TOEFL iBT™ Research

Insight

Series I, Volume 3

Reliability and Comparability of TOEFL iBT™ Scores

We are very excited to announce the TOEFL iBT™ Research Insight Series, a bimonthly publication to make important research on the TOEFL iBT available to all test score users in a user-friendly format.

The TOEFL iBT test is the most widely accepted English language assessment, used for admissions purposes in more than 130 countries including the United Kingdom, Canada, Australia, New Zealand and the United States. 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 the use of integrated tasks that engage multiple language skills to simulate language use in academic settings, and the use of test materials that reflect the reading and listening demands of real-world academic environments.

At ETS we understand that you use TOEFL iBT test scores to help make important decisions about your students, and we would like to keep you up-to-date about the research results that assure the quality of these scores. Through the TOEFL iBT Research Insight Series we wish to both communicate to the institutions and English teachers who use the TOEFL iBT test scores the strong research and development base that underlies the TOEFL iBT test, and demonstrate our strong, continued commitment to research.

We hope you will find this series relevant, informative and useful. We welcome your comments and suggestions about how to make it a better resource for you.

Ida Lawrence
Senior Vice President
Research & Development Division
Educational Testing Service

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, because it is only through a rigorous program of research that a testing company can demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. This is why ETS has made the establishment of a strong research base a consistent feature of the evolution of the TOEFL test.

The TOEFL test is developed and supported by a world-class team of test developers, educational measurement specialists, statisticians and researchers. Our test developers have advanced degrees in such fields as English, language education and linguistics. They also possess extensive international experience, having taught English in Africa, Asia, Europe, North America and South America. Our research, measurement and statistics team includes some of the world’s most distinguished scientists and internationally recognized leaders in diverse areas such as test validity, language learning and testing, and educational measurement and statistics.

To date, more than 150 peer-reviewed TOEFL research reports, technical reports and monographs have been published by ETS, many of which have also appeared in academic journals and book volumes. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), comprised of language learning and testing experts from the academic community, funds an annual program of TOEFL research by external researchers from all over the world, including preeminent researchers from Australia, the UK, the US, Canada and Japan.

In Series One of the TOEFL iBT Research Insight Series, we provide a comprehensive account of the essential concepts, procedures and research results that assure the quality of scores on the TOEFL iBT test. The six issues in this Series will cover the following topics:
Reliability and Comparability of TOEFL iBT™ Scores

Issue 1: TOEFL iBT Test Framework and Development
The TOEFL iBT test is described along with the processes used to develop test questions and forms. These processes include rigorous review of test materials, with special attention to fairness concerns. Item pretesting, try outs and scoring procedures are also detailed.

Issue 2: TOEFL Research
The TOEFL Program has supported rigorous research to maintain and improve test quality. Over 150 reports and monographs are catalogued on the TOEFL website. A brief overview of some recent research on fairness and automated scoring is presented here.

Issue 3: Reliability and Comparability of Test Scores
Given that hundreds of thousands of test takers take the TOEFL iBT test each year, many different test forms are developed and administered. Procedures to achieve score comparability on different forms are described in this section.

Issue 4: Validity Evidence Supporting Test Score Interpretation and Use
The many types of evidence supporting the proposed interpretation and use of test scores as a measure of English-language proficiency in academic contexts are discussed.

Issue 5: Information for Score Users, Teachers and Learners
Materials and guidelines are available to aid in the interpretation and appropriate use of test scores, as well as resources for teachers and learners that support English-language instruction and test preparation.

Contributors
The primary authors of this section are Mary Enright and Eileen Tyson.

The following individuals also contributed to this section by providing their careful review as well as editorial suggestions (in alphabetical order):

Cris Breining
Rosalie Szabo
Xiaofei Tang
Xiaoming Xi

Issue 6: TOEFL Program History
A brief overview of the history and governance of the TOEFL Program is presented. The evolution of the TOEFL test constructs and contents from 1964 to the present is summarized.

Future series will feature summaries of recent studies on topics of interest to our score users, such as “what TOEFL iBT test scores tell us about how examinees perform in academic settings,” and “how score users perceive and use TOEFL iBT test scores.”

The close collaboration with TOEFL iBT score users, English language learning and teaching experts and university professors in the redesign of the TOEFL iBT test has contributed to its great success. Therefore, through this publication, we hope to foster an even stronger connection with our score users by sharing the rigorous measurement and research base and solid test development that continues to ensure the quality of TOEFL iBT scores to meet the needs of score users.

Xiaoming Xi
Senior Research Scientist
Research & Development Division
Educational Testing Service

Reliability and Comparability of TOEFL iBT™ Scores
ETS has always been committed to the quality of its test scores. As an ETS assessment program, TOEFL strives to ensure score reliability and comparability through strict adherence to guidelines and practices established for the development and operational implementation of its products and services. Evidence of score reliability and comparability is important because it suggests that test scores will have the same meaning across test forms. ETS strives to ensure that the test scores of the TOEFL Internet based test (iBT) are reliable and comparable by:

• implementing and adhering to standardized administration and test security procedures
• using detailed test specifications to guide test development
• monitoring score reliability and generalizability
• employing an appropriate scale for reporting scores
• using equating and other means to maintain comparable scores across test forms

Standardized Administration and Security Procedures
In large-scale tests such as the TOEFL iBT test, a critical component to ensuring score validity and fairness is accomplished by implementing and adhering to standardized procedures for test administration and related test security.

As a result, the test scores reflect the test takers’ abilities and are not unduly influenced by other, unrelated factors. TOEFL iBT operational procedures for maintaining standardized conditions for test administration and security follow the requirements laid out in the ETS Standards for Quality and Fairness (ETS, 2002).

Test Specifications
Another important way to help ensure the comparability of scores across test forms is to create detailed specifications to guide the test development process. Test specifications are a detailed operational definition of test characteristics that form an exact content sampling plan. For example, test specifications may define the kind of content covered by the test, the number of test questions, the format of test questions and responses, and the response options. The Standards for Educational and Psychological Testing (AERA, APA, NCME, 1999, p. 43) provides general guidelines for developing and evaluating test specifications. When multiple forms of a test are developed according to well-defined test specifications, the test characteristics can be expected to remain very similar across the test forms and across test administrations.

TOEFL program also provides extensive material to test administrators and test takers so that violations of such procedures can be reported to ETS for investigation. The major procedures are:

• certifying all test centers facilities and equipment for administering TOEFL iBT tests, including hardware, software and Internet connections
• training test center staff on how to handle a test administration session, including test taker identity verification, test launch and incident and irregularity management
• providing online practice tests and other supporting information for test takers to become familiar with the test and test-taking conditions, including section sequence, test duration, use of headsets and microphones and navigating within and across test sections (information for test takers is available at www.toeflgoanywhere.org)
• using technology to control test delivery and transmission of test-related data to ensure the security of test content and test results
• informing test takers about how to report fraudulent behavior in a test session

www.toefl.org
www.toeflgoanywhere.org
www.madebyets.com
www.toefltesteverywhere.com
www.toefl.com
www.toefl.org
The TOEFL iBT test offers multiple test administrations each year, so it is critical to ensure that test forms used for these administrations are similar in content. This is accomplished by following the test specifications for the TOEFL iBT test in the test development process. Details of how these specifications were developed using Evidence Centered Design can be found in Pearlman (2008).

Score Reliability and Generalizability

An important measure of the quality of a test is how reliable the test scores are. Reliability is important because it indicates the replicability of the test scores when either a test could be given twice or more to the same group of people, or two tests constructed in the same manner could be given to the same group of people. Testing, like other measurement events, is subject to the influence of many factors that are not relevant to the ability being measured. Such irrelevant factors contribute to what is called "measurement error," which in turn determines how reliable test scores are.

In educational measurement, score reliability is a statistical index to quantify and evaluate the consistency of test scores. In essence, "the concern of reliability is to quantify the precision of test scores and other measurements" (Haertel, 2006, p. 65). A test score is a measurement outcome of a test scored test questions instead of selecting from a list of possible answers. Using an analysis of variance, generalizability theory separates different sources of variance (test taker ability, rater effects, and task effects) so that the effect of each source of variance (called a facet) can be evaluated. The index for score reliability in this framework is a generalizability coefficient (G-coefficient), which is also on a scale of 0 to 1 with a value closer to 1 being desired. A 'person by task' generalizability model is used for the Speaking scores, whereas a nested model is used for Writing (rating nested within task, and this is crossed with person) (see Lee & Kantor, 2005 for details).

The above-mentioned reliability and generalizability analyses are conducted for every test form. Table 1 presents the average section and total score reliability estimates and standard errors of measurement based on operational data from 2007.

Table 1. Reliabilities and Standard Errors of Measurement

<table>
<thead>
<tr>
<th>Score</th>
<th>Scale</th>
<th>Reliability Estimate</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>0-30</td>
<td>0.85</td>
<td>3.35</td>
</tr>
<tr>
<td>Listening</td>
<td>0-30</td>
<td>0.85</td>
<td>3.20</td>
</tr>
<tr>
<td>Speaking</td>
<td>0-30</td>
<td>0.88</td>
<td>1.62</td>
</tr>
<tr>
<td>Writing</td>
<td>0-30</td>
<td>0.74</td>
<td>2.76</td>
</tr>
<tr>
<td>Total</td>
<td>0-120</td>
<td>0.94</td>
<td>5.64</td>
</tr>
</tbody>
</table>

The reliability estimates for the Reading, Listening, Speaking, and Total scores are relatively high, while the reliability of the Writing score is somewhat lower. This is a typical result for writing measures composed of only two tasks (Breland, Bridgeman, & Fowles, 1999) and reflects one well-documented limitation of performance testing—reliability estimates for measures composed of a small number of time-consuming tasks are often lower than estimates for measures composed of many shorter, less time-consuming tasks. However, the construct of the TOEFL iBT test is such that this limitation is not a concern. The TOEFL iBT test is designed to be a comprehensive test of English language proficiency, and it is expected that examinees will be able to perform well on all sections of the test.

Scaling TOEFL iBT Scores

Reported test scores are derived from performance on a test through a statistical process called "scaling." In a simple example, a student who answered 55 questions correctly out of 60 questions on a test would receive a score of 55 if each correct answer was worth one score point. This score is the raw score. However, raw scores are not always the most meaningful way to report test results. Standardized scores, such as the scaled scores that are reported, take into account the difficulty of the test and the performance of other examinees. This allows for better comparison of scores across different test administrations and for making decisions based on the scores, such as admissions decisions to college or graduate school. The Total score provides the best information, both because it reflects all four language skills and because it is the most reliable. Nevertheless, there are circumstances under which decision makers may want to examine the profile of scores for test takers, such as the demands of the curriculum or a need for additional language training. Also note that ETS encourages score users to consider a number of other factors, when making admissions decisions, including grade point average, scores on other admissions exams, teacher recommendations, and interviews with individuals.

The reliability estimates in Table 1 are what are used for the TOEFL iBT operational test scores. Other types of reliability estimates also exist that take into account other sources of variability such as differences in test forms or changes in examinees' performances from day to day. Alternate form reliability, for example, is calculated based on examinees' scores on two different forms of a test. This requires examinees to take two different test forms, something only a few examinees would volunteer to do. But some examinees do take the test twice during a period of time too short for much learning to occur, for reasons of their own. An analysis of the scores of these repeat test takers on the two test forms provides an approximation of alternate form reliability. Zhang (February 2008) compared the test scores of more than 12,000 examinees who were identified as having taken two TOEFL iBT tests within a period of one month. The correlations of their scores on the two test forms were 0.77 for the Listening and Writing sections, 0.78 for the Speaking section, and 0.84 for the Total score. Because these measures of reliability take into account additional sources of variability, they are typically lower than internal consistency measures. Nevertheless, they indicate a high degree of consistency in the rank ordering of the scores of these test takers.

SEM = σ_x \sqrt{1 - r_{xx}} ,

where σ_x is the standard deviation of the scaled score, r_{xx} is the reliability estimate, and solving for σ_x will then obtain the estimated reliability. The SEM is on the same metric as the scaled scores that are reported.
is directly dependent on the specific items on a particular test form; this particular form may not have exactly the same difficulty level as other forms of the same test. As a result, the same raw scores from two different test forms may not represent the same level of performance or ability. Instead, the raw scale is transformed to a reporting score scale, which produces a scaled score.

The scales for the measures on the TOEFL iBT test were established such that the same scale range (0–30) for the four sections was chosen to indicate that all sections should be viewed as being equally important in measuring the construct of academic language ability. The total score would be the sum of the four section scores. The decision to use a 0-to-30 scale was based primarily on the need to provide reasonable raw to scale score mappings for each of the sections, which differed 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.

### Maintaining Score Comparability across Test Forms

For testing programs that have multiple administrations with different test forms, it is necessary to maintain score comparability across test forms for meaningful comparison of scores. Score comparability across test forms is typically maintained using certain statistical processes called equating.

**Equating** is a statistical process that is used to adjust scores on test forms so that scores on the forms can be used interchangeably. “Equating adjusts for differences in difficulty among forms that are built to be similar in difficulty and content” (Kolen & Brennan, 1995, p. 2).

For tests containing selected response items, equating is routinely carried out to produce reporting scores for a new test form, as is the case for TOEFL iBT Reading and Listening sections.

Such an equating process, however, is not practical or feasible for performance-based tests that contain only a few tasks. For example, a writing test may have only one or two writing tasks that are scored by human raters using a scoring rubric (rules or guidelines for assigning scores to constructed-response questions or performance tasks). Many equating procedures require repeating previously administered items in the current administration. This may not be feasible if a writing test has only one or two tasks that are easily remembered and shared with other examinees.

Threats to score comparability on such performance tests result from both differences in test form difficulty, and inconsistency in human raters’ scoring activities. In the absence of any feasible equating procedures, various other statistical analyses can be used to evaluate and control the quality of scores (Baldwin, Fowles, & Livingston, 2005).

### Comparability of TOEFL iBT™ Speaking and Writing Sections across Forms

Speaking and Writing scores are not equated statistically due to technical and test security constraints. The two test sections have six and two constructed response tasks, respectively. Because these few tasks are prone to memorization, test security concerns preclude the possibility of repeating tasks on every new test form, as is required by equating. To minimize differences in test form difficulty and potential inconsistency due to human scoring, a number of non-statistical procedures are put in place in test development and scoring.

Detailed task specifications guide the development of parallel tasks. Then, small-scale tryouts are used to screen out poorly performing tasks. Training is given to raters using well-defined and articulated scoring rubrics. In addition, raters are certified before they can begin scoring work and, prior to each scoring session, they must pass a calibration test. Furthermore, during all scoring sessions, raters are monitored and supervised by chief raters.

All the constructed response tasks in Writing and Speaking sections are analyzed after a test is given. The analysis examines such statistics as average scores on a task, distributions of scores on a task, and correlations between the Writing or Speaking sections with the Reading and Listening sections. The performance of raters is evaluated by statistics on rater agreement rates, which include both exact agreement (no score difference between two raters) and adjacent agreement (1 point difference). The average of all the scores a rater assigns in a scoring session is compared with the average score of all the raters participating in the same session. 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 were scored previously. If the task is occasionally included in a new test administration, these monitor papers are intermixed 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 the monitor papers are compared, and the agreement rates between the two sets of scores indicate cross-administration rater consistency in scoring.

Another type of statistical evidence of score consistency across forms comes from the analysis of repeat test takers. As noted in the section on reliability, Zhang (2008) analyzed the test scores of examinees who chose to take the test twice within a short period of time. The correlational analyses established that the examinees were rank-ordered consistently on the two test forms. Zhang also reported that the differences in scores on the two test forms for all four sections and for the total score were negligible for most examinees.

### A carefully developed score scale, together with an equating plan (see following section), is important in maintaining score comparability and meaningful interpretation of scores across test forms and over time.
Reliability and Comparability of TOEFL iBT™ Scores

Conclusion

Because different forms of the TOEFL iBT test are administered to test takers at different times and in different locations, score reliability and comparability are important criteria to evaluate the quality of the test. Therefore, ETS has implemented a variety of procedures to enhance test score reliability and comparability.

Evidence of score reliability and comparability allows decision makers to evaluate the trustworthiness of the test scores when the scores are used to indicate candidates’ abilities or performances.

Evidence of score reliability and comparability for TOEFL iBT scores comes both from statistical analyses and from the application of accepted test development, administration and scoring practices.

References


