Research Memorandum
ETS RM–23-05

Maintaining Score Quality on the Enhanced TOEFL iBT® Test

Lixiong Gu
Shuhong Li
Tongyun Li
John M. Norris

July 2023

ETS Research Memorandum Series

Maintaining Score Quality on the Enhanced TOEFL iBT® Test

Lixiong Gu, Shuhong Li, and Tongyun Li
ETS, Princeton, New Jersey, United States

John M. Norris
ETS Japan, Tokyo, Japan

July 2023

Corresponding author: Lixiong Gu, E-mail: lgu@ets.org

Suggested citation: Gu, L., Li, S., Li, T., & Norris, J. M. (2023). *Maintaining score quality on the enhanced TOEFL iBT® test* (Research Memorandum No. RM-23-05). ETS.

Action Editor: Larry Davis

Reviewers: Jonathan Schmidgall and Spiros Papageorgiou

Abstract

Beginning in July 2023, ETS and the TOEFL® program launched a new version of the TOEFL iBT® test that featured several important updates and enhancements, including a shortened reading section, replacement of the independent writing task with a shorter and more contextualized opinion task called Writing for an Academic Discussion, and removal of unscored pretest items. This memorandum reports on analyses of key psychometric properties of the TOEFL iBT test scores conducted prior to the launch of the enhanced version of the test. Findings revealed sufficiently high levels of reliability and low standard error of measurement values for the reading and writing section scores. In addition, we describe new procedures for ensuring listening and reading section item quality to be deployed during development and after test administration in lieu of pretesting.

Keywords: language assessment, reliability, standard error or measurement (SEM), pretesting, TOEFL iBT® test

Beginning in July 2023, ETS and the TOEFL® program launched a new version of the TOEFL iBT® test that featured several important updates and enhancements. Changes included simplified and shortened instructions throughout the test; removal of unscored, pretest items in the reading and listening sections; removal of one passage and its associated scored questions from the reading section; and the replacement of the current independent writing task with a shorter and more contextualized opinion task called Writing for an Academic Discussion. The enhanced TOEFL iBT test requires approximately 2 hours for a test taker to complete with no break, reduced from the current 3-plus hours with a break.

It is a corporate policy at ETS, following the *ETS Standards for Quality and Fairness* (ETS, 2014), that critical aspects of test quality (e.g., validity, reliability) be re-evaluated in conjunction with any substantial changes to a test prior to its use and that relevant information be provided to test score users. The current research memorandum reports on analyses of key psychometric properties of TOEFL iBT test scores conducted prior to the launch of the enhanced version of the test in order to estimate the potential impact of test revisions on the section and total scores. The particular focus of this report is reliability and standard error of measurement (SEM) for the TOEFL iBT Reading and Writing section scores. In addition, we describe new procedures for ensuring item quality for the TOEFL iBT Listening and Reading section to be deployed during development and after test administration in lieu of pretesting.

Reliability and SEM for the Reading Section

To support the continued high quality of the reported scores for the TOEFL iBT Reading section, ETS researchers and psychometricians evaluated the psychometric properties of the enhanced TOEFL iBT test using four test forms picked randomly from all of the TOEFL iBT administrations in 2022 and the first quarter of 2023. Reliability estimation for the multiple-choice reading sections of the TOEFL iBT is typically carried out using a method based on item response theory (Lord, 1980). Based on this method, Table 1 shows comparisons of the average reliability and SEM for the TOEFL iBT Reading section in its current longer version and in its enhanced shorter version. A comparison of the reliability estimates for the reading section shows only a slight drop in reliability from 0.88 to 0.84 and a small increase in SEM.[1] These small changes indicate that the precision of the test scores for the reading section still meets

ETS's rigorous quality standards. It should be noted that there was no change in reliability for the listening and speaking section scores or the overall test score.

Table 1. Average Reliability and Standard Error of Measurement (SEM) Estimates for the Reading Section Based on Four Test Forms in 2022 and 2023

Test form	Scale	Reliability	SEM
Current test	0–30	0.88	2.38
Simulated enhanced test	0–30	0.84	2.88

Reliability and SEM for the Writing Section

The updated TOEFL iBT Writing section replaced the 30-minute independent writing task with the 10-minute writing for an academic discussion task. Both tasks are scored using very similar 0–5 point rating scales that emphasize organization of ideas, variety and accuracy of lexical and grammatical structures, and appropriateness of the response to the given task. To evaluate the potential impact of the new task type on the reliability of writing section scores, we investigated a sample of 111 test takers who took the TOEFL iBT test and the TOEFL® Essentials™ test, which includes the writing for an academic discussion task, within 30 days of each other. For these analyses, we replaced the score on the independent writing task with the scores from the same test takers on the writing for an academic discussion task, which was performed in the context of the TOEFL Essentials test. Test score reliability for the writing section scores, which is based on an index known as coefficient alpha (Cronbach, 1951), and SEM were calculated for the new simulated writing section scores and compared with those of the current TOEFL iBT test. The results show a slight decrease in reliability and a slight increase in SEM for the new writing section. As with the reading section, these small changes indicate that the precision of the test scores for the writing section still meets ETS's rigorous quality standards.

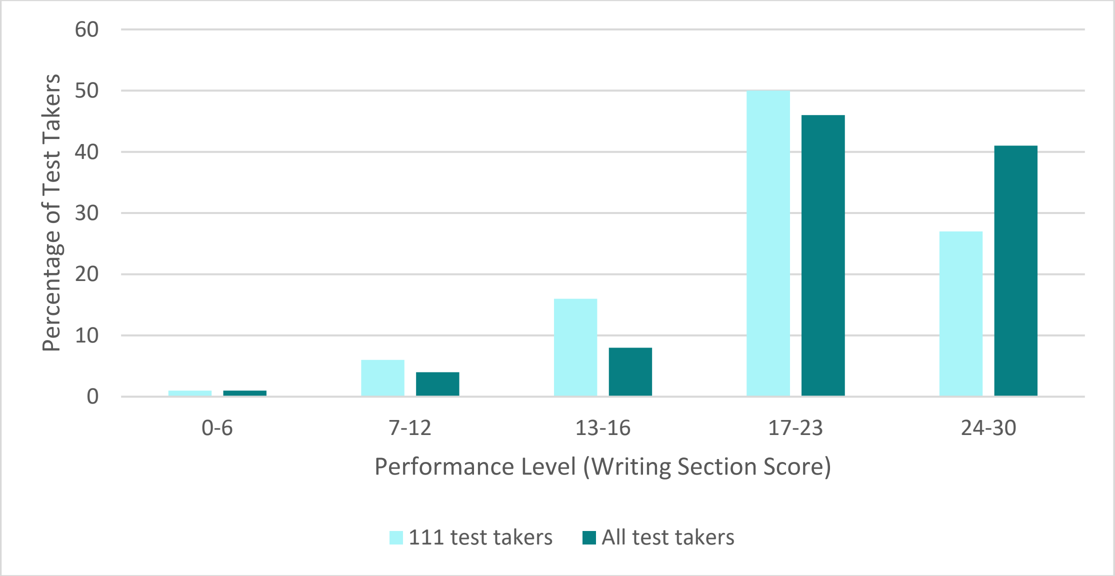
Table 2. Reliability and Standard Error of Measurement (SEM) Estimates for the Writing Section

Writing section	Scale	Reliability	SEM
Current test	0–30	0.80	2.30
Simulated enhanced test	0–30	0.76	2.52

It is worth noting that the 111 test takers in this convenience sample responded to a variety of prompts for both the TOEFL iBT independent writing task and TOEFL Essentials writing for an academic discussion task. For the current test, both writing tasks were taken at the same administration whereas for the simulated enhanced test, the writing for an academic discussion tasks were taken on different dates in a different test, introducing additional variation to the section score, which is reflected in slightly lower section score reliability.

These 111 test takers as a group also appeared to be a low performing subset of the TOEFL iBT test-taker population. Figure 1 shows score distributions based on actual (reported) TOEFL iBT writing scale scores categorized into writing performance levels identified for the TOEFL iBT test (available on the ETS official website: https://www.ets.org/toefl/test-takers/ibt/scores/understand-scores.html). In comparison with the total population of TOEFL iBT test takers in 2022, the distribution of scores for the sample of 111 test takers was shifted toward the lower performance levels. Therefore, it is reasonable to expect that the reliability obtained on a typical TOEFL iBT test-taker sample (covering the entire ability spectrum of the population) will be higher and the SEM will be lower. In operational use, scores from the enhanced test will be closely monitored to document actual score reliability and SEM for the writing section.

Figure 1. Test-Taker Distribution Across the TOEFL iBT Writing Performance Levels



Probability of Obtaining a High Score on the Reading and Writing Sections

A shorter reading section might be taken to imply that any incorrect items become more of a detriment to achieving the highest scores, but in fact there is no systematic difference between the current or enhanced versions of the TOEFL iBT test in terms of the probability of achieving a higher (or lower) section score. The enhanced TOEFL iBT Reading section still contains a sufficient number of high-quality questions to provide an accurate measure of reading comprehension ability, and the enhanced test retains items of similar difficulty, addressing all the same reading subskills that are covered in the current version of TOEFL iBT. Moreover, despite having fewer items in the reading section, the entire range of possible reading scale score points can be achieved with multiple forms in a test administration. Therefore, the 0–30 score scale does not change, nor is there any need to change institutional score requirements. To help ensure test score comparability, score distributions and trends will be closely monitored and compared with previous test data for both the total test-taking population and for subgroups of interest.

The communicative demands of the new writing task, Writing for an Academic Discussion, are similar to the current independent writing task in that both require stating and supporting an opinion. Initial analyses (see Davis & Norris, in press) suggest that both tasks are similar in difficulty as indicated by task scores and both provide similar evidence of effectiveness in the use of language (i.e., the ability to use language structures, conventions, and other features of writing to create a cohesive response that answers the question). In addition, the shorter length will likely reduce the degree of fatigue in completing both the reading and writing sections, decreasing the impact of a potential source of construct-irrelevant variance in scores.

Impact of Removing Items That Do Not Count Toward Scores

Prior to July 2023, the TOEFL iBT test included extra items in the reading or listening sections that did not count toward test taker’s scores. Those extra items served different purposes in the statistical analyses conducted for each administration, with some being new items that were pretested to determine how the items functioned under actual testing conditions. The enhanced TOEFL iBT test no longer includes extra items in the reading or listening sections; rather, evaluation of new reading and listening items will be conducted outside of operational testing as described below. All item development and review activities, including item-level and test-level analyses, will continue to abide by the same comprehensive and rigorous ETS procedures and standards (ETS, 2014).

Before new items are used in operational test forms, they first go through a rigorous review process conducted by experienced ETS assessment specialists, including multiple rounds of content review, fairness review, and editorial review. Items are then tried out with a small sample of the target test-taking population, and test-taker responses are reviewed to determine if items need to be revised or rewritten. Only items that pass the review process will appear in operational test forms.

Immediately following test administration, classical test theory item analysis is used to evaluate item difficulty, item discrimination, and raw score distributions. This analysis helps identify any items that might not perform as expected. Items flagged for poor performance are referred to TOEFL assessment specialists for further review. Assessment specialists then verify that the item keys are defensible and that none of the other response choices are plausible.

Items with serious flaws, if any, are removed prior to the calculation of operational scores to help ensure the quality of test scores reported to test takers.

Conclusion

The current report provides a summary of the psychometric properties of the enhanced TOEFL iBT test, derived from analyses of a small sample of test takers. Given the size and proficiency distribution of the sample, it is likely that the resulting reliability and SEM figures reported here represent a conservative estimation and test score users can interpret them as a lower threshold for the operational administrations of the test. Based on these findings, sufficiently high levels of reliability and low SEM values indicate that scores on the reading and writing sections (along with the unchanged listening and speaking sections) can continue to be trusted as highly consistent measures of test-taker abilities. Similarly, the chances of test takers achieving various scores along the score scale remain unchanged.

Following the launch of the enhanced test, psychometric properties of scores from operational administrations for the enhanced test will be calculated on a rolling basis, and these results will be summarized once sufficient data have accumulated. In addition, item quality characteristics will be calculated based on the performance of items derived from new development and postequating procedures, and they will be compared with the same characteristics of items previously produced through pretesting procedures. In keeping with *ETS Standards for Quality and Fairness* (ETS, 2014), information based on these analyses will be published and shared with test score users.

References

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Davis, L., & Norris, J. M. (in press). *A comparison of two TOEFL writing tasks* [Research memorandum]. ETS.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge. https://doi.org/10.4324/9780203056615

Note

[1] Note that with the existing TOEFL iBT, the SEM accounts for a small portion of the total variance as demonstrated by the high reliability. Hence, even though the SEM increases, error variance remains a very small portion of the total variance in scores and the reliability remains high.