000	-0.96

Note. ESS WAD = TOEFL Essentials Write for an Academic Discussion task; iBT IND = TOEFL iBT Independent Writing task; WAD-IND = Write for an Academic Discussion task – Independent Writing task.

^aEncodes the general connectivity structure of text. ^bIndicates the number of connections to other idea units in the text

Illustrative Example Texts

The findings so far suggest that, length aside, the IND and WAD writing tasks elicit similar performance characteristics from L2 writers in terms of a variety of linguistic features. However, the limited set of features analyzed in the current study does not fully represent the more complex aspects of written discourse such as the rhetorical structure of the response. Although detailed measurement and analysis of such complexities was beyond the scope of the current study (and the relatively small data set in particular), the holistic inspection of responses suggests that both tasks tended to elicit a similar organizational structure, as might be expected given that both have the same communicative goal. The longer responses to the IND task tended to include greater elaboration of arguments, whereas responses to the WAD task often employed a similar organization but in a condensed format. To illustrate this pattern, a few examples are provided where responses from both tasks can be compared, for test takers who received high, medium, or low scores on both tasks.

The first example is from an individual who received relatively high scores on both the IND and WAD writing tasks (Figure 13). The response to the IND task includes an introductory paragraph, two paragraphs that support the test taker's position, and a concluding paragraph. The first paragraph provides an introduction of the topic (the importance of preschool) followed by the test taker's opinion (governments should promote low-cost preschool education). Each of the two following paragraphs begins with a statement of an important skill developed in preschool, supported by relevant explanation of how preschool supports these skills, and associated benefits. In the final paragraph, the test taker restates their opinion and provides a final rationale (preschool supports both society and individuals). This brief essay largely conforms to common expectations for argumentative writing, and accordingly received a high score. Note also that the writing is quite accurate and linguistically complex, although spelling errors are apparent.

The response to the WAD task is considerably shorter, but follows the same overall structure of introduction, two supporting arguments, and a conclusion. In the introduction, an opinion and brief context are provided (agreement with a previous student, included in the task input, who argued for the importance of homework). This is followed by supporting arguments that discuss advantages of homework and dispute a claim made by the other student, from the task input, who argued that homework can be ineffectual for learning. In the concluding sentence, the test taker makes a suggestion regarding how the usefulness of homework can be optimized. Each of these rhetorical moves is accomplished in one or two sentences instead of a full paragraph, but the response is recognizable as a coherent argument. It is also notable that the test taker chose to incorporate content from the input materials and demonstrated the ability to express agreement and disagreement in a manner appropriate to the context. Interestingly, while this response is rhetorically similar to the IND response, it is completed in a

somewhat less formal writing style (e.g., use of contractions and first-person reference), which is appropriate to the context of the online discussion forum, though again with linguistically complex structures and overall grammatical accuracy.

The second example (Figure 14) comes from an individual who received a somewhat lower score for both writing tasks, with scores being near the mean for the sample of test takers included in the study. As with the previous example, the response to the IND task begins with a paragraph introducing the topic and stating a position. This is followed by two paragraphs that each introduce a supporting argument, with some elaboration. The final two sentences of the last paragraph may be an attempt at a conclusion (eating food and watching TV from home helps with homesickness), but the point made elaborates on the content of the previous paragraph, and the test taker chose not to start a new paragraph. The response to the WAD task follows a similar pattern, with a statement of a position followed by two supporting arguments, with each rhetorical move accomplished in two to three sentences. Like the response to the IND task, the final sentence of the WAD response may be intended as a conclusion, but a lack of clear signaling makes it challenging to know if this point is an elaboration of the second argument (homework will help them—the students who don’t like studying—to learn more) or potentially a restatement of the original position (“homework is a good tool helping the students to understand the class more”). Note that responses to both tasks are still relatively grammatically accurate though with less linguistic complexity compared to the higher scoring responses.

Figure 13. Response From a High-Scoring Test Taker

<p>Individual 1 Independent Writing Task, task score = 4.4</p>	
<p>Preschool is the essential stage for a child's early development, both academically and socially. Not only does preschool expose young students to a structured learning discipline and routine, but it also provides a chance for the children to develop their social skills. Therefore, Governments should make it possible, if not attempt to lower cost of early education for children.</p>	Introduction and statement of opinion
<p>We humans are social beings, meaning that one of the most crucial skills a person needs to be familiar with is their social skill. Social skills is not a skill that can be learned from reading books or lectures, It is cultivated from experience and exposure to society. That being said, having young children spend time with similarly aged peers at a young age significantly cultivates this skill. This makes preschool a very important step for the kids to increase their social-ness. Making the cost of preschool potentially free would increase the amount of socially developed citizens to a country, meaning a overall positive growth in society as a whole.</p>	Supporting argument #1
<p>Contrasting social skills, academic skills is far more practical in a student environment. skills like time-management, organization, and critical thinking is fundamental to a student's academic success. Providing a environment that enforces learning discipline, routine, and work, would correspond to the increase of these academic skills in children. This makes preschool a very important asset. Having a good foundational development of academic skills at this age would exponentially increase the discipline and academic success rate in the future. The academic success of a student would also provide scholarships to the student who would benefit if their family couldn't afford proper education, thus emphasizing even more on the importance of free preschools.</p>	Supporting argument #2
<p>In conclusion, as a societal standard, preschool should be free because not only does it promote a positive growth in the world but it also financially helps those who cannot afford quality education.</p>	Conclusion

Individual 1**Write for an Academic Discussion, score = 4.5**

I would have to agree with Kelly. Homework is definatly essential to being able to enforce learning concepts into students. Not only does homework force students to review certain ideas learnt in class, but it can also train the application of those ideas in the form of long term assignments and projects. Although paul has stated some reasonable points of how often homework can be presented in a ineffecteve format, I can't see how the removal of out-of-school tasks would improve a student's life, especially in the academic aspect. I do however stress that the school provide materials that vary in different mediums to assist in the student's learning so students can be more enthusiatic in doing the work.

Statement of opinion

Supporting argument #1

Supporting argument #2

Conclusion

Figure 14. Responses From a Medium-Scoring Test Taker**Individual 2****Independent Writing Task, task score = 3.9**

I am very used to feeling homesick since I am studying abroad to Hong Kong. As an international student in a foreign country, it is really precious to experience a completely different culture but always reminds me of home. Based on my personal experience, doing activities or eating foods that remind me of home is the most efficient way to feel less homesick among the options mentioned. Especially, cooking and eating foods related to my country is the best way for me.

Introduction and
statement of opinion

First, other options do not lessen the level of how I miss my home. I would take the fourth option as an example. Making friends in the new community is very crucial while staying in a foreign country. Spending time and getting closer with them helps me to easily adapt to Hong Kong and to learn its culture. However, although it is helpful for me to adapt, it does not stop missing my home. It applies the same to the third suggestion. Keeping in touch with people from home is less effective since sometimes it makes me feel more sad and miss more.

Argument #1

Next, cooking and eating foods that remind me of home somehow increase the level of my sadness. For my case, I like to cook myself and share it to my friends in order to feel better by the reaction of my friends. Also, I sometimes record a video of myself while cooking and send it to my family. It is one of the ways I show my recents and life in Hong Kong to my family. It generates one possible way for me to be connected to them despite the distance. One of the happiest things I like to do is eating Korean food and watching a Korean TV show at the same time. It feels like I am in my home country and not alone

Argument #2

Individual 2**Write for an Academic Discussion, score = 3.5**

I agree what Kelly says. I also believe that homework is a good tool helping the students to understand the class more.	Statement of opinion
The teaching in the limited class time is not sufficient for students to understand the whole lessons. The lack of learning and understanding may be filled with homeworks. If teachers give students relevant homework, they would be able to digest all the lessons.	Argument #1
Also, there must be some students who need their own time to learn the lessons. Without homework, they would not have separate time to study because studying is not fun for the majority of students. However, doing homework, at least, will help them to learn something and understand.	Argument #2

The final example comes from a test taker who received task scores well below the mean (Figure 15). The response to the IND task is noticeably shorter than the higher scoring responses to the IND task but still begins with a statement of a position (without introduction of the topic) and two paragraphs providing supporting arguments. Both supporting paragraphs begin with a clear topic sentence, and whereas the first paragraph is well developed with an extended example, the second paragraph is underdeveloped, and it appears the test taker was not capable of composing a complete response in the time available. The response to the WAD task also contains a statement of opinion followed by two supporting arguments, this time with the second argument more developed and the first argument incomplete, again suggesting the test taker was not capable of composing a fully complete response. Responses on both tasks are also noticeably less grammatically accurate and less linguistically complex than the higher scoring responses.

Figure 15. Responses From a Low-Scoring Test Taker**Individual 2****Independent Writing Task, task score = 1.5**

From my persfpective, I think havigng a paid part time job when they are off school time and days. There are several reasons supporting this view.

Statement of opinion

First, working after school or weekend benefits to the teens for manage their pocket money effectively. For example, one of my class mates in high school used his pocket money for a month during a week and asked extra money for their parents. Finally, his parent let him had a part time job and didnt give him money at all. they offered him only bus ticket and meal coupon. there is no choice for the classmate so he got a part time job at a hamberger shop. He had earn his expense by big efforts. even though he earned three times than his allowance, he started to make a plan how to manage his budget. Likewise, teenagers would learn about money is result about working result and not misusing it.

Argument #1

On the other hand, having experenices in real work place can inspire them to take their future vocation seriously. so move them to study hard or spend more time to be prepare for the future dream.

Argument #2

To be specific, the fried i mention becamas o

Individual 2**Write for an Academic Discussion, score = 1.5**

From my perspective, homework benefits to students to absob what they learend from school. This is because school class has limited time to adapt to

Statement of opinion

Argument #1

I believe when students do homework by theirselves I realize how much they undrestand and know exactly their class. After doing homework, they may have questions that would make their knoweldge firm and develop.

Argument #2

Although the examples shown here come from only three individuals, they provide an illustrative comparison of the writing elicited by the IND and WAD writing tasks. On the one hand, responses to the IND task include more elaboration of arguments and generally appear more similar to the commonly taught pedagogical genre of the opinion essay or the five-paragraph essay. Responses to the WAD task, by comparison, are shorter and less essay-like, but test takers appear to use a similar overall rhetorical structure to accomplish the same communicative goal. It is also clear that the two tasks elicit performances that vary in similar ways for learners at different levels of proficiency in terms of the accuracy and complexity of written language use. Higher scoring test takers produce writing that is linguistically more complex and grammatically more accurate on both tasks than do lower-scoring test takers.

Discussion

The WAD task has several potential advantages over the current IND task, including more efficient use of testing time, greater contextualization of the writing task to facilitate the writer's response, and simulation of a common type of writing found in diverse and contemporary academic contexts. However, such task differences may potentially influence the evidence of writing ability collected from test takers (i.e., the test taker's written response) and the manner in which this evidence is evaluated (i.e., scores). In keeping with recommended standards for large-scale tests, we investigated whether substitution of the TOEFL iBT IND writing task with the WAD task would impact the interpretation of test scores. Specifically, we investigated the performance of a small group of individuals who completed both tasks within operational test administrations. We compared both the scores received for each task, and the evidence upon which the scores are based, by analyzing test takers' written responses for selected textual features that are associated with proficiency in writing.

We found that test takers received similar human rater scores on both tasks, both individually and on average for the overall group. We also found that scores on the two tasks distributed test takers in similar patterns, with the large majority represented at the top three points on the rating rubric. Extrapolating writing section scores based on performances on the new WAD task, we found that CEFR level classifications remained largely unchanged. These findings suggest that the combination of scoring rubric, rater evaluation, and test taker response tend to produce comparable results. The current study did not specifically examine the details of scoring rubrics as interpreted and applied through rater perceptions, but similarity in scores was likely supported by analogous scoring criteria used in rubrics for both tasks (see the appendix) and similarity in communicative genre (express and support an opinion in a short response). We note that the broad and uncontrolled variety of topics that test takers wrote about may have led to uncertain variability in scores, potentially masking any subtle effect of task type. It was also likely that the somewhat truncated range of participant performances/scores depressed the correlation values between the scores on the two tasks. Future research will ideally employ a more controlled experimental design, including a broader sample of participant ability levels. Additionally, an evaluation of the impact of topic and other variable task elements on writing performance would provide useful insights for test development and interpretation. However, as an important initial step we have documented that the two tasks produce generally equivalent scores under operational testing conditions.

Turning to the qualities of writing performances on the two tasks, direct measures of various linguistic phenomena pointed to considerable similarity. The IND writing tasks elicited more writing as expected given the longer time allowed for drafting a response. However, the WAD task elicited more writing per unit of time. This difference is likely related to one or more elements of the writing process, such as a faster drafting phase in the WAD task (one goal of

the added context provided) or relative differences in the time test takers spent on particular activities (e.g., drafting versus editing). Investigation of test takers' writing processes was outside the scope of the current study, however the ability to quickly draft a response is of clear importance for both the IND and WAD task. The need for fluency in producing text is a typical feature of writing tasks used in standardized language assessments, where considerations of practicality often dictate the use of writing tasks that can be completed in a relatively short time.

A majority of the other linguistic measures that we evaluated did not vary systematically across task type, suggesting again that the two tasks elicit writing that is comparable for multiple dimensions of performance. We saw no difference in terms of the measures of cohesion that we evaluated, and overall very few differences in terms of specific measures of syntactic complexity, grammaticality and mechanics, or word use. A few linguistic measures differed across tasks in a manner that may suggest a slightly greater orientation toward academic register in the IND writing task. These measures included slightly greater use of academic vocabulary, as well as somewhat longer noun phrases and clauses, features typical of academic writing (Biber et al., 2011). On the other hand, responses to the WAD task showed marginally higher lexical density (relative frequency of content words) and somewhat fewer word usage errors, both of which may be associated with shorter responses. Initial qualitative inspection of test-taker responses also suggests there were many similarities in writing across the IND and WAD writing tasks, including both foundational elements of writing such as command of syntax and lexis, and higher-order phenomena such as rhetorical structure. Consistent with patterns seen in the direct linguistic measures, these features differed noticeably across responses written by more- or less-proficient test takers, whereas within

responses from individual test takers there was a clear resemblance in the writing produced for each task.

However, questions remain regarding the extent to which the shorter WAD task elicits certain features of writing associated with longer stretches of text, particularly related to the organization and cohesion of content across an elaborated chain of reasoning. The measures of elaboration we used, while challenging to interpret, seemed to indicate that the longer responses elicited by the IND task might be associated with a more complex arrangement of ideas. These same measures also showed that the IND responses contained more discourse units—likely a function of response length—and a larger number of units would logically support a more complex arrangement of content. However, the NLP tools we used were limited in their ability to evaluate discourse-level features of organization and coherence, and further investigation of how these phenomena manifest in the two task types should be a priority for future research. It should also be noted that, for reasons of practicality, both tasks produced texts that were shorter than many important types of academic writing, and supporting inferences regarding test takers’ ability to produce extended academic genres is a broader challenge for standardized language assessment.

Limitations

Findings from the current study should be interpreted with caution, due to a variety of limitations, the most critical of which we highlight here. First, the test-taker participant sample was not controlled in any way and simply reflected the available data from test takers who chose to register for and complete both the TOEFL iBT and the TOEFL Essentials tests within a relatively proximate period of time. It is unknown to what extent participant motivation, test-taking goals, or other individual differences within this participant pool may have affected

performance on one or both of the tasks under investigation. Second, and related, the size and proficiency distribution of the sample limits generalization of the study findings in specific ways. The sample was relatively small, thereby curtailing the number of statistical analyses that might be trusted to produce interpretable comparisons and hence the number of linguistic feature measures that were included. The distribution of test-taker proficiency levels was also clearly condensed, leading to writing task performances that were mostly scored at the upper three score points of the rating scales on both tasks. This distributional artifact no doubt affected the degree of correlation between scores on the two tasks, and the inclusion of lower proficiency and lower scoring test takers would be desirable in future studies. Third, other uncontrolled factors might have affected patterns in the results, including differential amount of time between completion of the two tests, potential variability in the specific task prompts that test-takers were assigned on each test, and possible differences in the testing environments (i.e., at-home versus test center).

Future studies should strive to address these limitations in an effort to produce more generalizable results. One obvious goal would be to conduct an experimental study where both tasks are administered at the same time, are scored using the operational pipelines used in the TOEFL iBT test, and other variables are better controlled. Additionally, aspects of writing related to organization and text-level coherence could only be evaluated at a relatively superficial level by the automated analysis tools used in this study. Further analyses of text-level discourse features will be important for comparing the way test takers establish an effective argument when responding to each task type. Finally, the current study did not investigate the processes and strategies used by test takers when writing their responses. Further research to document the writing process, ideally using both objective measures (e.g., keystroke logs) and subjective

approaches (e.g., think-aloud protocols), will provide useful insights to understand the nature of writing ability elicited by each task.

Conclusion

The current study provides initial evidence in support of using the WAD task as a replacement for and update of the IND task on the TOEFL iBT test. While the two tasks differ in important ways, including the amount of writing time available as well as the provision of key aspects of writing purpose, context, and audience, findings from the study suggest that performances on both tasks can support similar interpretations of test-takers' English writing abilities. Overall, the communicative goal and genre of the tasks are very similar (expressive opinion writing), and the expectations for task accomplishment and writing quality—as exhibited in rating rubrics and performance descriptors—point to largely shared criteria for evaluating test takers' L2 writing in a high-stakes assessment context. Of course, the IND task elicits obviously longer and more elaborated essay writing, and this is one important difference for score users to bear in mind. However, based on comparisons of performances by test takers who completed both tasks within a proximate period of time, scores assigned to the two tasks are very similar, and they distribute test takers into a similar range of abilities related to overall English language proficiency. In addition, direct measures of various dimensions of the writing performances indicate that the tasks elicit fairly similar patterns of writing fluency and organization and considerable similarity in syntactic and lexical complexity, linguistic accuracy, and discourse cohesion. While future research is needed to investigate these patterns within larger and more carefully controlled data sets, and to explore additional questions related to writing processes and other aspects of performance, these initial findings serve as an initial warrant in support of adopting the new writing task.

References

- Abdel Latif, M. M. M. (2013). What do we mean by writing fluency and how can it be validly measured? *Applied Linguistics*, 34(1), 99–105. <https://doi.org/10.1093/applin/ams073>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association. https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf
- Attali, Y. (2011). *A differential word use measure for content analysis in automated essay scoring* (Research Report No. RR-11-36). ETS. <https://doi.org/10.1002/j.2333-8504.2011.tb02272.x>
- Barrot, J. S., & Agdeppa, J. Y. (2021). Complexity, accuracy, and fluency as indices of college-level L2 writers' proficiency. *Assessing Writing*, 47, Article 100510. <https://doi.org/10.1016/j.asw.2020.100510>
- Biber, D. (1985). Investigating macroscopic textual variation through multifeature/multidimensional analyses. *Linguistics*, 23(2), 337–360. <https://doi.org/10.1515/ling.1985.23.2.337>
- Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, 36(1), 9–48. <https://doi.org/10.2307/3588359>
- Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT® test: A lexico-grammatical analysis* (Research Report No. RR-13-04). ETS. <https://onlinelibrary.wiley.com/doi/10.5054/tq.2011.244483>

- Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, 45(1), 5–35. <https://doi.org/10.5054/tq.2011.244483>
- Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, 37(5), 639–668. <https://doi.org/10.1093/applin/amu059>
- Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated evaluation of discourse coherence quality in essay writing. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 289–302). Routledge.
- Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 77–89). Routledge.
- Byrnes, H., Maxim, H. H., & Norris, J. M. (2010). Realizing advanced foreign language writing development in collegiate education: Curricular design, pedagogy, assessment. *The Modern Language Journal*, 94(s1), i–235. <https://doi.org/10.1111/j.1540-4781.2010.01136.x>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. Routledge. <https://doi.org/10.4324/9780203937891>
- Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, 24(1), 67–87, <https://www.jstor.org/stable/40171446>

- Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.
<https://rm.coe.int/1680459f97>
- Covelli, B. J. (2017). Online discussion boards: The practice of building community for adult learners. *The Journal of Continuing Higher Education*, 65(2), 139–145.
<https://doi.org/10.1080/07377363.2017.1274616>
- Crossley, S. A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, 46(2), 256–271. <https://doi.org/10.1017/S0261444812000547>
- Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3), 415–443. <https://doi.org/10.17239/jowr-2020.11.03.01>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016a). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, 32, 1–16. <https://doi.org/10.1016/j.jslw.2016.01.003>
- Crossley, S. A., Kyle, K., & McNamara, D. S. (2016b). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227–1237. <https://doi.org/10.3758/s13428-015-0651-7>
- Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing*, 18(2), 119–135.
<https://doi.org/10.1016/j.jslw.2009.02.002>
- Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35(2), 115–135. <https://doi.org/10.1111/j.1467-9817.2010.01449.x>

- 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>
- Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10(1), 5–43. <https://doi.org/10.1016/j.asw.2005.02.001>
- Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL® 2000 writing framework: A working paper* (TOEFL® Monograph No. MS-18). ETS.
- Davis, L., Norris, J., Papageorgiou, S., & Sasayama, S. (2023). Balancing construct coverage and efficiency: Test design and validation considerations for a remote-proctored online language test. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology mediated language assessment* (pp. 49–63). Routledge. <https://doi.org/10.4324/9781003292395-5>
- Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, 18(1), 7–24. <https://doi.org/10.1016/j.asw.2012.10.002>
- ETS. (2017). *About the TextEvaluator® technology*. <https://textevaluator.ets.org/TextEvaluator/Docs/AboutTextEvaluator.pdf>
- ETS. (2023). *TOEFL iBT® Writing practice questions*. <https://www.ets.org/pdfs/toefl/toefl-ibt-writing-practice-sets.pdf>
- ETS. (in press-a). *TOEFL® steps: Building the learning path of the TOEFL® family*. *TOEFL® Research Insight Series*, 11.
- ETS. (in press-b). *TOEFL iBT® test and score data summary 2022*.

- Fehrman, S., & Watson, S. L. (2021). A systematic review of asynchronous online discussions in online higher education. *American Journal of Distance Education*, 35(3), 200–213.
<https://doi.org/10.1080/08923647.2020.1858705>
- Ferris, D. R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly*, 28(2), 414–420. <https://doi.org/10.2307/3587446>
- Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, 80, 176–187. <https://doi.org/10.1016/j.system.2018.12.001>
- Garner, J., Crossley, S., & Kyle, K. (2020). Beginning and intermediate L2 writer's use of N-grams: An association measures study. *International Review of Applied Linguistics in Language Teaching*, 58(1), 51–74. <https://doi.org/10.1515/iral-2017-0089>
- Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193–202. <https://doi.org/10.3758/BF03195564>
- Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9(2), 123–145.
[https://doi.org/10.1016/S1060-3743\(00\)00019-9](https://doi.org/10.1016/S1060-3743(00)00019-9)
- Harrington, S. (2005). Learning to ride the waves: Making decisions about placement testing. *Writing Program Administration*, 28(3), 9–29.
- Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, 59(3), 338–361. <https://doi.org/10.1111/jedm.12335>
- Jung, Y., Crossley, S., & McNamara, D. (2019). Predicting second language writing proficiency in learner texts using computational tools. *Journal of Asia TEFL*, 16(1), 37–52.
<https://doi.org/10.18823/asiatefl.2019.16.1.3.37>

- Kim, E.-Y. J. (2017). The TOEFL iBT® writing: Korean students' perceptions of the TOEFL iBT® writing test. *Assessing Writing*, 33, 1–11. <https://doi.org/10.1016/j.asw.2017.02.001>
- Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. *EUROSIA Yearbook*, 5(1), 195–222. <https://doi.org/10.1075/eurosla.5.10kui>
- Kuiken, F., & Vedder, I. (2009). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing*, 17(1), 48–60. <https://doi.org/10.1016/j.jslw.2007.08.003>
- Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. ScholarWorks @ Georgia State University. <https://doi.org/10.57709/8501051>
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4), 757–786. <https://doi.org/10.1002/tesq.194>
- Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. <https://doi.org/10.1075/ijcl.15.4.02lu>
- Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, 45(1), 36–62. <https://onlinelibrary.wiley.com/doi/abs/10.5054/tq.2011.240859>

- Malvern, D., & Richards, B. (2012). Measures of lexical richness. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–29). Wiley.
<https://doi.org/10.1002/9781405198431.wbeal0755>
- Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, 30(4), 555–578.
<https://doi.org/10.1093/applin/amp044>
- O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, 18(1), 83–108.
<https://doi.org/10.1075/ijcl.18.1.07odo>
- Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, 24(4), 492–518.
<https://doi.org/10.1093/applin/24.4.492>
- Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL® Essentials™ test 2021* (Research Memorandum No. RM-21-03). ETS. <https://www.ets.org/Media/Research/pdf/RM-21-03.pdf>
- Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT® test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). ETS.
<https://www.ets.org/Media/Research/pdf/RM-15-06.pdf>
- Plakans, L. (2015). Integrated second language writing assessment: Why? What? How? *Language and Linguistics Compass*, 9(4), 159–167.
<https://doi.org/10.1111/lnc3.12124>

- Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing*, 36(2), 161–179.
<https://doi.org/10.1177/0265532216669537>
- Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning*, 47(1), 101–143. <https://doi.org/10.1111/0023-8333.31997003>
- Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing*, 26, 10–27.
<https://doi.org/10.1016/j.jslw.2014.09.003>
- Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater® scoring engine* (Research Report No. RR-09-01). ETS. <https://doi.org/10.1002/j.2333-8504.2009.tb02158.x>
- Romova, Z., & Andrew, M. (2011). Teaching and assessing academic writing via the portfolio: Benefits for learners of English as an additional language. *Assessing Writing*, 16(2), 111–122. <https://doi.org/10.1016/j.asw.2011.02.005>
- Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.
- Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal*, 115(2), 184–209. <https://doi.org/10.1086/678294>
- Somasundaran, S., Riordan, B., Gyawali, B., & Yoon, S. Y. (2016). Evaluating argumentative and narrative essays using graphs. In Y. Matsumoto & R. Prasad (Eds.), *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers* (pp. 1568–1578). Association for Computational Linguistics.
<https://aclanthology.org/C16-1148.pdf>

- Tian, Y., Kim, M., Crossley, S., & Wan, Q. (2021). Cohesive devices as an indicator of L2 students' writing fluency. *Reading and Writing*. Advance online publication.
<https://doi.org/10.1007/s11145-021-10229-3>
- Wagner, E. (2020). Test review: Duolingo English test, revised version July 2019. *Language Assessment Quarterly*, 17(3), 300–315.
<https://doi.org/10.1080/15434303.2020.1771343>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511732997>
- Weigle, S. C. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, 18(1), 85–99.
<https://doi.org/10.1016/j.asw.2012.10.006>
- Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity* (National Foreign Language Center Technical Reports, No. 17). University of Hawaii Press.
- Yang, W., Lu, X., & Weigle, S. C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. *Journal of Second Language Writing*, 28, 53–67.
<https://doi.org/10.1016/j.jslw.2015.02.002>
- Yasuda, S. (2011). Genre-based tasks in foreign language writing: Developing writers' genre awareness, linguistic knowledge, and writing competence. *Journal of Second Language Writing*, 20(2), 111–133. <https://doi.org/10.1016/j.jslw.2011.03.001>
- Zhang, X., & Li, W. (2021). Effects of *n*-grams on the rated L2 writing quality of expository essays: A conceptual replication and extension. *System*, 97, Article 102437.
<https://doi.org/10.1016/j.system.2020.102437>

Appendix. Scoring Rubrics for the TOEFL iBT® Independent Writing Task and the TOEFL® Essentials™ Write for an Academic Discussion Task

TOEFL iBT® Independent Writing Rubrics

SCORE	TASK DESCRIPTION
5	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> Effectively addresses the topic and task Is well organized and well developed, using clearly appropriate explanations, exemplifications and/or details Displays unity, progression and coherence Displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice and idiomaticity, though it may have minor lexical or grammatical errors
4	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> Addresses the topic and task well, though some points may not be fully elaborated Is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications and/or details Displays unity, progression and coherence, though it may contain occasional redundancy, digression, or unclear connections Displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form or use of idiomatic language that do not interfere with meaning
3	<p>An essay at this level is marked by one or more of the following:</p> <ul style="list-style-type: none"> Addresses the topic and task using somewhat developed explanations, exemplifications and/or details Displays unity, progression and coherence, though connection of ideas may be occasionally obscured May demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning May display accurate but limited range of syntactic structures and vocabulary
2	<p>An essay at this level may reveal one or more of the following weaknesses:</p> <ul style="list-style-type: none"> Limited development in response to the topic and task Inadequate organization or connection of ideas Inappropriate or insufficient exemplifications, explanations or details to support or illustrate generalizations in response to the task A noticeably inappropriate choice of words or word forms An accumulation of errors in sentence structure and/or usage
1	<p>An essay at this level largely accomplishes all of the following:</p> <ul style="list-style-type: none"> Serious disorganization or underdevelopment Little or no detail, or irrelevant specifics, or questionable responsiveness to the task Serious and frequent errors in sentence structure or usage
0	<p>An essay at this level merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.</p>

ets.org/toefl

ETS® TOEFL iBT®



TOEFL® Essentials™ Test

Writing Scoring Guide

WRITE FOR AN ACADEMIC DISCUSSION

SCORE	DESCRIPTION
5	<p>A fully successful response</p> <p>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:</p> <ul style="list-style-type: none"> • 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	<p>A generally successful response</p> <p>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:</p> <ul style="list-style-type: none"> • Relevant and adequately elaborated explanations, exemplifications and/or details • A variety of syntactic structures and appropriate word choice • Few lexical or grammatical errors
3	<p>A partially successful response</p> <p>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 exhibits the following:</p> <ul style="list-style-type: none"> • 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	<p>A mostly unsuccessful response</p> <p>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:</p> <ul style="list-style-type: none"> • 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	<p>An unsuccessful response</p> <p>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:</p> <ul style="list-style-type: none"> • 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	<p>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.</p>

Notes

¹ Note that the name of the task, Write for an Academic Discussion, is slightly different in its original format on the TOEFL Essentials test compared with its name on the updated TOEFL iBT test, Writing for an Academic Discussion.

² Additional information regarding construct interpretations about writing ability can be found in the TOEFL iBT and TOEFL Essentials performance descriptors:

<https://www.ets.org/pdfs/toefl/toefl-ibt-performance-descriptors.pdf>

<https://www.ets.org/content/dam/ets-org/pdfs/toefl/toefl-essentials-performance-descriptors.pdf>