TOEFL iBT™ Test Framework and Test Development

The TOEFL iBT™ test’s design is the result of years of research and development that involved experts in the fields of measurement and language teaching and testing. Test developers applied Evidence-Centered Design (ECD) in the test’s creation. ECD is a test design methodology, developed at ETS, that requires test developers to achieve consensus on an explicit definition of measurement claims and evidence. ECD, originally developed by Mislevy, Almond, & Lukas (2003), was adapted to finalize the design of the TOEFL iBT test (Pearlman, 2008). As part of the ECD process, a team of ETS assessment specialists and statisticians reviewed the original working papers and evidence gathered through developmental research (described in Chapelle, Enright, & Jamieson, 2008) to finalize the TOEFL iBT test framework. The 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. The 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® framework document (Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000, pp.10–11): “The purpose of the … test will be to measure the communicative language ability of people whose first language is not English … The test will measure examinees’ English-language proficiency in situations and tasks reflective of university life …” where instruction is conducted in English.

Testing Format

The TOEFL iBT test includes many technological and assessment innovations.

• It is administered via computer from a secure, international, Internet-testing network.

• It includes tasks that require test takers to integrate their use of two or more language skills when responding to a question.

• For speaking tasks, test takers wear noise-reducing headphones and speak into a microphone. Responses are digitally recorded and sent to ETS’s Online Scoring Network.

• For writing tasks, test takers type their responses; the typed responses are also sent to ETS’s Online Scoring Network.

• 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® (www.ets.org/erater/about) automated scoring system is a second rater on TOEFL independent and integrated writing tasks. A discrepancy between the human rater and the e-rater system would be resolved by a second human rater.

Test Structure

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. Detailed descriptions of the content of each section follow.

Table 1. The Structure of the TOEFL iBT Test

<table>
<thead>
<tr>
<th>Section</th>
<th>Number of Items/Tasks</th>
<th>Testing Time</th>
<th>Score Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>36–70</td>
<td>60–100 minutes</td>
<td>0–30</td>
</tr>
<tr>
<td>Listening</td>
<td>34–51</td>
<td>60–90 minutes</td>
<td>0–30</td>
</tr>
<tr>
<td>Break</td>
<td></td>
<td>10 minutes</td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td>6 tasks</td>
<td>20 minutes</td>
<td>0–30</td>
</tr>
<tr>
<td>Writing</td>
<td>2 tasks</td>
<td>50 minutes</td>
<td>0–30</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>Approximately 4 hours</td>
<td>0–120</td>
</tr>
</tbody>
</table>

Test Content

Reading

The Reading section measures test takers’ ability to understand university-level academic texts. TOEFL test takers read 3–5 passages of approximately 700 words each and answer 12–14 questions about each passage. The passages contain all of the information needed to answer the questions; they require no special background knowledge.

The questions are mainly intended to assess the test taker’s ability to comprehend factual information, infer information from the passage, understand vocabulary in context, and understand the author’s purpose. These questions are multiple-choice questions with a single correct answer. Other types of questions are to assess the test taker’s ability to recognize relationships among facts and ideas in different parts of a passage. These questions have more than four choices and more than one answer, allowing for partial-credit scores.
Listening

The Listening section measures test takers’ ability to understand spoken English in an academic setting. Test takers listen to 4–6 lectures, each 3–5 minutes long, and listen to 2–3 conversations, each about 3 minutes long. The questions are 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. Listening questions are mostly multiple-choice questions with a single correct answer, and some questions have more than one answer, allowing for partial-credit scores.

Speaking

The Speaking section measures test takers’ ability to speak English effectively in educational environments, both inside and outside of the classroom. The Speaking section consists of six tasks: Two of these tasks are independent; that is, test takers receive no oral or written test materials. On this task, test takers respond to a relatively general question on a familiar topic. The other four tasks assess integrated skills. On two of these tasks, test takers respond to both an oral and a written stimulus; in the other two integrated tasks, they respond to an oral stimulus. The tasks follow this format:

**Independent Speaking Tasks**

For these two questions, test materials are designed so as not to constrain examinee responses. On one task, test takers respond to a question concerning a personal preference. On the other task, they answer a question that asks them to make a choice.

**Four Integrated Speaking Tasks**

These tasks assess integrated skills, requiring test takers to respond orally both to oral and to written stimuli. The types of integrated tasks are as follows:

- **Read/Listen/Speak (Campus situation).** Test takers read a passage, listen to a speaker express an opinion about the passage topic, and then give an oral summary of the speaker’s opinion.
- **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. Test takers then respond orally, combining and conveying important information from both the reading passage and the lecture.
- **Listen/Speak (Campus situation).** Test takers listen to a conversation about a student-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 and gives concrete examples to illustrate it. Test takers must then orally summarize the lecture and demonstrate their understanding of how the examples relate to the overall topic.

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**

For this task, test takers receive no oral or written stimulus materials; instead, they respond to a relatively general question that allows them to tap their own knowledge and experience. They must write an essay
that states, explains, and supports their opinion on an issue, and must develop support for their opinions rather than simply listing personal preferences or choices.

**Integrated Writing Task**

On this task, test takers first read a passage; then they listen to a lecture that takes a position that is somehow different from the position presented in the reading passage. Test takers must then, in connected English prose, write a summary of the important points in the lecture and explain 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 test development process involves a complex series of steps, which typically may take from 6 to 18 months, to ensure that tests and items meet strict quality standards and that test forms are similar to each other in content and difficulty.

**Selection of Item Writers**

The TOEFL program recognizes the importance of using qualified item writers. All internal test development staff members, known as assessment specialists, receive training in language learning or other related subjects at the university level, and the majority of them have taught at schools, colleges, or universities abroad. These ETS assessment specialists formulate the test stimuli (e.g., passages, lectures) and items (test questions) as the test takers eventually see them. ETS carefully selects and trains outside item writers who have experience in teaching English as a second or foreign language to conduct the initial development of test materials before the assessment specialists refine them. To the greatest degree possible, ETS considers item writers’ experience and backgrounds so that the pool of item writers reflects the diversity of the TOEFL test’s international test-taking population.

**Item Writing**

Test question writers follow detailed guidelines when selecting material for reading passages and lectures and writing test questions, with the goal of making test questions as comparable as possible from one administration to another. Writers consider whether the passages or lectures (and the questions based on them):

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

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

**Item Review Process**

ETS assessment specialists review test materials a number of 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 may suggest 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 stage. At each stage, reviewers must sign off before moving the stimuli and items to the next stage. The TOEFL iBT test review process has three main stages: a content review, a fairness review, and an editorial review that focuses on both content and formatting. Additionally, when required, a subject matter expert checks the accuracy and currency of the content.
Content Review

At this stage, assessment specialists review 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
- 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, lack of ambiguity 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.

This peer review process is linear; reviewers move all test materials through predetermined stages. Test materials move to the next stage of review only after a reviewer signs off, signifying approval.

Fairness Review

ETS Standards for Quality and Fairness (2002) mandates fairness reviews. This fairness review must take place before using materials in a test.

Because attention to fairness is such an integral part of the test design, all assessment specialists undergo fairness training — in addition to item writing training — relatively soon after their arrival at ETS. As part of their training to develop TOEFL test materials, item writers must become familiar with the ETS Guidelines for Fairness Review of Assessments (2009) and the ETS International Principles for Fairness Review of Assessments (2007) and use them when reviewing items and stimuli. The content review process itself, therefore, always includes fairness as an aspect of development.

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 may be 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 into the review. When fairness reviewers find unacceptable content in the test materials, they issue fairness challenges. The content reviewer assigned to immediately follow the fairness reviewer must resolve the challenge to the satisfaction of both reviewers. For rare cases in which the reviewers cannot reach agreement, there is a process in place known as fairness adjudication, in which a panel that includes the content and fairness reviewers adjudicates the issues at hand and comes to a resolution.

Validity concerns underlie all aspects of fairness review. To ensure the validity of a test, it is paramount that only construct-relevant factors affect test takers’ scores. The construct can be defined as all of the knowledge, skills, and abilities that a test is supposed to measure. A primary goal of fairness review, then, is to identify and reduce construct-irrelevant aspects of stimuli or
items that might hinder test-taker performance or even, in construct-irrelevant ways, enhance test-taker performance. Minimizing the influence of construct-irrelevant test content enhances fairness and thus also the validity of test scores.

Editorial Review

All TOEFL test materials receive an editorial review. This review’s purpose 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 must 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.

Assembly of New Test Forms

After assessment specialists approve individual stimuli and associated test questions for use, the materials are assembled into a test. Each TOEFL iBT test form must meet the same content and statistical specification as previous test forms. It is essential that each test form be similar to other test forms so that test takers taking 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.

In assembling an operational form, assessment specialists consider a number of complex variables, including topic variety among test questions, the number of questions of a particular type, and the difficulty of the questions; test assemblers work according to test assembly specifications that describe the parameters for these variables; this ensures content comparability and supports statistical comparability across different forms of the test.

Item Pretesting and Tryout

Item Pretesting for Reading and Listening

As is true on other standardized tests, some TOEFL iBT test items are pretested. 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 early in the development cycle and remove them from the operational item pool. Pretested items that perform poorly are never included in scored sections, so the scored sections can be shorter and more efficient without sacrificing quality. This approach has an additional benefit: By reviewing data on item performance, assessment specialists calibrate their understanding of what may make a good test item.

Item Tryout 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 will appear in operational test forms.
Scoring for Speaking and Writing Sections

Scoring Guidelines

The scoring guidelines or rubrics for Speaking (ETS, 2004a) and Writing (ETS, 2004b) were 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 ESL and 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 rank-order speaking or writing samples and identify features that differentiated 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.

Scoring Processes

Constructed-response scoring presents challenges that multiple-choice testing does not. Assessment specialists and psychometricians 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 Scoring Network (OSN) to distribute test takers’ responses to raters, to record ratings, and to constantly monitor rating quality.
- Raters must be qualified. In general, they must be experienced teachers and specialists in the field of English or teaching English as a second language. In addition to teaching experience, ETS prefers raters who have master’s degrees and experience assessing spoken and written language.
- Nonnative speakers of English may be raters, but they must submit TOEFL iBT or TSE® (Test of Spoken English™) scores when they apply.
- 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.

At the beginning of each rating session, raters must pass a calibration test before they proceed to operational scoring. Scoring leaders — the scoring session supervisors — have the capabilities to monitor raters in real time, throughout the day. These supervisors also work as raters 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.

Review of Items after Test Administration

After each TOEFL test administration, each test form undergoes a preliminary item analysis (PIA) to evaluate the performance of items in terms of their difficulty levels 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 process 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 “Use of Item Analysis in Test Development.”

**Ongoing Oversight**

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

The TOEFL Committee of Examiners (COE) consists of twelve individuals from around the world, each of whom has achieved professional recognition in an academic field related to English as a Foreign or Second Language. The COE provides guidance and oversight for research and development related to the TOEFL test. On other matters, such as major test redesign and policy changes, the COE has power of recommendation to the TOEFL Board.

The TOEFL Board consists of sixteen renowned professionals involved in international education, including admissions officers, graduate deans, foreign student advisors, and specialists in the fields of language testing, teaching, learning, and research. The TOEFL Board advises the TOEFL program on policy issues.

**References**


ETS, 2004 (a). TOEFL iBT Scoring Guides (Rubrics) for Speaking Responses.  

ETS, 2004 (b). TOEFL iBT Scoring Guides (Rubrics) for Writing Responses.  


