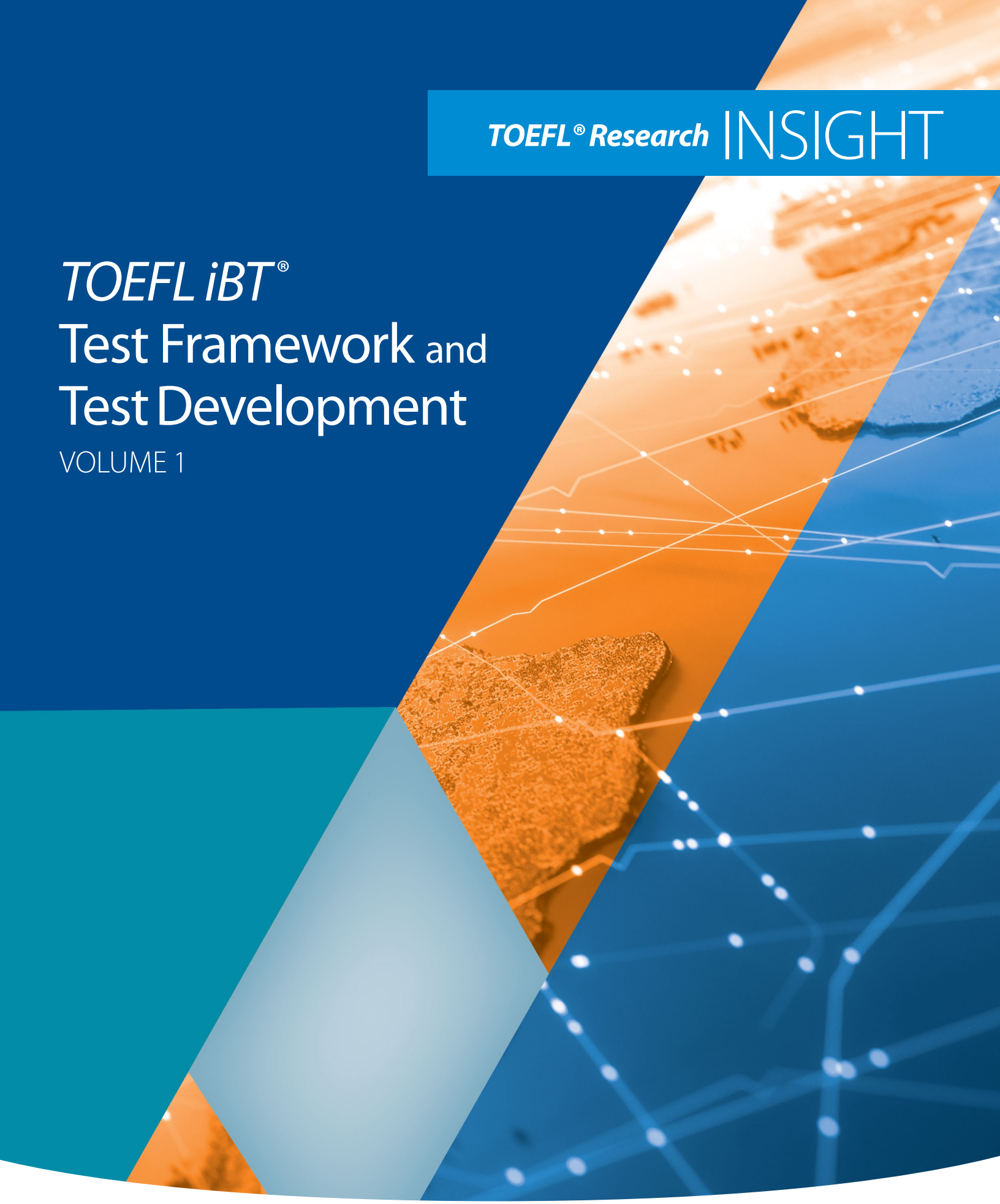
TOEFL® Research INSIGHT

TOEFL iBT® Test Framework and Test Development

VOLUME 1



ETS TOEFL

Quality Beyond Measure.

TOEFL® Research Insight Series, Volume 1: *TOEFL iBT*® Test Framework and Test Development

Preface

The *TOEFL iBT*® test is the world's most widely respected English language assessment and is used for admissions purposes in more than 130 countries, including Australia, Canada, New Zealand, the United Kingdom, 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 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 has been expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL Primary*® and *TOEFL Junior*® tests, which are designed to help teachers and learners of English in school settings. The *TOEFL ITP*® program offers colleges, universities, and others affordable tests for placement and progress monitoring within English 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 assure the quality of these scores. Through the publication of the *TOEFL Research Insight Series*, we wish to communicate to institutions and English teachers who use any/all of the TOEFL tests about the strong research and development base that underlies the TOEFL Family of Assessments and to 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. Beyond demonstrating this critical evidence of test quality, research is also important for enabling innovations in test design and ensuring that the needs of test takers and test score users are persistently met. This is why ETS has made the establishment of a strong research base a fundamental feature underlying the evolution of the TOEFL Family of Assessments.

The TOEFL Family of Assessments 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 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 related research projects are conducted by ETS's Research & Development staff each year and 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 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 include issues of core interest to test users, including how tests were designed, evidence for the reliability and validity of test scores, and research-based recommendations for best practices.

The close collaboration with TOEFL 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. Therefore, through this publication, we hope to foster an ever-stronger connection with our test users by sharing the rigorous measurement and research base and solid test development that continues to ensure the quality of the TOEFL Family of Assessments.

Dr. John Norris
Senior Research Director
English Language Learning and Assessment
Research & Development Division
Educational Testing Service (ETS)

The following individuals contributed to the second edition (2018) by providing careful reviews and revisions as well as editorial suggestions (in alphabetical order): Terry Axe, Ian Blood, Michelle Hampton, Susan Nissan, Eileen Tyson, Jennifer Wain, and Yuan Wang.

TOEFL iBT Test Framework and Test Development

The TOEFL iBT test design is the result of years of research—both investigation of the language-related knowledge, skills, and abilities (KSAs) that English language learners need to succeed in academic environments where English is the medium of instruction and research to identify the most effective methods of assessing these KSAs (described in Chapelle, Enright, & Jamieson, 2008). Leading experts from both inside and outside ETS in the fields of educational measurement, language testing, and language teaching contributed to the design of the TOEFL iBT test using an assessment design methodology known as *evidence-centered design* (ECD), originally developed at ETS by Mislevy, Steinberg, and Almond (2003) and now applied in a wide range of testing contexts across the globe. ECD is a process that requires explicit definitions of measurement claims and close examination and questioning of the strength of the evidence that supports them. As part of the ECD process, a team of ETS assessment specialists and statisticians reviewed a series of working papers defining the language use domains of the TOEFL iBT test along with evidence gathered through developmental research, resulting in the TOEFL iBT test framework (Pearlman, 2008). This framework established the test's format, structure, and content.

The TOEFL iBT Test Framework

Test Purpose

The purpose of the TOEFL iBT test is to evaluate the English proficiency of people whose native language is not English. TOEFL iBT scores are primarily used as a measure of the ability of international students to use English in an academic environment. To quote the original TOEFL working paper, the purpose of the test is "to measure the communicative language ability of people whose first language is not English . . . in situations and tasks reflective of university life" (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000, p. 10).

Test Structure

The TOEFL iBT test is administered via computer from a secure, worldwide, internet-based testing network. Some tasks on the test require the use of two or more language skills. Test takers wear noise-reducing headphones and speak into a microphone to record their responses to Speaking tasks and type their responses to Writing tasks. The spoken and written responses are digitally recorded and sent to the ETS online scoring network (for details, see *Scoring the Speaking and Writing Sections* below).

As Table 1 illustrates, each test form includes four sections: Reading, Listening, Speaking, and Writing. Each section is scored on a scale of 0–30, resulting in a total score of 120. The test takes about 4 hours to complete.

Table 1. The Structure of the TOEFL iBT Test

Section	Number of Items/Tasks	Testing Time	Score Scale
Reading	35–56 questions	60–80 minutes	0–30
Listening	34–51 questions	60–90 minutes	0–30
Break		10 minutes	
Speaking	6 tasks	20 minutes	0–30
Writing	2 tasks	50 minutes	0–30
Total		Approximately 4 hours	0–120

Test Content

Reading

The Reading section measures test takers' ability to understand university-level academic texts. TOEFL iBT test takers read three or four passages of approximately 700 words each and answer thirteen or fourteen questions about each passage. The passages represent a variety of academic areas and contain all of the information needed to answer the questions; they require no special background knowledge.

The questions are intended to assess the test takers' ability to comprehend factual information, infer information from the passage, understand vocabulary in context, and understand the author's purpose. Other types of questions assess the test taker's ability to recognize relationships among facts and ideas in different parts of a passage.

Listening

The Listening section measures test takers' ability to understand spoken English in an academic setting. Test takers listen to four to six lectures representing different academic areas, each about five minutes long, and listen to two or three conversations representing typical campus interactions with faculty, staff, and fellow students, each about three minutes long. Each listening passage is associated with a set of questions intended to assess test takers' ability to understand main ideas or important details, recognize a speaker's attitude or function, understand the organization of the information presented, understand relationships between the ideas presented, and make inferences or connections among pieces of information.

Speaking

The Speaking section measures test takers' ability to use spoken English effectively in educational environments, both inside and outside the classroom. There are six tasks in the Speaking section. Two "independent" tasks require the test taker to draw on personal experiences and opinions to answer. The other four tasks, referred to as "integrated" tasks, require the test taker to use information presented in a short spoken text or both a short spoken text and a related short written text.

Independent Speaking Tasks

These two tasks are designed to allow test takers to draw on their own personal experience and opinions. On one task, test takers describe or explain their personal experiences or opinion about a familiar topic. On the other task, they state and support an opinion from two opposing views.

Integrated Speaking Tasks

These four tasks assess integrated skills, requiring test takers to respond orally about things they listen to and read. The four types of integrated tasks are as follows:

- *Read/Listen/Speak (campus situation).* Test takers read a short passage communicating a typical campus situation or policy and then listen to a conversation in which a speaker expresses an opinion about the situation or policy. Test takers are then asked to give an oral summary of the speaker's opinion. A full response will require the test taker to combine and convey key information from both the Reading and the Listening tasks.

- *Read/Listen/Speak (academic course topic)*. Test takers read a passage that broadly defines a term, process, or idea from an academic subject. They then listen to a lecture that provides specific examples to illustrate the term, process, or idea expressed in the reading passage. Finally, they are asked to explain how the illustration presented in the lecture supports the broader concept defined in the reading. A full response will require test takers to combine and convey key information from both the Reading and the Listening tasks.
- *Listen/Speak (campus situation)*. Test takers listen to a conversation about a student's school-related problem and two possible solutions. Test takers must demonstrate understanding of the problem and orally express an opinion about the best way to solve it.
- *Listen/Speak (academic course topic)*. Test takers listen to an excerpt from a lecture that explains a term or concept (often by explaining two aspects or perspectives) and gives concrete examples to illustrate it. Test takers must then demonstrate understanding of the concept by providing a brief oral summary of the explanation and the related examples.

Writing

The Writing section measures test takers' ability to write in an academic environment and includes two tasks—one independent and one integrated.

Independent Writing Task

This task requires test takers to draw on their own knowledge and experience to write a short essay that states, explains, and supports their opinion on a specific issue.

Integrated Writing Task

In this task, test takers first read a passage on an academic topic. They then listen to part of a lecture that evaluates and criticizes the information and arguments presented in the reading. Finally, test takers must write a summary, in connected English prose, of the important points in the lecture, explaining how these points relate to those in the reading passage.

For both the Speaking and the Writing sections, test developers carefully design integrated tasks to ensure that a successful response will consider information from both the listening and reading materials.

Test Development Process

The development of a test form involves a complex series of steps and typically may take from 6 to 18 months. The steps in this process are designed to ensure that tests and items meet strict quality standards and that test forms are similar to each other in content and difficulty.

Content Development Staff

The TOEFL program recognizes the importance of using qualified staff to create test content for the TOEFL iBT test. All internal test development staff members, known as assessment specialists, have been trained in language learning or related subjects at the university level, and the majority of them have taught at schools, colleges, or universities internationally. Many TOEFL assessment specialists are themselves English language learners who have achieved graduate-level degrees from universities where English is the language of instruction. These ETS assessment specialists formulate the test stimuli (e.g., reading passages, lectures) and items (test questions) as the test takers eventually see them.

ETS carefully selects and trains outside item writers (who have experience teaching English as a second or foreign language or other academic content areas) to develop an initial draft of test questions. ETS considers item writers' experience and backgrounds so that the pool of item writers reflects, to the greatest degree possible, the diversity of the TOEFL iBT test's international test-taking population.

Item Writing

To ensure that test content is as comparable as possible from one TOEFL iBT administration to another, test developers follow detailed guidelines when selecting material for reading passages, lectures, and writing test questions. They consider whether the passages or lectures (and the questions based on them):

- are clear, coherent, at an appropriate level of difficulty, and culturally accessible;
- do not require background knowledge in order to be comprehensible;
- align with ETS fairness guidelines (discussed below); and
- contain sufficient testable content.

These considerations are fundamental to the TOEFL iBT test development process.

Item Review Process

ETS assessment specialists review test materials multiple times before using them in tests. Four or more assessment specialists sequentially review each stimulus and its associated items. They may suggest revising a stimulus or an associated item or rejecting an item or a stimulus entirely. Stimuli and items only become eligible for use in a test if all reviewers judge them to be acceptable. This linear peer-review process includes discussion between and among reviewers at each of three main stages: content review, fairness review, and editorial review. Additionally, when required for a given test stimulus or item, a subject matter expert checks the accuracy and currency of embedded information.

Content Review

At this stage, assessment specialists conduct multiple reviews of stimuli and items for both language and content, considering questions such as these:

- Is the language in the test materials clear? Is it accessible to a nonnative speaker of English who is preparing to study or is studying at a university where English is a medium of instruction?
- Is the content of the stimulus accessible to nonnative speakers who lack specialized knowledge in a given field (e.g., geology, business, or literature)?

For multiple-choice questions, reviewers also consider factors such as the following:

- the appropriateness of the point tested;
- the uniqueness of the answer or answers (the item keys);
- the clarity and accessibility of the language used; and
- the plausibility and attractiveness of *distracter* choices—the incorrect options.

For constructed-response items (Speaking, Writing) the process is similar but not identical. Reviewers tend to focus on accessibility, clarity in the language used, and on how well they believe the particular Speaking or Writing item will generate a fair and scorable response. It is also essential that reviewers judge each Speaking or Writing item to be comparable with others in terms of difficulty. Expert judgment, then, plays a major role in deciding whether a Speaking or Writing item is acceptable and can be included in an operational test.

Fairness Review

The *ETS Standards for Quality and Fairness* (ETS, 2014a) mandate fairness reviews. This fairness review must take place before using materials in a test.

All assessment specialists undergo fairness training—in addition to item writing training—soon after their arrival at ETS. As part of their training, item writers become familiar with the *ETS Guidelines for Fairness Review of Assessments* (ETS, 2016a) and the *ETS International Principles for Fairness Review of Assessments* (ETS, 2016b) and use them when developing and reviewing test stimuli and items. Fairness issues are thus considered at each stage of the development process.

In addition, specially trained and periodically calibrated fairness reviewers conduct a separate and independent review of all TOEFL test materials. TOEFL assessment specialists may not perform this official fairness review of TOEFL materials; the official fairness reviewer is typically an assessment specialist who works on other ETS tests. In this way, the fairness review is more objective and the reviewer brings no sense of ownership of the test to the review. When fairness reviewers find unacceptable content in the test materials, they issue a *fairness challenge*. The content reviewer assigned to the review step immediately after the fairness reviewer must resolve the challenge to the satisfaction of both reviewers. For rare cases in which the reviewers cannot reach agreement, a panel that includes the content and fairness reviewers adjudicates the issues at hand and comes to a resolution.

Editorial Review

All TOEFL test materials receive an editorial review. The purpose of this review is to ensure that language in the test materials (e.g., usage, punctuation, spelling, style, and format) is as clear, concise, and consistent as possible. Editors ensure that established ETS test style is followed. In addition, when warranted, editors check facts in stimuli for accuracy or to ensure that the stated facts are currently true; in areas such as physics or geography, for example, changes in facts occur periodically.

A Typical Test Review Chronology

The chronology of a typical review chain is: first content review, second content review, fairness review, editorial review, third content review, editorial review, and a final content review. Reviewers carefully analyze each set or item before signing off. A subsequent reviewer knows who the previous reviewer is and will usually consult with the previous reviewer on suggested changes to the set or item. Thus, the test development process for the TOEFL iBT test is collaborative.

Item Pretesting for Reading and Listening

As is true for other standardized tests, TOEFL iBT test items are pretested. Pretest items are included in operational forms and data are collected on real TOEFL test takers' ability to answer the items. Test takers cannot identify pretest items because they do not differ in any distinguishable way from the operational (scored) questions on the test. Pretesting items allows assessment specialists to identify poorly functioning items and revise them or exclude them from the operational item pool. Assessment specialists review data from item pretesting and use the information to refine their understanding of what makes a good test item.

Item Tryouts for Speaking and Writing

In operational administrations, the TOEFL iBT test's constructed-response sections do not contain embedded pretest items. Instead, both sections have small-scale tryout processes. ETS conducts tryouts of Speaking and Writing prompts (items) among members of the TOEFL test's target population. Assessment specialists review and evaluate spoken or written responses to these tryout questions. These specialists use expert judgment to determine which prompts are likely to elicit scorable responses from test takers across the range of proficiency levels; these *viable* prompts are the ones that appear in operational test forms.

Assembly of New Test Forms

After assessment specialists approve individual stimuli and associated test questions for use, and after the items have been successfully pretested (in the case of Reading and Listening items) or successfully tried out (in the case of Speaking and Writing items), the materials enter a database of items that are available for assembly into a test. Each TOEFL iBT test form is assembled and reviewed to ensure it meets the same content and statistical specifications as previous test forms. Each test form is comparable to other test forms so that test takers who take different test forms receive tasks that are similar in nature and in difficulty. This similarity, in turn, facilitates score equating, which is the statistical process used to calibrate the results of different forms of the same test.

Scoring the Speaking and Writing Sections

Scoring Guidelines

The scoring guidelines or rubrics for Speaking (ETS, 2014b) and Writing (ETS, 2014c) are the products of a careful, iterative development process. Many individuals with experience in evaluating the speaking and writing abilities of second-language learners contributed their expertise in developing the rubrics; among these individuals were English as a second language (ESL) and English as a foreign language (EFL) instructors, oral proficiency raters, applied linguists outside of ETS, and ETS assessment specialists. They employed a variety of methods in the rubric development process, including:

- having groups of experts order speaking or writing samples to identify features that differentiate performance at high, middle, and low levels of proficiency;
- investigating raters' decision-making processes to develop models of rater behaviors; and
- comparing holistic and analytical rating scales.

Rubric developers created 4-point rubrics for the Speaking section and 5-point rubrics for the Writing section; all of the rubrics are holistic, meaning that they require the rater to consider the overall quality of the response.

Scoring Processes

Constructed-response scoring presents challenges that multiple-choice testing does not. Assessment specialists and psychometricians—experts in the design and statistical quality of standardized tests—are fundamentally concerned with the difficulty of constructed-response items as well as raters' scoring consistency. ETS supports scoring quality for the TOEFL Speaking and Writing sections in a number of ways:

- The scoring process is centralized, and it is performed separately from the test center administration in order to ensure that test data are not compromised. Through centralized, separate scoring, each scoring step is closely monitored to ensure its security, fairness, and integrity.
- ETS uses its patented Online Network for Evaluation to distribute test takers' responses to raters, record ratings, and monitor rating quality constantly.
- Raters must be qualified. In general, they must be experienced teachers, ESL or EFL specialists, or in possession of other relevant experience. In addition to teaching experience, ETS prefers raters who have master's degrees and experience assessing spoken and written language.
- If they have the formal qualifications, raters are then trained. ETS trains raters using a web-based system. Following their training, raters must pass a certification test in order to be eligible to score. To assure reliability of constructed-response scoring, ETS monitors raters continuously as they score.
- Nonnative speakers of English may be raters, and, in fact, contribute a much needed perspective to the rater pool, but they must pass the same certification test as native-speaking raters.

At the beginning of each rating session, raters must pass a calibration test for the specific task type they will rate before they proceed to operational scoring. Scoring leaders—the scoring session supervisors—monitor raters in real time, throughout the day. These supervisors also regularly work as raters on different scoring shifts and are subject to the same monitoring. No rater, no matter how experienced, scores without supervision. ETS assessment specialists also monitor rating quality and communicate with scoring leaders during rating sessions.

For each administration, ETS's online scoring network sends Speaking and Writing responses to multiple independent raters for scoring. Each test taker's responses are scored by more than a single rater. The *e-rater*® automated scoring system (**https://www.ets.org/erater/about**) is a second rater on TOEFL independent and integrated Writing tasks. When a discrepancy between the human rater and the *e-rater* system arises, it is resolved by a second human rater.

Review of Items after Test Administration

After each TOEFL test administration, Reading and Listening items undergo a preliminary item analysis (PIA) to evaluate their performance in terms of their difficulty and how well they differentiate test takers of different ability levels. The PIA helps measurement specialists and assessment specialists to identify items that are too difficult or that fail to distinguish test takers of high and low proficiency in the skill being measured. Such problematic items are not scored. The PIA is thoroughly collaborative: Assessment specialists and psychometricians work together to make informed decisions about item performance and analysis. After the PIA, items go into an item pool with their accompanying statistics. For further information about statistical analysis of item performance, see Volume 3 of the *TOEFL Research Insight Series*, "Reliability and Comparability of TOEFL iBT Scores" (ETS, 2018).

Ongoing Oversight

Ongoing oversight is essential to the TOEFL program. As with all ETS tests, the TOEFL test undergoes an internal audit every 3 years. The auditors report directly to the ETS Board of Trustees.

The COE consists of twelve individuals from around the world, each of whom has achieved professional recognition in an academic field related to ESL or EFL. The COE provides guidance and oversight for research and development related to the TOEFL test.

The TOEFL Board consists of renowned professionals involved in international education, including admissions officers, graduate deans, international student advisors, and specialists in the fields of language testing, teaching, learning, and research. The TOEFL Board advises on the policies under which ETS administers the TOEFL test.

References

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

Educational Testing Service. (2014a). *ETS standards for quality and fairness.*

Educational Testing Service. (2014b). *TOEFL iBT scoring guides (rubrics) for speaking responses.*

Educational Testing Service. (2014c). *TOEFL iBT scoring guides (rubrics) for writing responses.*

Educational Testing Service. (2016a). *ETS guidelines for fairness of assessments.*

Educational Testing Service. (2016b). *ETS international principles for fairness of assessments.*

Educational Testing Service (2018). Reliability and comparability of TOEFL iBT scores. *TOEFL Research Insight Series* (Vol. 3, 2nd ed.).

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (1999). *TOEFL 2000 framework: A working paper* (TOEFL Monograph No. 16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–62. https://doi.org/10.1207/S15366359MEA0101_02

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.