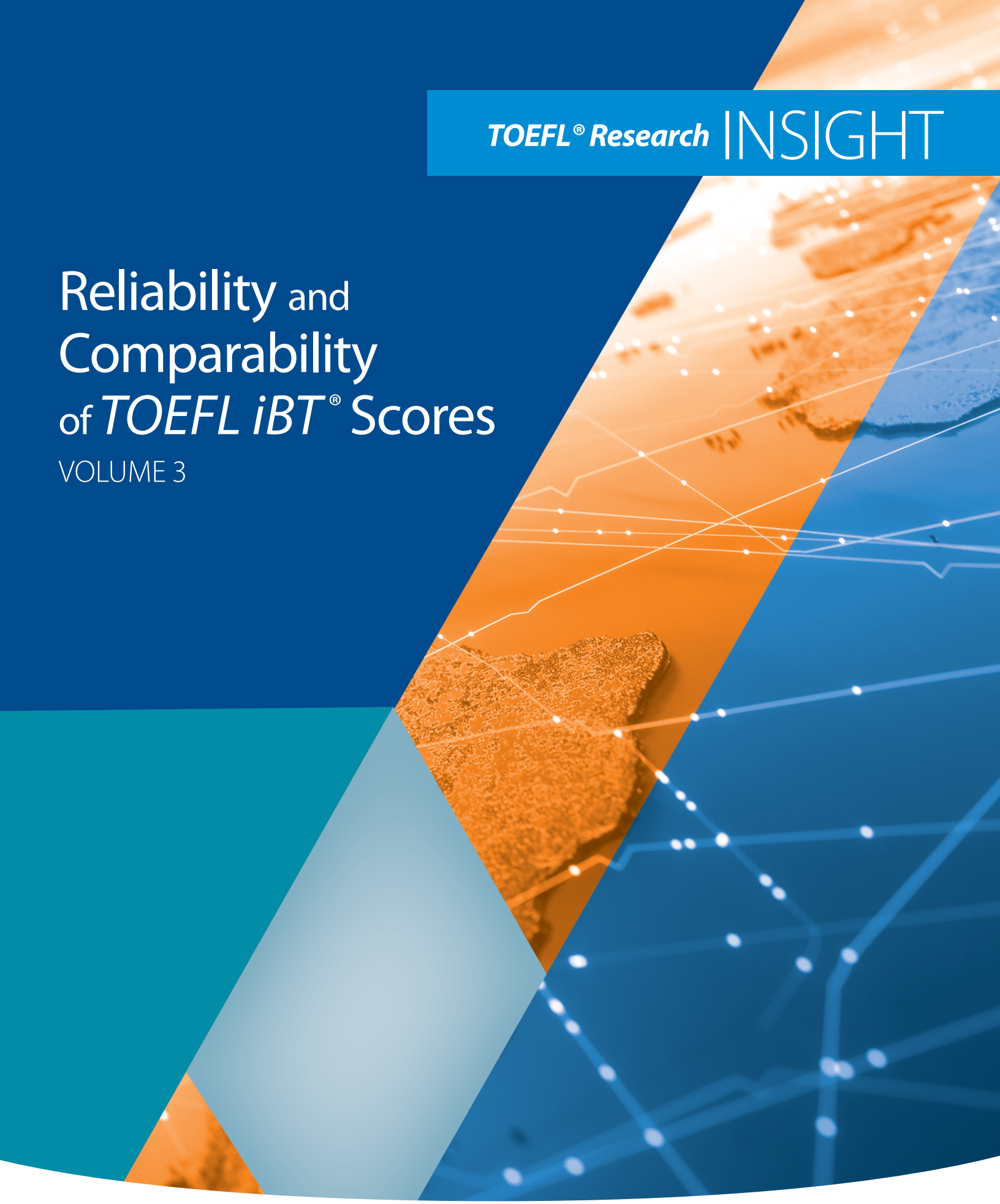
TOEFL® Research INSIGHT

Reliability and Comparability of *TOEFL iBT*® Scores

VOLUME 3



ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Insight Series, Volume 3: Reliability and Comparability of *TOEFL iBT*® Scores

Preface

The *TOEFL iBT*® test is the world's most widely respected English language assessment and used for admissions purposes in more than 150 countries, including Australia, Canada, New Zealand, the United Kingdom, and the United States (see test review in Alderson, 2009). Since its initial launch in 1964, the *TOEFL*® test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the TOEFL iBT test, was launched in 2005. It contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments.

In addition to the TOEFL iBT test, the *TOEFL*® Family of Assessments was expanded to provide high-quality, English proficiency assessments for a variety of academic uses and contexts. The *TOEFL*® Young Students Series features the *TOEFL Primary*® and *TOEFL Junior*® tests, which are designed to help teachers and learners of English in school settings. In addition, the *TOEFL ITP*® program offers colleges, universities, and others affordable tests for placement and progress monitoring within English programs as a pathway to eventual degree programs.

At ETS, we understand that scores from the TOEFL Family of Assessments are used to help make important decisions about students, and we would like to keep score users and test takers up-to-date about the research results that help assure the quality of these scores. Through the publication of the *TOEFL*® *Research Insight Series*, we wish to communicate to the institutions and English teachers who use the TOEFL tests the strong research and development base that underlies the TOEFL Family of Assessments and demonstrate our continued commitment to research.

Since the 1970's, the TOEFL test has had a rigorous, productive, and far-ranging research program. But why should test score users care about the research base for a test? In short, it is only through a rigorous program of research that a testing company can substantiate claims about what test takers know or can do based on their test scores, as well as provide support for the intended uses of assessments and minimize potential negative consequences of score use. Beyond demonstrating this critical evidence of test quality, research is also important for enabling innovations in test design and addressing the needs of test takers and test score users. This is why ETS has established a strong research base as a fundamental feature underlying the evolution of the TOEFL Family of Assessments.

This portfolio is designed, produced, and supported by a world-class team of test developers, educational measurement specialists, statisticians, and researchers in applied linguistics and language testing. Our test developers have advanced degrees in fields such as English, language education, and applied linguistics. They also possess extensive international experience, having taught English on continents around the globe. Our research, measurement, and statistics teams include some of the world's most distinguished scientists and internationally recognized leaders in diverse areas such as test validity, language learning and assessment, and educational measurement.

To date, more than 300 peer-reviewed TOEFL Family of Assessments research reports, technical reports, and monographs have been published by ETS, and many more studies on TOEFL tests have appeared in academic journals and book volumes. In addition, over 20 TOEFL test-related research projects are conducted by ETS's Research & Development staff each year and the TOEFL Committee of Examiners — comprising language learning and testing experts from the global academic community — funds an annual program of TOEFL Family of Assessments research by independent external researchers from all over the world.

The purpose of the *TOEFL Research Insight Series* is to provide a comprehensive, yet user-friendly account of the essential concepts, procedures, and research results that assure the quality of scores for all members of the TOEFL Family of Assessments. Topics covered in these volumes feature issues of core interest to test users, including how tests were designed; evidence for the reliability, validity, and fairness of test scores; and research-based recommendations for best practices.

The close collaboration with TOEFL test score users, English language learning and teaching experts, and university scholars in the design of all TOEFL tests has been a cornerstone to their success and worldwide acceptance. Therefore, through this publication, we hope to foster an ever-stronger connection with our test users by sharing the rigorous measurement and research base, as well as solid test development, that continues to help ensure the quality of the TOEFL Family of Assessments.

John Norris, Ph.D.
Senior Research Director
English Language Learning and Assessment
Research & Development Division
ETS

The following individuals contributed to the second edition (2018) and the third edition (2020) by providing careful reviews and revisions as well as editorial suggestions (in alphabetical order): Terry Axe, Ian Blood, Terran Brown, Lin Gu, Michelle Hampton, Ching-Ni Hsieh, Marcel Ionescu, Yanmei Li, Spiros Papageorgiou, Eileen Tyson, and Lin Wang. The primary authors of the first edition were Mary K. Enright and Eileen Tyson. Cristiane Breining, Daniel Eignor, Rosalie Szabo, Xiaofei Tang, and Xiaoming Xi also contributed to the first edition.

Reliability and Comparability of TOEFL iBT Scores

ETS has always been committed to the quality of its test scores. A fundamental part of test score quality is ensuring that scores reported on different versions (e.g., test forms) of the same test can be interpreted in the same way. In other words, to be useful to score users, test scores must be reliable and comparable. As an ETS assessment program, the TOEFL program strives to ensure score reliability and comparability through strict adherence to the guidelines and practices established for the development and operational implementation of the TOEFL iBT test. Evidence of score reliability and test score comparability is important because this evidence supports an argument that test scores will have the same meaning across test forms. ETS ensures that TOEFL iBT scores are reliable and comparable through five major areas of practice:

- Implementing standardized administration and test security procedures
- Using detailed test specifications to guide test development
- Monitoring score reliability
- Employing an appropriate scale for reporting scores
- Using equating and other means to maintain comparable scores across test forms

In this volume of the *TOEFL Research Insight Series*, we describe the general procedures and guidelines that are used to achieve score reliability and comparability for the TOEFL iBT test, with the aim of helping score users and test takers understand how score quality is ensured by ETS.

Standardized Administration and Security Procedures

In large-scale tests such as the TOEFL iBT test, standardization is a critical part of ensuring score validity and fairness. Standardized test administration and test security measures ensure that the TOEFL iBT test is given under comparable conditions to all test takers, no matter where or when they take the test. The purpose of standardization is to ensure that test scores reflect the test takers' language proficiency rather than other irrelevant factors.

Standardized test administration and test security measures ensure that the TOEFL iBT test is given under comparable conditions to all test takers, no matter where or when they take the test.

The TOEFL iBT test's operational procedures for maintaining standardized conditions for test administration and security follow the requirements laid out in the *ETS Standards for Quality and Fairness* (Educational Testing Service, 2014). The TOEFL program also has provided extensive material to test administrators and test takers so that violations of standard procedures can be reported to ETS for investigation. The major procedures are:

- Certifying all test centers' facilities and equipment (such as hardware, software, and internet connections) for administering the TOEFL iBT test

- Training test center associates on handling test administration sessions, including test-taker identity verification (through biometric data, among other requirements), test launch, and incident and irregularity management
- Providing online practice tests and other supporting information to allow test takers to become familiar with the test and test-taking conditions (e.g., the test section sequence, test duration, use of headphones and microphones, and navigating within and across test sections); information for test takers is available at **https://www.TOEFLGoAnywhere.org**
- Using technology to deliver many different versions (forms) of the TOEFL iBT test per test administration, and to transmit test-related data, in order to ensure the security of the test contents and the test results
- Informing test takers about how to report fraudulent behaviors in a test session

In addition, after each test administration, ETS staff conduct comprehensive statistical analyses of all test takers' response data using advanced techniques to identify test takers with questionable responses. This information is further evaluated and investigated by the Office of Testing Integrity at ETS.

Test Specifications

The TOEFL iBT test offers multiple test administrations each year, and new test forms containing new items are assembled regularly. In order to maintain consistency in the interpretation of TOEFL iBT scores, it is critical to ensure that the test forms used for these different test administrations have comparable content and difficulty. This is accomplished through the use of detailed test specifications to guide the development of new TOEFL iBT test content.

Test specifications are a technical description of test characteristics used to guide the development of new test content. For example, test specifications may define such things as the number of test questions or the content and format of test questions, multiple-choice options, open-ended speaking or writing prompts, or reading and listening passages. *Standards for Educational and Psychological Testing* (American Educational Research Association®, American Psychological Association, & National Council on Measurement in Education, 2014, p. 85) provides general guidance for developing and evaluating test specifications. When multiple forms of a test are developed according to well-defined test specifications, the test characteristics are expected to remain very similar across different test forms and test administrations. Details of how the test specifications for the TOEFL iBT test were developed using a methodology known as *evidence-centered design* can be found in Pearlman (2008).

Score Reliability

A critical aspect of any test's quality is the reliability of its scores. Reliability is crucially important in testing because it indicates the replicability of the test scores. As discussed above, each form of the TOEFL iBT test may be composed of a different set of questions that are written to a common set of design specifications in order to measure the same construct (English language proficiency) and to have the same level of difficulty.

To illustrate the concept of reliability, imagine that a group of 100 English learners takes two TOEFL test forms with no time to study or practice between Form 1 and Form 2, so their level of English proficiency stays exactly the same. If these two different TOEFL test forms were *identical* in their level of difficulty—a feat that is impossible to achieve in real life—we would expect each of the 100 English learners to receive exactly the same score on Form 1 and Form 2. Now, what if these same 100 English learners all received very *different* scores on Form 1 and Form 2, with some scoring very high on the first form and very low on the second form, or vice versa? If we maintain our assumption that the learners' proficiency stayed exactly the same between taking Form 1 and Form 2, then we would have to conclude that their scores are very *un*reliable. Of course, these unreliable scores wouldn't tell us anything about the learners' true English language proficiency.

In the real world, there is no such thing as a perfectly reliable test score. Test results are always influenced to some degree by factors that have nothing to do with the targeted proficiency construct. Imagine, for example, that a test taker is unusually tired or distracted on testing day and performs below his or her true level of language proficiency, or a question on a test that fails to adhere to relevant quality standards is poorly written such that getting the correct answer depends not on the test taker's language proficiency but on random chance. Such irrelevant factors contribute to what is called *measurement error*, which in turn determines how reliable test scores are. The more reliable scores are, the smaller the amount of measurement error. In the field of educational measurement, various methods have been developed for estimating score reliability and expressing it as a statistical index, allowing us to quantify and evaluate the consistency of test scores.

The more reliable scores are, the smaller the amount of measurement error.

In essence, "the concern of reliability is to quantify the precision of test scores and other measurements" (Haertel, 2006, p. 65). Since tests are imperfect, a person's "real" or "true" language proficiency can never be perfectly measured on a test. The observed test score is instead a composite of a true score component and a measurement error component. A well-developed test is expected to yield scores that reflect the test takers' real proficiency as much as possible and minimize measurement error. This is what reliable test scores really mean.

Since a person's true score is never obtainable, the best we can do is to estimate from the observed score using statistical methods. One way that the precision of test scores can be expressed is with a statistical index called a reliability coefficient. A reliability coefficient's values can range from 0 (not at all reliable) to 1 (perfectly reliable). Reliability coefficients are estimated in different ways depending on their intended use and the underlying theoretical framework of the assessment. Reliability estimation for the multiple-choice Reading and Listening sections of the TOEFL iBT test is carried out using a method based on item response theory (IRT; Lord, 1980). For the Speaking and Writing sections of the test, reliability estimates are based on an index known as coefficient alpha (Cronbach, 1951). High reliability is considered a prerequisite for drawing useful inferences from test scores.

Another statistical index used to express the precision of test scores is the standard error of measurement (SEM). To illustrate SEM, imagine that a Super Examinee can take a large number of repeated tests that are

designed to the exact same specifications. This Super Examinee would receive many "observed" test scores, but because these observed test scores always contain some measurement error, none of them would be the Super Examinee's *true* score. This is the case for any reported test score—we can never be certain of a given examinee's true language proficiency score. However, using an *observed* score together with SEM, it is possible to estimate a range above and below the observed score and the chance (typically 68% or 95%) that the *true* score may fall within this range. Generally speaking, one SEM indicates a 68% chance and two SEMs indicate a 95% chance (two SEMs are most often used in practice).

To illustrate, if the test has a score range of 1 to 30 and one SEM equals 2 score points, and if an examinee receives a score of 20 on the test, we now know with 95% certainty that the examinee's *true* score lies somewhere between 16 and 24 (20 plus or minus 2 SEMs). Similarly, if one SEM equals 1 score point, the range would be narrower—we could say with 95% certainty that the Super Examinee's true score lies between 18 and 22. The smaller the value of SEM, the higher the quality of measurement and the more precise the test scores will be.[1]

Table 1 presents the section and total-score reliability estimates and standard errors of measurement (SEMs) based on data from a typical TOEFL iBT test administration.

Table 1. Reliability Estimates and Standard Errors of Measurement

Score	Scale	Reliabilty Estimate	SEM
Reading	0–30	0.87	2.34
Listening	0–30	0.87	2.38
Speaking	0–30	0.86	1.57
Writing	0–30	0.80	2.14
Total	0–120	0.95	4.26

Readers may notice that the reliability estimate for the Writing scores is somewhat lower than that of the Reading, Listening, Speaking, and Total scores. This is because these reliability estimates are computed in a way that tends to yield high reliability coefficients for tests composed of many shorter, less time-consuming tasks with a large number of items (such as typical multiple-choice based reading or listening tasks in most standardized tests). On the other hand, reliability coefficients computed in this way tend to be low for tests composed of a small number of time-consuming tasks (such as the TOEFL iBT Writing section, which consists of only two tasks that measure a similar underlying construct with somewhat different foci). However, it is argued that "A test with written responses to two prompts is not really a two-item test. A score based on judgments by highly trained raters looking at dozens of sentences and hundreds of words does not equate to one based on two multiple-choice questions" (Bridgeman, 2016, p. 21). In fact, the construct of academic writing as defined for the TOEFL iBT test requires the production of *extended* writing samples (Cumming, Kantor, Powers, Santos, & Taylor, 2000). Bridgeman (2016) also recommended reliability estimates from parallel forms (or alternate forms) from repeating test takers.

1 Readers who are interested in deepening their technical understanding of reliability and SEM in the context of educational testing are encouraged to consult the Instructional Topics in Educational Measurement Series (ITEMS) Modules 8 and 9, published by the National Council on Measurement in Education: https://members.ncme.org/ncme/NCME/NCME/Publication/ITEMS.aspx.

Alternate form reliability is calculated based on test takers' scores on two different forms of a test. In practice, only a few test takers would volunteer to take two different versions of the test in two different administrations. However, for reasons of their own, some test takers take the test twice during a period of time that is too short for much learning to occur. An analysis of the scores of these repeating test takers on the two test forms provides an approximation of alternate form reliability. A repeater analysis was conducted on the scores from a sample of about 20,000 test takers who took the TOEFL iBT test twice within 30 days (a time interval deemed unlikely for improving English proficiency in any substantial way) in 2015–2016. The correlations between the two scores of the test takers were the alternate-forms reliability estimates provided in Table 2.

Table 2. Alternate-forms reliability estimates of the TOEFL iBT test (2015 and 2016 data)

Score	Reliabilty Estimate
Reading	0.81
Listening	0.83
Speaking	0.83
Writing	0.81
Total	0.93

The Writing alternate forms reliability was comparable to that of the other three measures. This finding was consistent with the results from previous repeater studies. For example, the test-retest reliability estimates based on 2011 data were 0.76 for the Reading section, 0.75 for the Listening section, 0.84 for the Speaking section, 0.80 for the Writing section, and 0.90 for the Total test score.[2] Because these measures of reliability take into account additional sources of variability, they are typically lower than coefficient alpha or IRT-based estimates of reliability. Consequently, the above test-retest reliability estimates indicate a high degree of consistency in the rank order of the scores of these test repeaters.

A final note to understand these reliability indices is that for making high-stakes decisions, such as admissions to college or graduate school, the TOEFL iBT test total score provides the best information — both because it reflects all four language skills and because it is the most reliable measurement. Nevertheless, there are circumstances under which decision makers may want to examine individual section scores for test takers, such as when studying the success of a particular curriculum, when evaluating the possible need for additional language training, or when success in an academic program requires a specific language skill to be well-developed. When making high-stakes decisions, score users should always also consider a number of non-TOEFL test-related factors, such as grade point average, scores on other admissions exams, teacher recommendations, or interviews with individuals.

2 The test-retest correlation coefficients were adjusted for range restriction.

Evaluating *e-rater*® Engine Performance

The TOEFL iBT test implements ETS's *e-rater*® automated essay scoring system to produce a contributory score for both the *independent* and *integrated* TOEFL iBT Writing tasks. A candidate's score on each Writing task is based on one human rater's score and one *e-rater* score. For each test administration, a random sample of about 300 responses to each Writing task is scored by two human raters so that the engine's performance can be evaluated in relation to the two human raters' scores. ETS also conducted research in 2011 to assess how well *e-rater* and human rater scores on one occasion may predict TOEFL iBT test takers' Writing scores on a later occasion; the results were published in the previous version of this series. Following the same design as in the 2011 study, ETS researchers obtained a repeater sample (i.e., test takers who took the test twice within 30 days), from the 2015 and 2016 TOEFL iBT testing years and correlated the test takers' *e-rater* and human rater scores from Time 1 (first test) with the human rater scores from Time 2 (second test). The Time 2 score was the sum of Human Rater 1 scores on the integrated and independent tasks; Time 1 scores were computed in three different ways: a) using only human rater scores, b) using only e-rater scores, and c) combining human and e-rater scores. Table 3 displays these correlations for each Writing task and the total raw score. The results were consistent with those of the 2011 study, and those obtained by Bridgeman, Trapani, and Williamson (2011). For the independent task, the *e-rater* scores predicted Time 2 Writing scores better than the human rater scores did, while for the integrated task *e-rater* scores and human rater scores at Time 1 had the same correlation with the Time 2 Writing scores. The combination of *e-rater* and human rater scores on the integrated task at Time 1 appeared to be the best predictor of Writing performance at Time 2. For the independent task, however, the *e-rater* scores predicted the Time 2 Writing scores better than the combined *e-rater* and human rater scores did. Moreover, the total raw Writing scores, based on adding combined *e-rater* and human rater scores on each task, yielded the highest correlation with Writing scores at Time 2.

Table 3. Correlation of Time 1 Writing Scores with Time 2 (Independent + Integrated) Human Rater Scores in a Sample of 2015–2016 Repeat Test Takers

Task	Time 1 Score	Correlation with Time 2 Human Rater Scores
Integrated	*e-rater* score	0.64
	Human Rater 1 score	0.64
	Combined *e-rater* & Human Rater 1 score	0.69
Independent	*e-rater* score	0.68
	Human Rater 1 score	0.57
	Combined *e-rater* & Human Rater 1 score	0.64
Total Raw Score	Human Rater 1 score on integrated task + Human Rater 1 score on independent task	0.70
	Combined *e-rater* & Human Rater 1 score on integrated task + Human Rater 1 score on independent task	0.72
	Human Rater 1 score on integrated task + combined *e-rater* & Human Rater 1 score on independent task	0.72
	e-rater score on integrated task + *e-rater* score on independent task	0.70
	Combined *e-rater* & Human Rater 1 score on integrated task + Combined *e-rater* & Human Rater 1 score on independent task	0.74

Evaluating *SpeechRater*® Scoring System Performance

The TOEFL iBT test implements ETS's *SpeechRater*® automated speech scoring system to produce a contributory score for both *independent* and *integrated* TOEFL iBT Speaking tasks (a Speaking section contains one independent and three integrated tasks). A candidate's score on each Speaking task is based on one human rater's score and one *SpeechRater* score. For each test administration, a random sample of about 300 responses to each Speaking task is scored by two human raters, which enables the *SpeechRater* system's performance to be evaluated in relation to the two human raters' scores.

ETS has also conducted research to assess and compare the assessment reliability for the Speaking section under three scenarios: a) using only *SpeechRater* scores; b) using only human rater scores, and c) combining human and *SpeechRater* scores in a contributory approach (the operational scenario). Table 4 below provides the assessment reliability for these three scenarios when true scores estimated based on entire Speaking forms are used. Table 4 shows that the combination of human and machine scores provides better assessment reliability than using either only human rater scores or only machine scores.

Table 4. Comparison of assessment reliability (provided as the PRMSE value, i.e., the proportional reduction of mean squared error, a measure of reliability) for TOEFL iBT test Speaking sections for three different scenarios.

Scenario	Assessment Reliabilty
SpeechRater scores only	0.76
Human rater scores only	0.75
Contributory scoring (both human and machine scores are used)	0.83

In 2019, operational scoring was updated to include the *SpeechRater* engine as a provider of contributory scores to be combined with human rater scores. As operational test data accumulate, analyses of *SpeechRater* scores will be conducted in the same manner as has been done for *e-rater* scores to monitor the performance of the *SpeechRater* engine in the scoring of TOEFL iBT spoken responses.

Scaling TOEFL iBT Scores

The scores that appear on TOEFL iBT score reports are derived from performance on the test through statistical processes called *scaling* and *equating*. The purpose of scaling is to help score users interpret the meaning of scores. A carefully developed score scale, together with an equating plan (described in the following section), is important in maintaining score comparability and meaningful interpretation of scores across test forms and over time.

A carefully developed score scale, together with an equating plan, is important in maintaining score comparability and meaningful interpretation of scores across test forms and over time.

Imagine a student, Juana, who answers 55 questions correctly on a 60-question test. If each correct answer is worth 1 score point, Juana's score is 55. This is Juana's number-correct or *raw* score. Now imagine a second student, Lina, takes a different form of the same test. Lina's test form is designed to the same specifications as Juana's test form, and Lina also gets 55 questions correct. Both Juana and Lina got the same number of questions correct, but can we say that their performance is equal? The answer is no, because despite the best efforts of test designers, no two test forms can ever be exactly alike. Each form is composed of different questions, which means that each test form differs slightly from other test forms in its level of difficulty. Therefore, number-correct (raw) scores are bound to a specific test form and are not directly comparable across forms, and for this reason they should not be reported when there is a need to compare scores from different forms of a test.

Applying scaling and equating procedures is one way to deal with the problem of comparability of scores across test forms. Through scaling and equating, a test's raw-score scale is transformed into a reporting score scale. Scores derived from the reporting score scale are called *scale* scores, and they have the property of comparability — that is, any scale score can be meaningfully compared to another scale score, even if the two scores were derived from different test forms.

Each TOEFL iBT section score is reported on a 0–30 scale. The scales for each section were established in a field trial in 2003–2004. A TOEFL iBT test form was administered to participants from 31 countries representing the typical TOEFL test-taker population. The same scale score range (0–30) for the four test sections was chosen to indicate that all sections should be viewed as equally important in measuring the construct of academic English language proficiency. The TOEFL iBT total score is calculated as the sum of the four section scores (0–120). The decision to use a 0–30 scale was based primarily on the need to provide reasonable raw-to-scale score mappings for each of the test sections, which differ in their maximum raw scores. The maximum number of raw score points on the four sections of the form used in the field study ranged from 20 for Writing to 44 for Reading. Although the scale scores of the four TOEFL iBT sections all range from 0 to 30, they do not have the same meaning and cannot be directly compared. Each section of the TOEFL iBT test is a separate measure of language proficiency, and each measure is on its own scale.[3]

3 For additional information on the TOEFL iBT section score scales and to view performance descriptions for different scale score levels, visit https://www.ets.org/toefl/ibt/scores/understand/.

Maintaining Score Comparability Across Test Forms

For testing programs like the TOEFL iBT test that have multiple administrations with different test forms, it is necessary to maintain score comparability across test forms. Score comparability across test forms is typically maintained using a statistical process called *equating*. Kolen and Brennan (2004) defined equating as follows:

> *Equating* is a statistical process that is used to adjust scores on test forms so that scores on different forms can be used interchangeably. Equating adjusts for differences in difficulty among forms, which occurs even though they are built to be similar in difficulty and content. (p. 2)

For tests containing selected response items, such as the TOEFL iBT Reading and Listening sections, equating is routinely carried out to produce reported scores for a new test form. For tests composed of items that require human scoring of spoken or written responses, as is the case for the TOEFL iBT Speaking and Writing sections, a variety of statistical and nonstatistical procedures have been put in place to minimize differences in test form difficulty and potential inconsistency due to human scoring.

Equating TOEFL iBT Reading and Listening Sections

A nonequivalent group anchor test design (also called a common item nonequivalent group design) is used as the equating data collection design for the TOEFL iBT Reading and Listening sections. This means that each new form of the TOEFL iBT test contains an "anchor block," which is a set of items that have been pretested in previously administered forms. This design enables an adjustment for possible proficiency differences between the group to whom the items were administered in pretesting and the group to which the items are administered with the current new test form. This is possible because the same items are given to the two groups of candidates, and the differences in the item statistics for the two groups reflect the proficiency differences between the two groups. Such differences need to be adjusted during the equating process.

The scale scores for the test forms are directly comparable, as they indicate the same levels of proficiency.

A statistical model within the IRT framework is used to analyze the characteristics of items and the test takers' proficiency. In the IRT-based analyses, item parameters (such as difficulty) and test takers' proficiency levels are estimated together. The estimated item parameters and proficiency levels are put on the same metric as the TOEFL iBT Field Trial form's IRT scales that were established with data from the original TOEFL iBT Field Trial and later updated with operational test data. This way, the item parameters and test takers' proficiency estimates from different test administrations can be directly compared.

The IRT true-score equating method (see detailed descriptions about this method in Kolen & Brennan, 2004) is implemented to establish the relationship between scores on a current form and the previous or "base" form. After equating, raw scores on the new form are adjusted to be equivalent to raw scores on the base form. Because each raw score on the base form already corresponds to a scale score between 0 and 30, each raw score on the new form can now be related to a scale score. The scale scores for the test forms are directly comparable, as they indicate the same levels of proficiency.

Comparability of TOEFL iBT Speaking and Writing Sections Across Forms

The process used to equate multiple-choice reading and listening tests is not practical or feasible for constructed-response writing and speaking tests, which typically contain a much smaller number of tasks that take a longer amount of time to complete. As described above, many equating procedures require including previously administered items in a current test administration. Including such "linking" items is often not feasible if a writing test has only one or two tasks that are easily remembered and shared with other test takers, as doing so would constitute a threat to test security.

Threats to score comparability on constructed-response speaking and writing tests result from both differences in test form difficulty and from inconsistency in human raters' scoring. In the absence of applicable equating procedures, an innovative statistical linking procedure has been implemented to achieve the comparability of TOEFL iBT Speaking and Writing scores across forms. This new procedure uses a weighted equi-percentile approach to link the Speaking and Writing scores on a new form such that the linked scores achieve statistical equivalency to the scores on any previous test forms for the TOEFL iBT test-taker population.

In addition, a number of nonstatistical procedures have been put in place to minimize differences in test form difficulty and potential inconsistency due to human scoring. Careful test development effort and rigorous scoring standards are used to maintain score quality for the TOEFL iBT Speaking and Writing sections. Detailed task specifications guide the development of parallel tasks, and small-scale tryouts are used to screen out poorly performing tasks.

To ensure fairness, operational scoring of both TOEFL iBT Speaking and Writing responses is accomplished through a centralized scoring model, which means scoring does not take place at each test site but rather through a centralized scoring network that implements and ensures consistent scoring standards. To minimize rater bias, a test taker's responses are never all scored by the same rater. All raters are required to complete training for using well-defined and articulated scoring rubrics and to pass the Speaking or Writing scoring certification before they become eligible for the rater pool. In addition, prior to each live scoring session, raters must pass a topic-specific calibration test; on any given day, raters cannot begin to score unless they first pass this calibration test. Furthermore, during each scoring session, raters are monitored and supervised by scoring leaders. When problems arise, raters are retrained or replaced.

Careful test development effort and rigorous scoring standards are used to maintain score quality.

Several statistical methods are also implemented to monitor and evaluate the performance of Speaking and Writing tasks on each test form and the performance of the raters. All the constructed-response tasks in the TOEFL iBT Speaking and Writing sections are analyzed after a test is given. The analysis yields such statistics as average scores on a task, distributions of all the scores on a task, and the correlations between the Writing or Speaking sections with the Reading and Listening sections. The performance of raters is statistically evaluated using rater agreement rates, which include both exact agreement (no score difference between two raters) and adjacent agreement

(1 point difference) rates. The average of all the scores a rater assigns to a particular task in a scoring session is compared with the average score of all the raters participating in the same session who scored that task. A large difference between these two average scores may alert a scoring leader to a possible problem in a rater's performance.

Whenever possible, "monitor papers" are also used to evaluate cross-administration scoring consistency. Monitor papers are selected responses on a task from a prior test administration that have already been scored. After a new test administration, these monitor papers are mixed with the responses to the task on the new test for scoring. Because these monitor papers are indistinguishable from the responses to the task on the new test, raters will score them in the same way as they score the new responses. Then the old and new scores on monitor papers are compared. The agreement rates between the two sets of scores indicate cross-administration raters' consistency in scoring.

Several statistical methods are also implemented to monitor and evaluate the performance of Writing and Speaking tasks on each test form and the performance of the raters.

Another type of statistical evidence used to evaluate score comparability across forms comes from the analysis of repeat test takers. As noted in the section on reliability, an analysis was conducted on test takers that chose to take the test twice within a short period of time. The correlational analyses established that the test takers were rank ordered consistently on the two test forms and that, for most test takers, differences in scores across the two test forms were negligible for all four test sections and the total score. This finding was consistent with that of an earlier study conducted by Zhang (2008).[4]

Conclusion

Because different versions (test forms) of the TOEFL iBT test are administered to test takers at different times and different locations, score reliability and comparability are important criteria for evaluating the quality of the test. ETS implements a variety of statistical and non-statistical procedures to monitor and enhance test score reliability and comparability. Evidence of score reliability and comparability for TOEFL iBT scores comes both from statistical analyses and from the application of best practices of test development, administration, and scoring. This evidence allows decision makers to have confidence in the trustworthiness of TOEFL iBT scores.

4 For more information about measures ETS takes to ensure the quality of assessments that use constructed-response item types and human ratings, see *Guidelines for Constructed-Response and Other Performance Assessments* (Baldwin, Fowles, & Livingston, 2005).

References

Alderson, J. C. Test review: Test of English as a Foreign Language™: Internet-based Test (TOEFL iBT®). *Language Testing*, 26(4), 621-631. doi:10.1177/0265532209346371

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

Baldwin, D., Fowles, M., & Livingston, S. (2005). *Guidelines for constructed response and other performance assessments*. Princeton, NJ: Educational Testing Service.

Bridgeman, B. (2016). Can a two-question test be reliable and valid for predicting academic success? *Educational Measurement: Issues and Practice, 35*(4), 21–24.

Bridgeman, B., Trapani, C, & Williamson, D. (2011, April). *The question of validity of automated essay scores and differentially valued evidence*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

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

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph No. 18). Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2014). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pearlman, M. (2008). Finalizing the test blueprint. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 227–258). New York, NY: Routledge.

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT* (Research Memorandum No. RM-08-05). Princeton, NJ: Educational Testing Service.