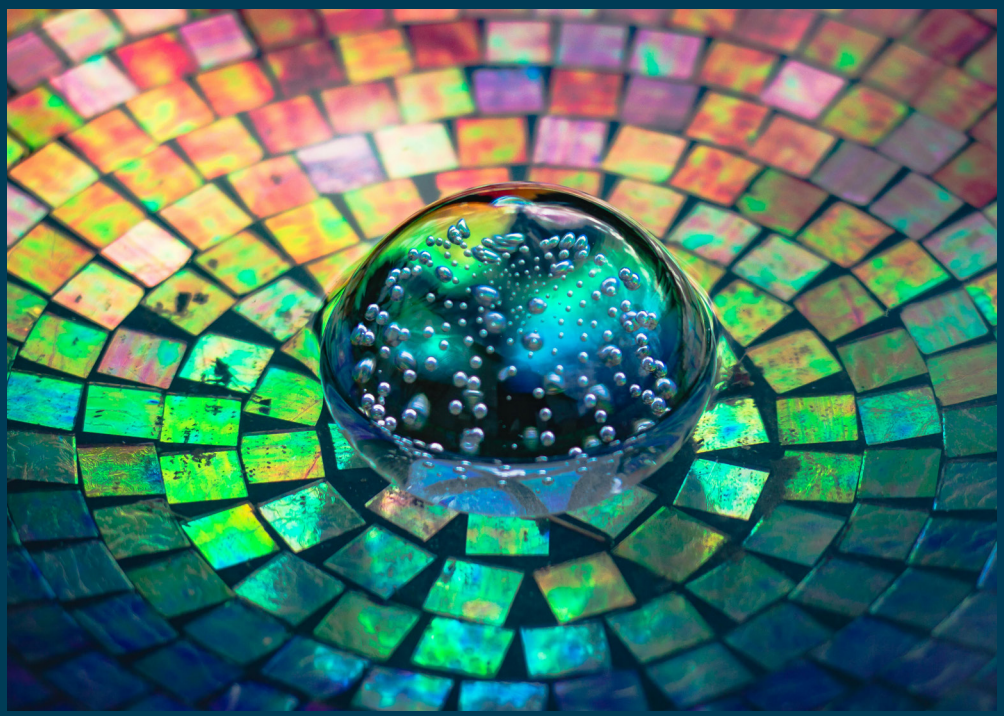
*ets research institute

OCTOBER 2025

TOEFL® RESEARCH SERIES

TOEFL iBT® Technical Manual



Venessa F. Manna, Shuhong Li, Spiros Papageorgiou, and Lixiong Gu

TOEFL–RR-106

ETS Research Report No. RR–25-12

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

TOEFL iBT® Technical Manual

Venessa F. Manna[1], Shuhong Li[1], Spiros Papageorgiou[2], Lixiong Gu[1]

[1] ETS, Princeton, New Jersey, United States

[2] ETS Research Institute, ETS, Princeton, New Jersey, United States

Abstract

This technical manual describes the purpose and intended uses of the TOEFL iBT test, its target test-taker population, and relevant language use domains. The test design and scoring procedures are presented first, followed by a research agenda intended to support the interpretation and use of test scores. Given the updates to the test starting January 2026, this technical manual is intended to serve as an overview and rationale for the test design as well as a reference point for informing investigations of validity evidence to support the intended test uses over time. Designed as a living document, this manual will be updated as the test's design, administration, scoring, and evidence of measurement quality (including reliability, validity, and fairness) evolve, along with its intended uses.

Keywords: English language proficiency, test tasks, Test design, TOEFL iBT®, validity, fairness

Corresponding author: L. Gu Email: lgu@ets.org

Contents

Section I. Introduction

I-1. Test Purpose and Intended Users

The TOEFL iBT® test, hereinafter referred to as "TOEFL," measures foundational language skills and communication abilities needed in academic and daily life settings. The test evaluates the four language skills of reading, listening, writing, and speaking and is intended to offer academic institutions and other score users reliable insights into a test taker's English language ability.

Since its 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. In its current iteration, the TOEFL test is designed for efficient measurement of both foundational aspects of language proficiency (lexical and grammatical competence) as well as the ability of language learners to communicate in English through a range of language knowledge activities and communicative language tasks. These activities and tasks are drawn from both academic and daily life contexts, and they provide test takers with brief but authentic opportunities to demonstrate their language skills. Some examples of communicative language tasks represented in the test include

- reading passages from academic and daily life sources, such as textbooks, newspapers and magazines, websites, and social media;
- listening to academic talks and lectures, public announcements, and personal interactions;
- writing responses for common situations such as emails and academic online discussions; and
- speaking to a simulated interviewer, or fluently and intelligibly retelling spoken input.

The TOEFL test is designed to optimize both convenience and quality. It can be taken either in a test center or at home, and official test scores are available in 72 hours. Test security during the administration of the test is provided by a combination of trained human proctors and artificial intelligence (AI). AI technology monitors activity and settings on the test taker's computer and sends alerts to proctors about unusual behavior or room conditions. A variety of security measures before and after the administration of the test are also used to minimize content exposure and detect misconduct.

This technical manual adheres to the professional guidelines outlined in the Standards for Educational and Psychological Testing (AERA et al., 2014) and the ETS Standards for Quality and Fairness (ETS, 2014). These guidelines represent the consensus of measurement professionals and reflect ETS's commitment to these standards.

The purpose and intended uses of the test, its target test-taker population, and relevant language use domains are described first. The test design and scoring procedures are presented next, followed by a research agenda intended to support the interpretation and use of test scores. This technical manual is intended to serve as an overview and rationale for the test design as well as a reference point for informing investigations of validity evidence to support the intended test uses over time. Designed as a living document, the manual will be updated as the test's design, administration, scoring, and evidence of measurement quality (including reliability, validity, and fairness) evolve, along with its intended uses.

I-2. Target Population, Language Domains, and Intended Uses

The TOEFL test is intended for older adolescents and adults who wish to provide evidence of their overall English language proficiency level in academic and daily life settings. The multistage adaptive test (MST) methodology of the test, explained in more detail later, helps to ensure accurate and efficient measurement of the test taker's language ability by matching the difficulty of the test tasks with the proficiency level of the test taker. Using MST methodology, the TOEFL test is suitable for language learners with a wide range of proficiency levels. In terms of proficiency levels described in the Common European Framework of Reference (CEFR; Council of Europe, 2001, 2020), the TOEFL test is designed to cover the full range from A1 to C2 (see Section III: Scoring and Score Reporting).

The CEFR defines four domains in which communicative language activities take place: public, personal, occupational, and educational. The public domain refers to language activities as part of ordinary social interaction, including business and public services and leisure activities. The personal domain focuses on the immediate family environment and the individual. The occupational domain refers to activities related to one's professional life. The educational domain is concerned with contexts where people learn or receive training. The TOEFL test is designed to efficiently measure foundational language skills and general communication abilities relevant to academic and general (daily life) contexts. These contexts coincide with domains described in the CEFR, with emphasis on the educational and public domains.

Extensive market research was conducted by ETS in late 2020 and early 2021, with nearly 250 score users from institutions in the United States, Canada, and the United Kingdom and 7,200 test takers of TOEFL iBT and other English language tests around the world. The market research identified a need for a language proficiency test that is affordable and convenient to access. In response, the TOEFL test provides academic programs and other scores users with valid and reliable information about an individual's English proficiency. It offers a relatively brief test-taking experience, using a format that is both test-taker friendly and engaging. Recommended uses of the TOEFL test include

- to inform decisions about the English language proficiency of international students who apply for admission into higher education institutions and international high schools;
- to inform decisions about students' placement in, progress through, and exit from English language proficiency classes or English pathway programs;
- and to inform other decisions where an overall indication of English language proficiency is required.

Section II. Test Constructs, Design, and Content Development

II-1. Construct Definition

Considering the intended uses and administration requirements for the TOEFL test outlined in the previous section, the construct that guided assessment task development and test design reflected the following dimensions. Overall, the test measures both (a) selected foundational skills underlying English learners' proficiency and (b) the ability to communicate effectively in listening, reading, writing, and speaking tasks in English language academic and daily life communication contexts. This construct is, therefore, a hybrid combination of foundational aspects of English language competence—and associated cognitive capacities—and contextualized higher order communicative abilities (Hulstijn, 2015; Norris & Ortega, 2012; Xi & Norris, 2021).

On the one hand, foundational aspects of second language (L2) competence are generalizable (i.e., they apply across contexts of language use) and useful for differentiating the overall English language proficiency levels typical of adolescent and adult learners. This dimension of the construct emphasizes skills that underlie, and also predict, other communicative aspects of language proficiency. Importantly, rather than attempting to measure comprehensively all of the many foundational skills that constitute L2 competence (e.g., Bachman & Palmer, 2010), the TOEFL test focuses on a handful of these skills that are highly predictive of global language proficiency. The test thus measures aspects of English language vocabulary knowledge, which has been shown to predict language proficiency in general (Qian & Lin, 2020) and reading ability in particular (Qian, 2002). The test also measures knowledge of English language syntax and associated word order rules, a useful predictor of overall L2 proficiency (Norris, 2005) and writing ability (Crossley et al., 2014). Additionally, the test measures the ability to process aural and written English input for both semantic meanings and linguistic forms and to reproduce the input with accuracy and fluency. These phenomena, too, provide strong predictions of general L2 proficiency (Yan et al., 2016) and speaking ability in particular (Van Moere, 2012). Test tasks associated with this dimension of the construct are designed to efficiently predict global L2 English proficiency across the full spectrum of the CEFR proficiency levels.

On the other hand, a second construct dimension addresses test takers' abilities to engage in higher order communication tasks that call upon contextualized listening, reading, writing, and speaking. This dimension of the construct emphasizes how learners marshal their linguistic

competencies and apply them to solve a range of communication challenges that represent English as it is used in academic and daily life contexts. This task-based dimension of the construct is essential for informing interpretations about test takers' abilities to use English effectively and authentically (Norris, 2018). The Reading section measures the ability to read and comprehend information presented in a variety of formats, including short informational graphics as well as extended passages. The Listening section measures the ability to listen to and comprehend both conversational and extended monologic (e.g., lecture) speech. The Writing section measures the ability to write effectively in common genres such as writing an email and responding to an academic discussion. The Speaking section also measures the ability to speak spontaneously and meaningfully in response to questions in an interview format. Test tasks associated with this dimension of the construct are designed to situate learners in real-life settings that require specific types of receptive and productive language performance.

This hybrid approach to construct definition, which covers both selected foundational aspects of L2 competence and task-based communicative language ability, is operationalized through a test design that can efficiently level a test taker's global proficiency (i.e., through the foundational dimension of the construct) while simultaneously probing their communicative competence in relevant performance situations (i.e., through the task-based dimension of the construct). Construct operationalization for the TOEFL test focuses on predicting overall English ability and discerning the likelihood that learners can accomplish real-life English communication tasks.

II-2. Test Design Process

ETS brings over 60 years of experience in developing and administering English language assessments and more than 20 years in designing tasks that utilize automated scoring technology. Leveraging this expertise, the TOEFL test was developed through a collaborative effort involving researchers, test developers, and psychometricians. The design team worked closely with ETS business directors to establish requirements ensuring the assessment meets the needs of score users, English language learners, and other stakeholders.

Key requirements for the test design included the following:

- Measuring and reporting scores for all four language skills: reading, listening, writing, and speaking

- Assessing a wide range of abilities, from novice to advanced users of English (CEFR A1 to C2)
- Measuring language ability in academic and general (daily life) contexts
- Offering contexts that reflect use of the English language beyond North American contexts.
- Using the same reporting score scale for all four language skills
- Completing automated scoring and score report delivery within 72 hours

With these requirements in mind, the team adopted a principled approach to test design, which involved evaluating an extensive catalog of assessment tasks for appropriateness and drafting an initial blueprint.

The design of the test reflected the need to combine test-taker convenience and efficiency with trustworthy measurement of language ability across a broad range of proficiency levels and yet be relevant to a wide range of language use contexts (Davis, Norris, et al., 2023). The test was designed to balance these demands by employing MST, an efficient test administration model, and by combining task types that address both foundational language abilities and communication skills. Tasks measuring foundational abilities, such as providing missing letters of words or the ability to repeat sentences that one hears, were selected to provide rapid and reliable information regarding general language proficiency. These tasks were then integrated with tasks that require the test taker to understand spoken or written input or produce spoken or written responses. The integration of these task types represents the hybrid approach to construct operationalization mentioned in the previous section, which is intended to quickly determine a test taker's general level of language proficiency as well as provide information regarding the ability to use English to communicate.

The designers of the TOEFL test selected all the questions from a previously conducted prototyping study of writing and speaking questions and a pilot study, as well as a field study, of reading, listening, writing, and speaking tasks, which led to the development of the TOEFL Essentials® test (Papageorgiou et al., 2021).

The prototyping study initially focused on iterative development of concept demos illustrating tasks that were specifically designed to collect evidence of ability in a brief period of time; these demos were then presented to an advisory panel of university language program administrators who gave their reactions regarding the usefulness of the tasks for measuring

language ability. This step was followed by development of working prototypes of writing and speaking tasks, which were trialed with language learners over several iterations to evaluate the usability of different design features and confirm that useful evidence of ability was elicited. Once the general design of the speaking and writing tasks had been confirmed, a large-scale prototyping study was conducted where these new task types were administered to an international sample of English learners ($N = 570$). After the prototype tasks were administered and responses were evaluated, scoring criteria were developed for each task based on expected response features as well as review of responses collected. At this stage, several task types were dropped from further consideration due to challenges in delivery and/or scoring, and design features of the remaining tasks were refined as needed.

Next, a pilot administration was conducted, incorporating the refined speaking and writing tasks as well as listening and reading tasks adapted for efficient language proficiency assessment. The pilot administration included a population of English learners from diverse regions of the world ($N = 700$). Both the prototype and pilot administrations included more task types than were needed for the final test design. Based on the pilot results, a subset of the highest performing task types was selected for the operational test design and their specifications were further refined.

The final step in operational test design involved field testing of a pool of questions on a population similar to the expected operational test-taker population and large enough to produce stable question statistics ($N \approx 5{,}000$). The field test-taker population covered the full spectrum of CEFR levels.

A core design principle of the TOEFL test is that assessment tasks, scoring guides, and delivery systems should support fairness and equity by providing all test takers the needed opportunities to demonstrate their English language proficiency. As a first step, at-home delivery is expected to increase access to the test compared to traditional test delivery limited to test centers. At the same time, test takers have the option of test centers, for example when they do not wish to deal with setting up their own computer for at home testing, or finding a room appropriate for test administration. Additionally, the test developers used MST design with the intention to present each test taker with test tasks that are appropriate for their proficiency level so they have the best opportunity to demonstrate their ability. Finally, empirical analyses were

conducted during pilot and field testing to confirm absence of bias toward specific test-taker groups identified on the basis of gender and first language.

Preliminary analysis by psychometricians and researchers determined the number of questions per task type needed to facilitate reliable test scores. For the Writing and Speaking sections, rubrics and AI scoring models were evaluated and refined.

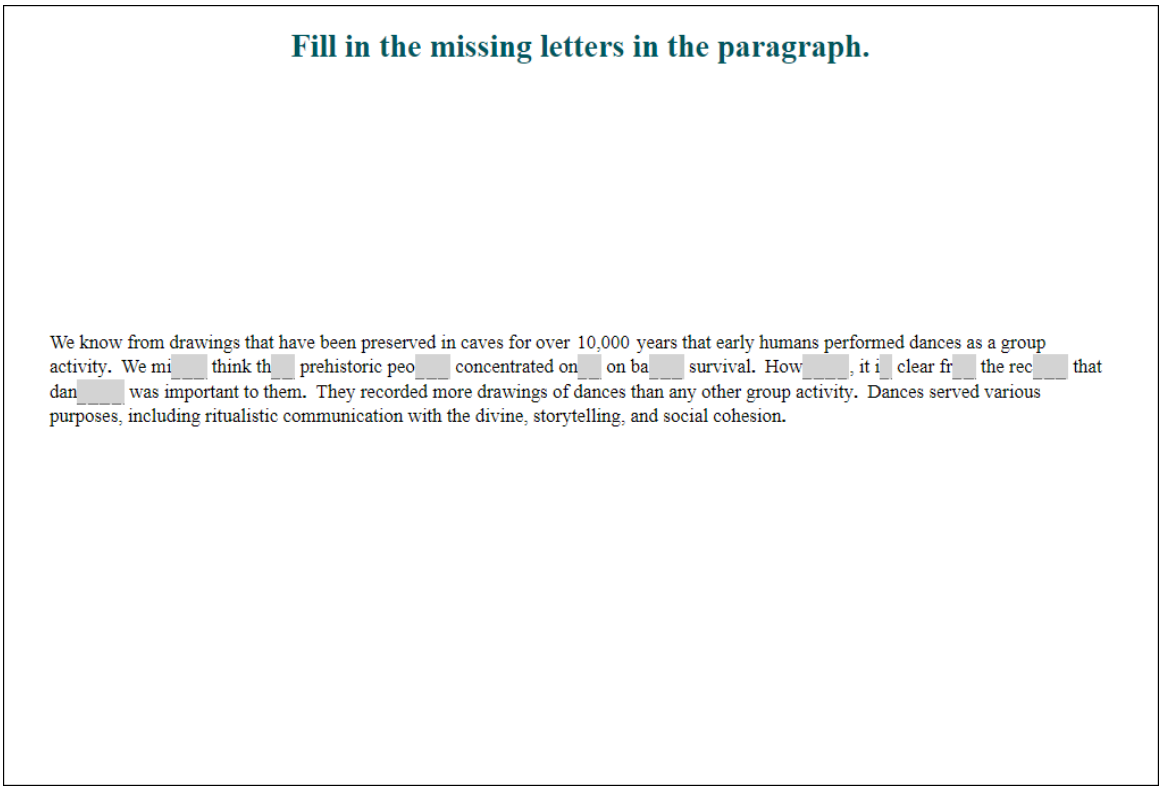
II-3. TOEFL Reading Task Types

People worldwide learn from academic texts and other academic materials in English. In their daily lives, people also need to navigate a wide range of reading material they encounter, from concise information like receipts, schedules, signs, and menus to more expanded informal texts such as webpages, news and magazine articles, emails, and text messages. The reading questions assess a test taker's ability to comprehend both academic and nonacademic texts from various English-speaking contexts. Reading skills are measured with the following task types: *Complete the Words* (C-test), *Read in Daily Life*, and *Read an Academic Passage.*

Complete the Words (C-test)

Reading—or more precisely, the ability to process written texts for meaning and form—is tested on the basis of the C-test format (see Figure 1). The C-test presents test takers with paragraph-length texts drawn from authentic sources. Following an intact first sentence, the second half of every second word is deleted, and the examinee must provide the missing letters. Each text contains 10 truncated words. Each text is a passage that presents a coherent and self-contained meaning unit. In other words, text meaning should not depend on information contained in other preceding or following passages. Texts are sampled and adapted from authentic, first-language sources. Texts should reflect common, widely accessible topics that are not highly specialized, do not rely on technical vocabulary or jargon, and do not feature excessive use of proper nouns. Texts should be based on standard, grammatically accurate written English, and not on hybridized forms of written communication (e.g., chat) or transcribed/reported dialogue.

Figure 1. Example of *Complete the Words* Task Type



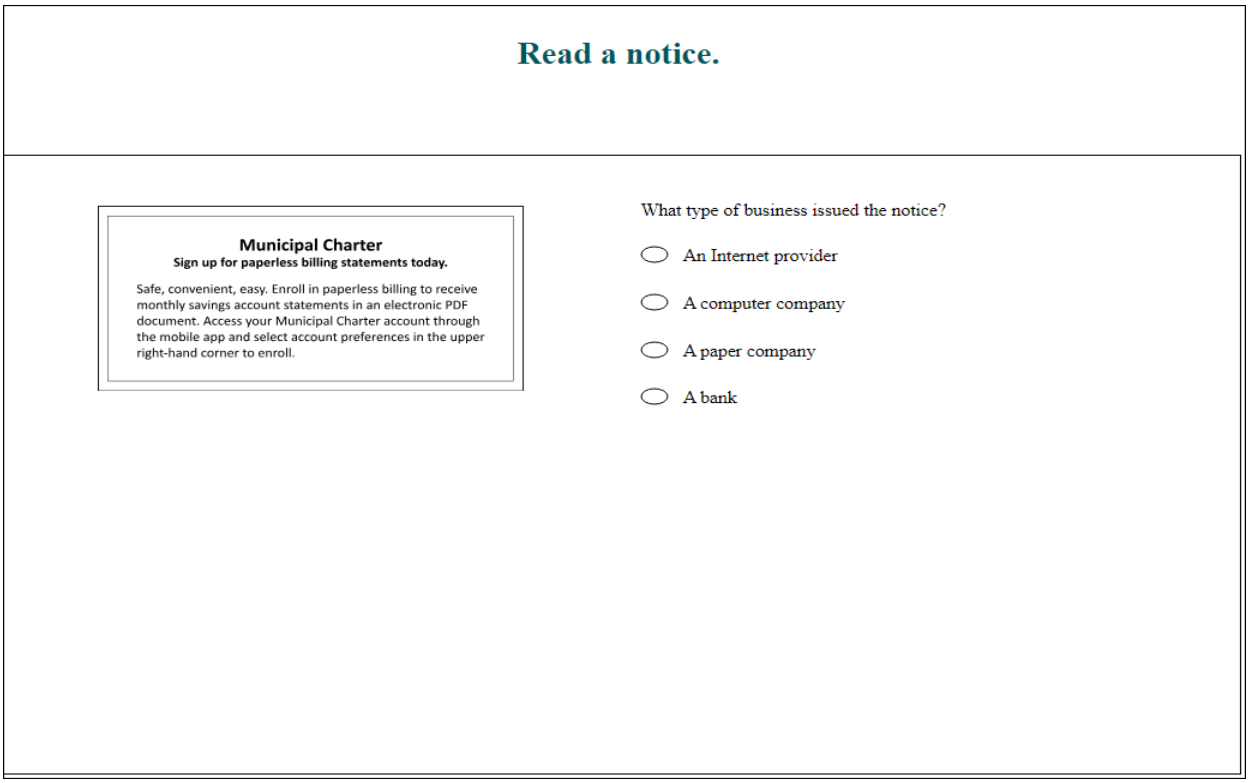
Source: TOEFL iBT® test, ETS

Read in Daily Life

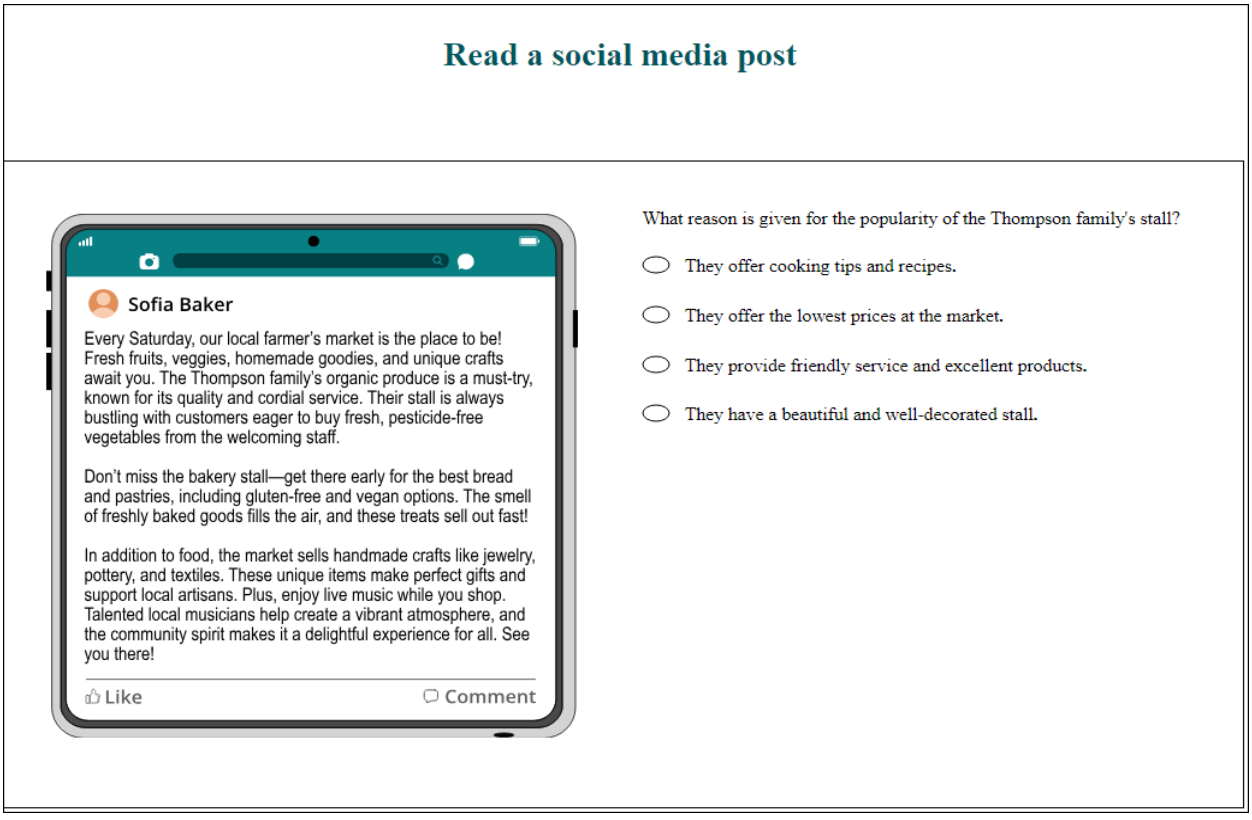
The *Read in Daily Life* task includes short, nonacademic texts commonly encountered in daily life around the world (see Figure 2). Examples of texts include a poster, sign, or notice; menu; social media post or webpage; schedule; email; chain of text messages; advertisements; news article; form; invoice; or receipt. The texts can be anywhere from 15 to 150 words and include two or three multiple-choice questions depending on the length of the text. The questions require test takers to

- understand information in common, nonlinear text formats;
- identify the main purpose of a written communication;
- understand informal language, including common idiomatic expressions;
- make inferences based on text;
- understand telegraphic language; and
- skim and scan for information.

Figure 2. Examples of *Read in Daily Life* Task Type



Source: TOEFL iBT® test, ETS



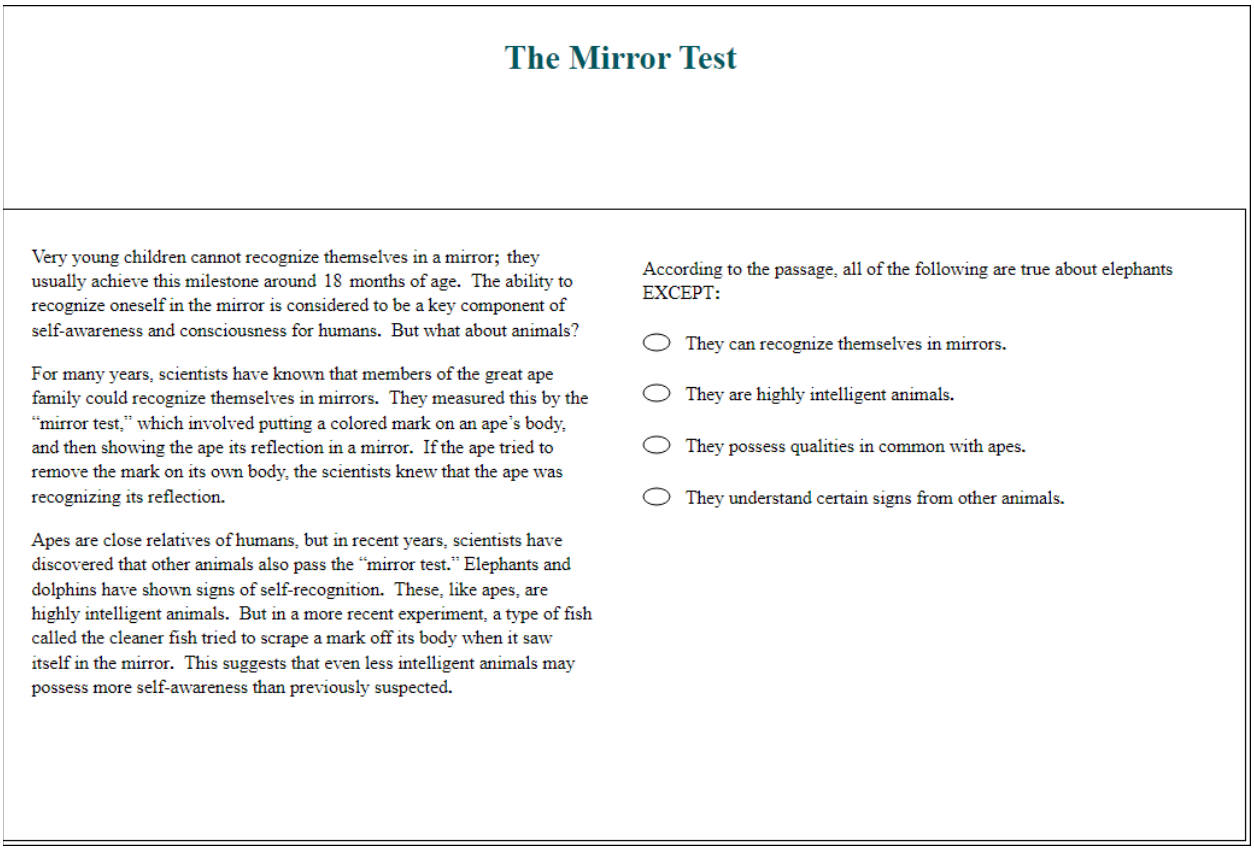
Source: TOEFL iBT® test, ETS

Read an Academic Passage

The *Read an Academic* Passage task includes short expository passages typical of those in secondary and higher education (see Figure 3). The task is designed so that background knowledge is not required. The passages cover topics drawn from subject areas such as history, art and music, business and economics, life science, physical science, and social science. The texts are approximately 200 words and are typically followed by five questions that may ask about factual information, vocabulary in context, inferences, relationships between ideas, and the purpose of part or all of the text. The questions require test takers to

- identify the main ideas and basic context of a short, linear text;
- understand the important details in a short text;
- understand the range of grammatical structures used by academic writers;
- infer meaning from information that is not explicitly stated;
- understand a broad range of academic vocabulary;
- understand a range of figurative and idiomatic expressions;
- understand ideas expressed with grammatical complexity;
- understand the relationship between ideas across sentences and paragraphs; and
- recognize the rhetorical structure of all or part of a written text.

Figure 3. Example of *Read an Academic Passage* Task Type



Source: TOEFL iBT® test, ETS

II-4. TOEFL Listening Task Types

People around the world use English for daily life listening activities and may also need to understand orally delivered academic subjects in English. Input in such listening activities is encountered in both monologic and dialogic format. The questions in the Listening section measure the test taker's ability to understand conversations and talks set in academic and daily life contexts. The speakers in the tasks have accents from four regions of the world: North America, the United Kingdom, Australia, and New Zealand. Listening skills are measured with the following task types: *Listen and Choose a Response, Listen to a Conversation, Listen to an Announcement,* and *Listen to an Academic Talk.*

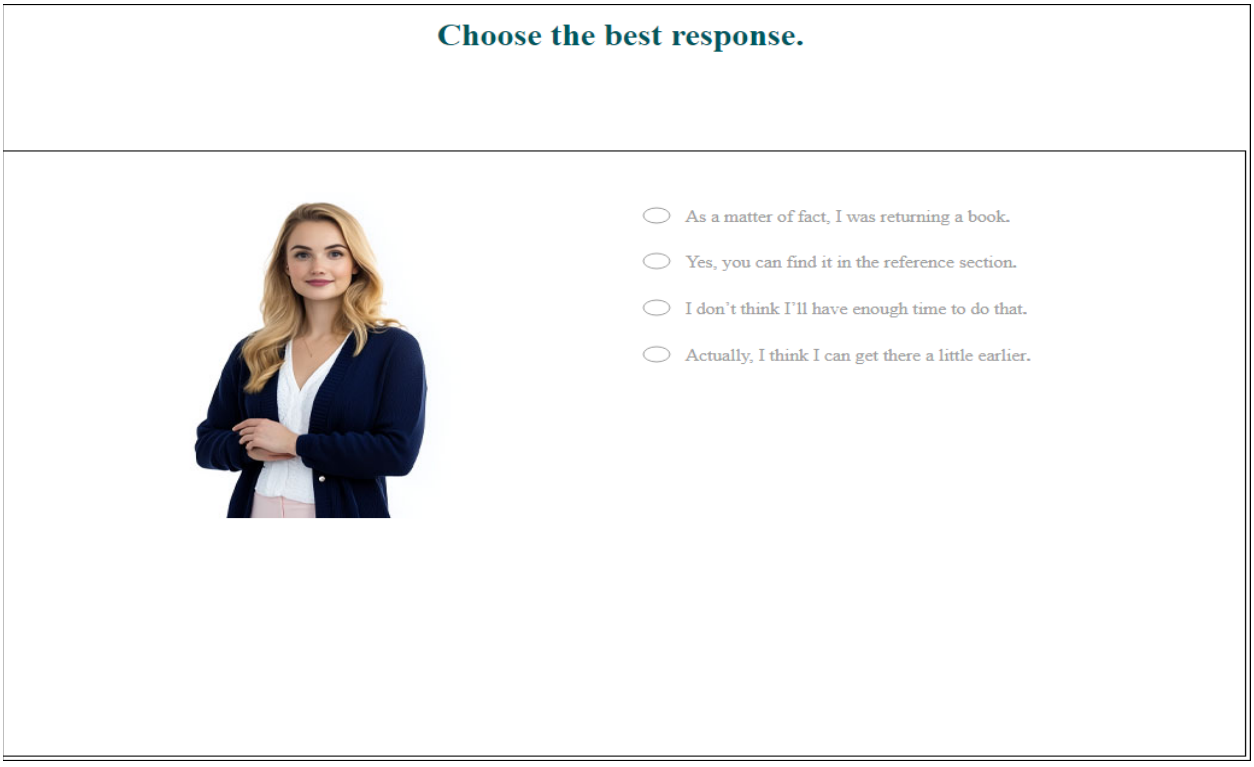
Listen and Choose a Response

The *Listen and Choose a Response* task is designed to measure the test taker's ability to understand a short, spoken question or statement and recognize an appropriate response in short

dialogues on topics related to everyday life. Selecting the appropriate response requires understanding both the literal and implied meaning of the speaker, a skill that is important for social interactions. The test taker hears a question or statement, which forms the first part of a short exchange between two speakers (see Figure 4). The question or statement is only heard, and it is not written on the screen. The test taker then reads four possible responses to the question or statement. The test taker must select the most appropriate response to the first speaker's question or statement. Test questions require test takers to

- understand common vocabulary and formulaic phrases;
- understand simple grammatical structures, including question-formation patterns;
- recognize socially appropriate responses in short spoken exchanges;
- recognize and distinguish English phonemes and the use of common intonation and stress patterns to convey meaning in carefully articulated speech; and
- infer implied meaning, speaker role, or context in short spoken exchanges.

Figure 4. Example of *Listen and Choose a Response* Task Type



Source: TOEFL iBT® test, ETS

Note. Test takers hear the following:

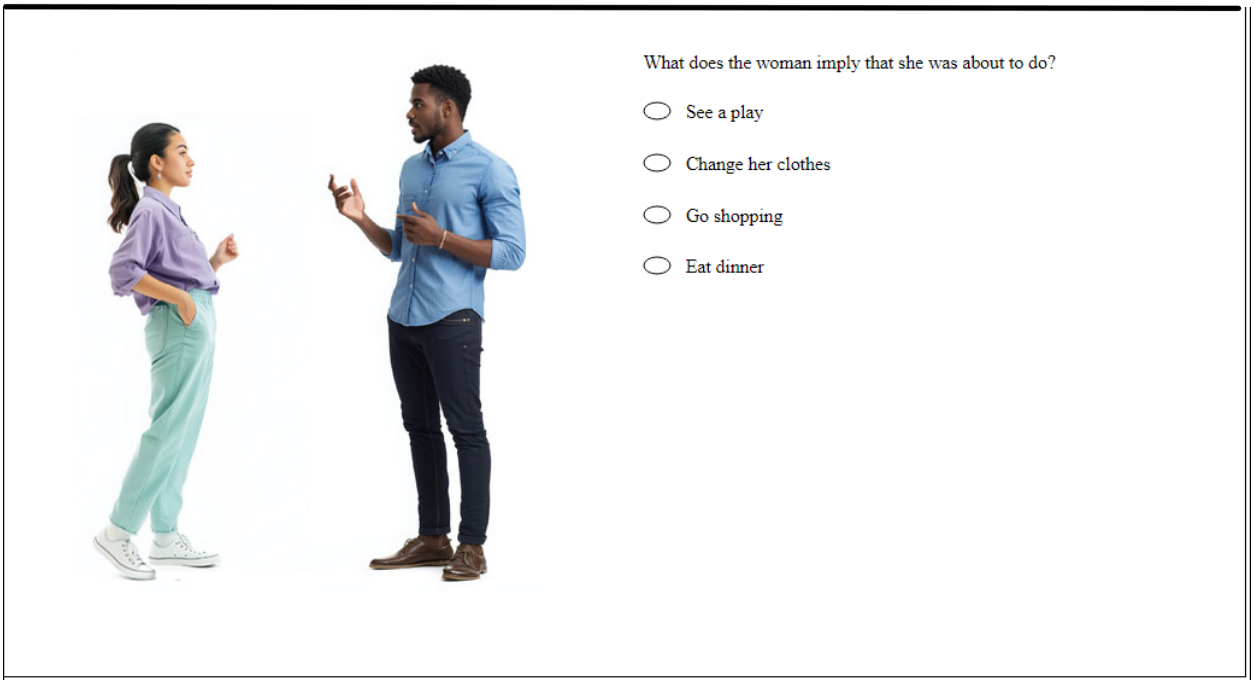
" Didn't I just see you in the library an hour ago?"

Listen to a Conversation

The *Listen to a Conversation* task (see Figure 5) is designed to measure the ability to fully comprehend a conversation in everyday situations. This ability involves more than just recognizing the spoken words; listeners must be able to make inferences, recognize speaker roles and purposes, and make predictions. The test taker listens to a short conversation between two speakers and answers questions about the conversation. The conversation may be on everyday topics in the public domain such as dining, social activities, education, entertainment, services, health, hobbies, home, shopping, communications, and travel. The questions require test takers to

- identify the main ideas and basic context of a conversation,
- understand the important details in a conversation,
- understand the range of grammatical structures used by proficient speakers,
- understand a wide range of vocabulary including idiomatic and colloquial expressions,
- infer meaning from information that is not explicitly stated,
- recognize the purpose of a speaker's utterance,
- make simple predictions about the further actions of the speakers, and
- follow the connection between ideas across speaker turns.

Figure 5. Example of *Listen to a Conversation* Task Type



Source: TOEFL iBT® test, ETS

Note. Test takers hear the following:

(F) Need anything from the supermarket?

(M) Huh? Aren't we getting ready to go see that play in a few minutes?

(F) That's tomorrow.

(M) Oh. Wow, I'd forget my head if it wasn't screwed on…. Guess I don't need to change my clothes after all.

(F) So, you weren't planning to prepare dinner?

(M) No, but I can. What do you want?

(F) Just something light and healthy. So, can you go shopping instead?

(M) Yeah, sure. How about salmon and salad? Want anything else?

(F) No, that's good. Thanks!

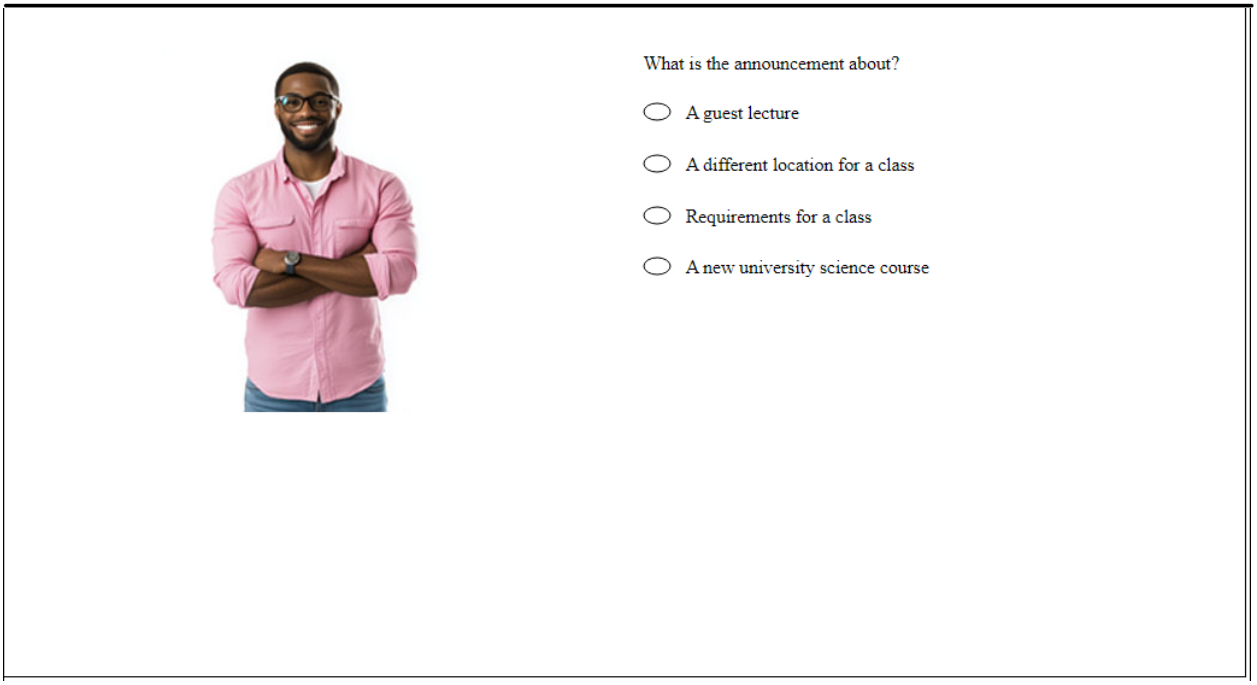
Listen to an Announcement

The *Listen to an Announcement* task is designed to simulate what a listener would hear either during an in-person or a broadcasted message in an academic context, for example, in a classroom or at a school-related event (see Figure 6). The test taker listens to a short academic-related announcement and then answers questions about it. The announcement may include information about schedules, directions, rules and regulations, or student achievements. The questions require test takers to

- identify the main ideas and basic context of a short message,

- understand the important details in a short message,
- understand the range of grammatical structures used by proficient speakers,
- understand a wide range of vocabulary including idiomatic and colloquial expressions,
- infer meaning from information that is not explicitly stated,
- predict future actions based on what a speaker has said, and recognize the purpose of a speaker's message.

Figure 6. Example of *Listen to an Announcement* Task Type



Source: TOEFL iBT® test, ETS

Note. Test takers hear the following:

Good afternoon, everyone. I am excited to inform you that Dr. Cynthia Palmer, a renowned expert in environmental science, will be giving a guest lecture next Monday at 2 pm in Waldman Auditorium. Dr. Palmer will discuss the latest advancements in sustainable energy solutions and their impact on global climate change. Due to her popularity and the high interest in her work, I highly recommend arriving early to secure a seat.

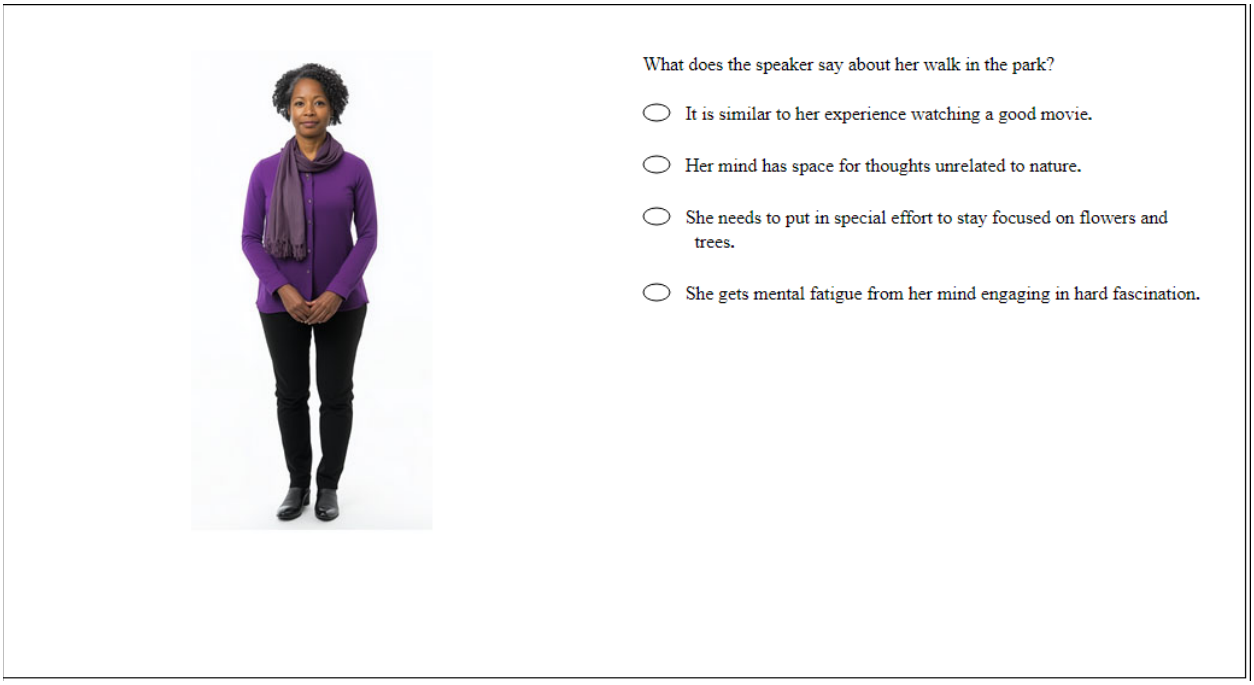
Listen to an Academic Talk

The *Listen to an Academic Talk* task is designed to simulate academic talks given by educators (see Figure 7). The test taker listens to a short (175–250 words) academic-related talk and answers questions about it. The task is designed so that background knowledge is not required. Topics are taken from fields such as history, art and music, life science, physical

science, business and economics, and social science. Test questions require test takers to

- understand the main and supporting ideas of a short academic talk,
- understand a range of grammatical structures,
- make inferences based on what is said,
- recognize the organizational features of the talk, and
- understand vocabulary that is sometimes uncommon, colloquial, or idiomatic.

Figure 7. Example of *Listen to an Academic Talk* Task Type



Source: TOEFL iBT® test, ETS

Note. Test takers hear the following: Did you see that new thriller movie that came out last week? I did and loved it. The action, the plot twists… I was totally captivated. Time just flew by. Not a single thought occurred to me that was unrelated to the movie. What I experienced is what psychologists call hard fascination. Hard fascination means intense focus and concentration. Whether it's TV programs, video games... hard fascination is all too easy to come by in this modern world.

There's another type of fascination—soft fascination. There's still effortless attention, meaning that no special effort is required for you to stay focused, but there's still room for other thoughts. When I take a walk in the park and look at the flowers and trees, for example, I might be thinking in the back of my mind about my dinner plans.

Now, one thing to know is hard fascination causes mental fatigue. The mind is so intensely focused that it gets tired fast. What follows mental fatigue? You might find yourself easily distracted, irritable, and stressed. Soft fascination, in contrast, engages a different part of the brain—the DMN, or Default Mode Network, which soothes the mind and helps combat mental fatigue. So next time you feel like your mind is on overload, turn off the TV, put down your phone. Take a walk, or simply sit and stare at the clouds.

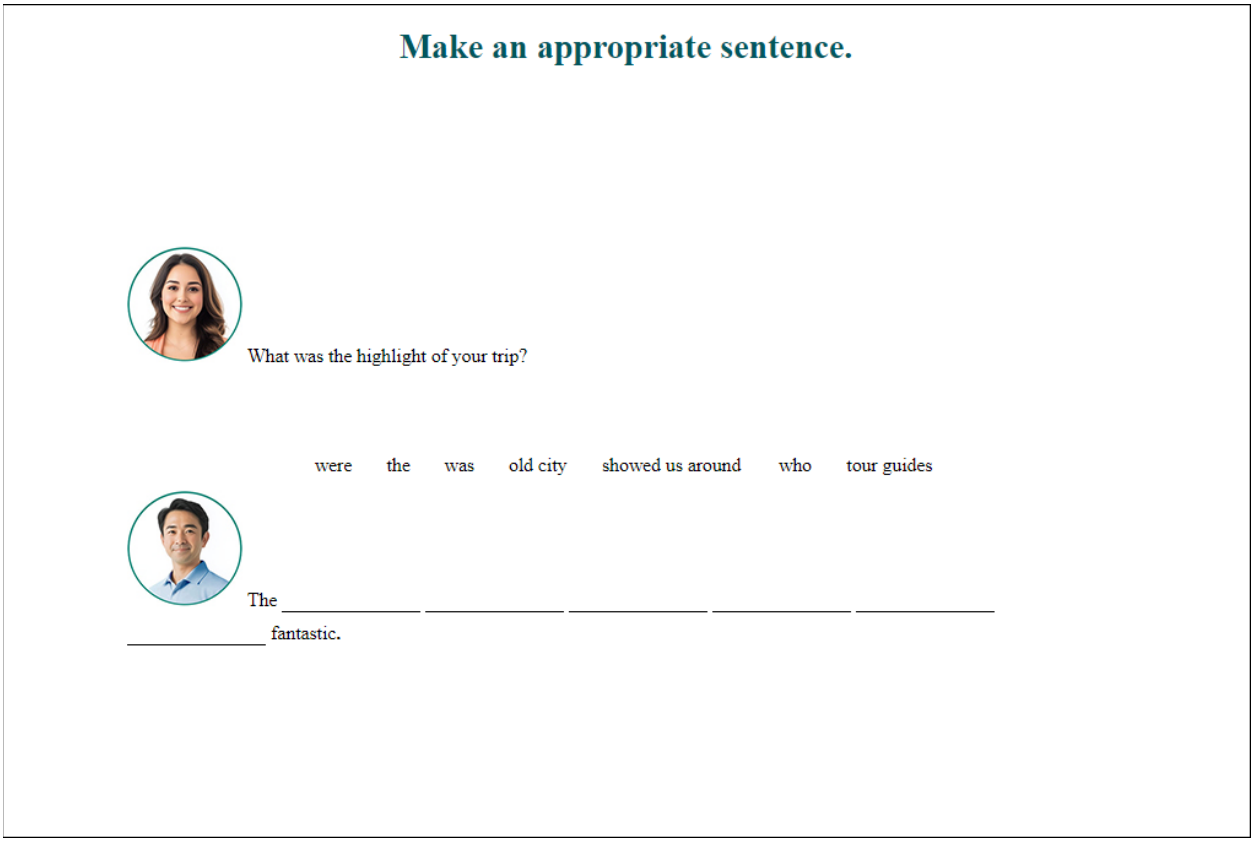
II-5. TOEFL Writing Task Types

Every day, people need to write, review, and edit texts in English for communication purposes that take place in a variety of settings, such as offices, labs, and classrooms. Such writing may take a variety of forms, including social media posts, instant messages, emails, and written course assignments. Writing skills are measured with the following task types: *Build a Sentence, Write an Email,* and *Write for an Academic Discussion.*

Build a Sentence

In the *Build a Sentence* task, test takers see several sentences with words or phrases in the wrong order and move them to form a grammatical sentence or question (see Figure 8). This task measures the test taker's command of sentence structures, a skill that is essential for all written communication.

Figure 8. Example of *Build a Sentence* Task Type



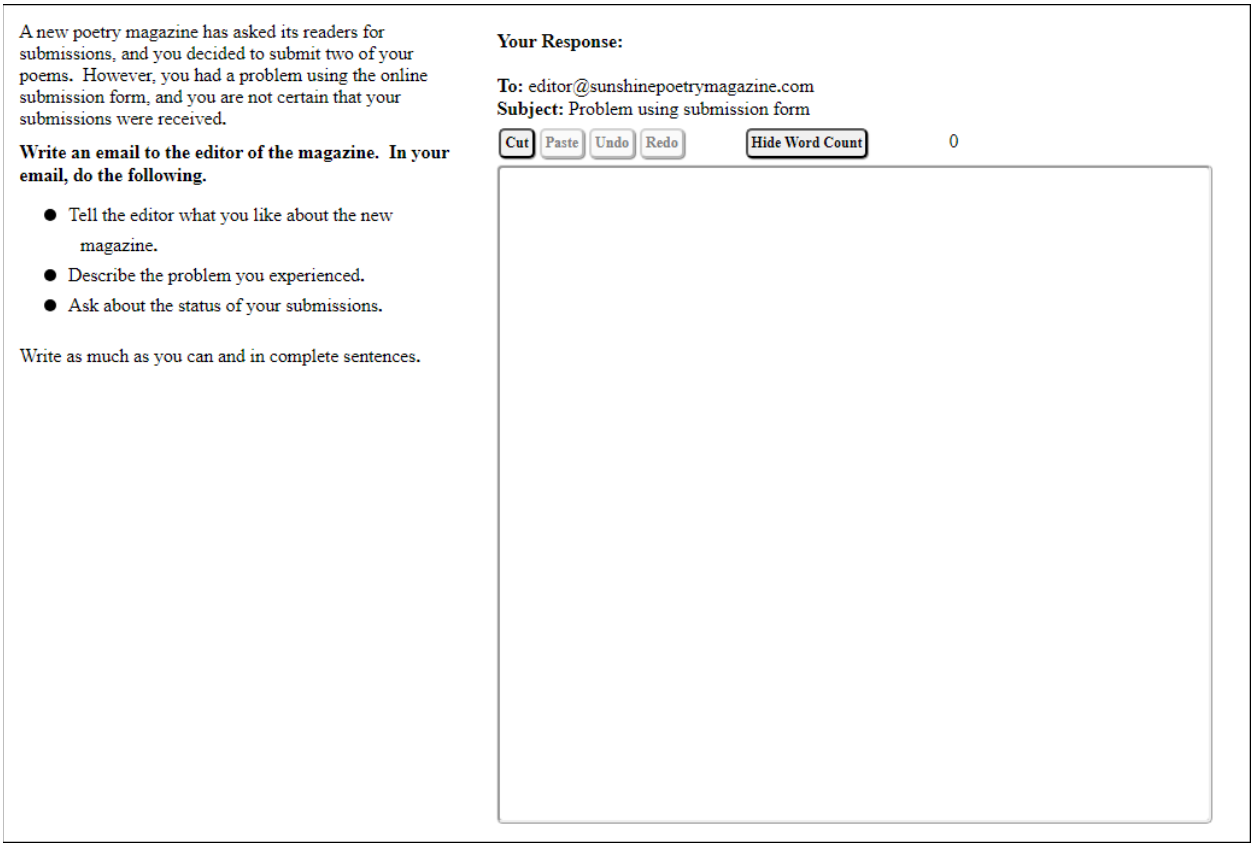
Source: TOEFL iBT® test, ETS

Write an Email

In the *Write an Email* task, test takers are presented with a scenario in text regarding either an academic or social setting (see Figure 9). A written explanation of the scenario and visual graphics are used to provide context to the task. Test takers are asked to share information in writing for a specific communicative purpose—for example, making a recommendation, extending an invitation, or proposing a solution to a problem. This writing task measures the test taker's ability to produce a multisentence written text that

- achieves the designated communication goal, following basic social conventions;
- is adequately elaborated, clear, and cohesive;
- makes accurate and appropriate use of a range of grammatical structures and vocabulary; and
- follows mechanical conventions of English (spelling, punctuation, and capitalization).

Figure 9. Example of *Write an Email* Task Type



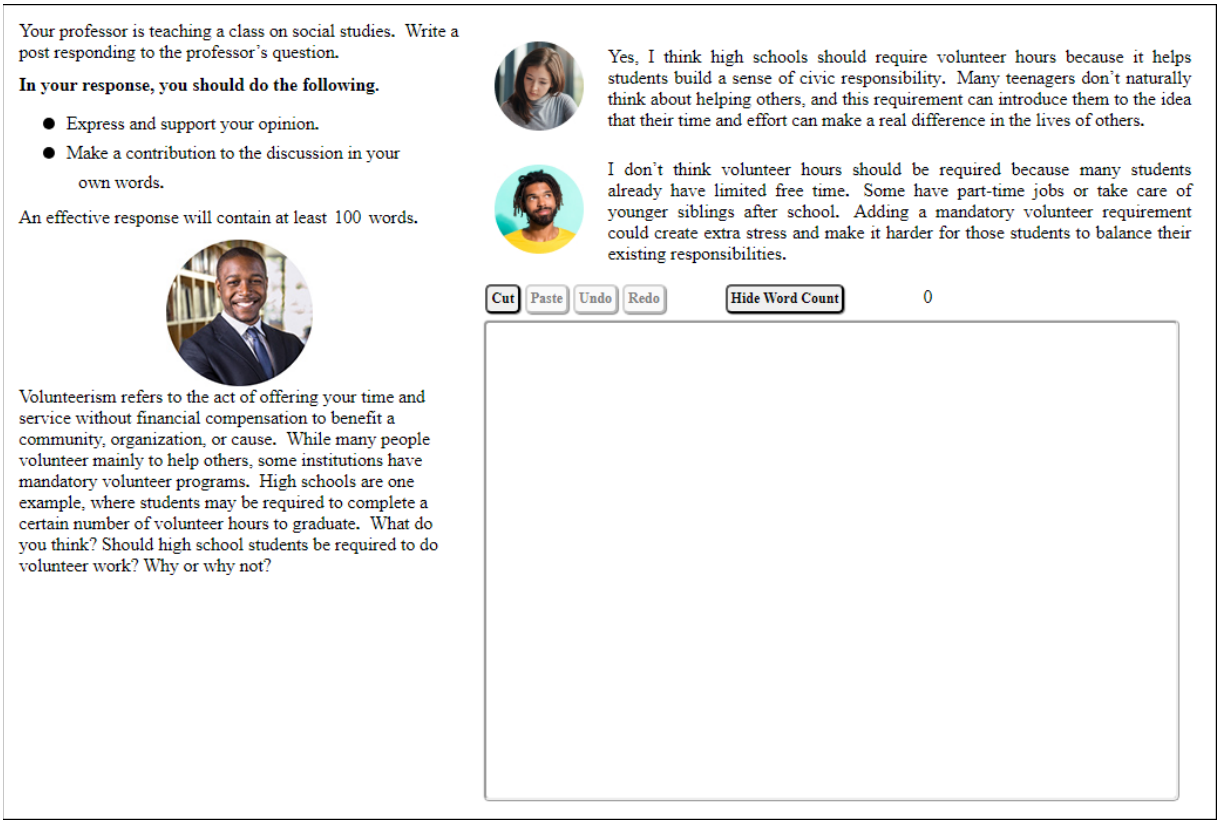
Source: TOEFL iBT® test, ETS

Write for an Academic Discussion

In the *Write for an Academic Discussion* task, test takers are asked to state and support an opinion within the context of an online class discussion forum (see Figure 10). A post from the professor briefly frames the topic and poses a related opinion question for discussion. Brief posts from other students then provide different positions on the issue. The test takers contribute their own position on the question, supporting their opinion with their own reasoning, experiences, or knowledge. This task measures the test taker's ability to produce a multisentence written text that

- clearly elaborates an argument for a position by responding to arguments and/or using information provided in short texts;
- is adequately supported, clear, and cohesive;
- makes accurate and appropriate use of a range of grammatical structures and vocabulary; and
- follows the mechanical conventions of English (spelling, punctuation, and capitalization).

Figure 10. Example of *Write for an Academic Discussion* Task Type



Source: TOEFL iBT® test, ETS

II-6. TOEFL Speaking Task Types

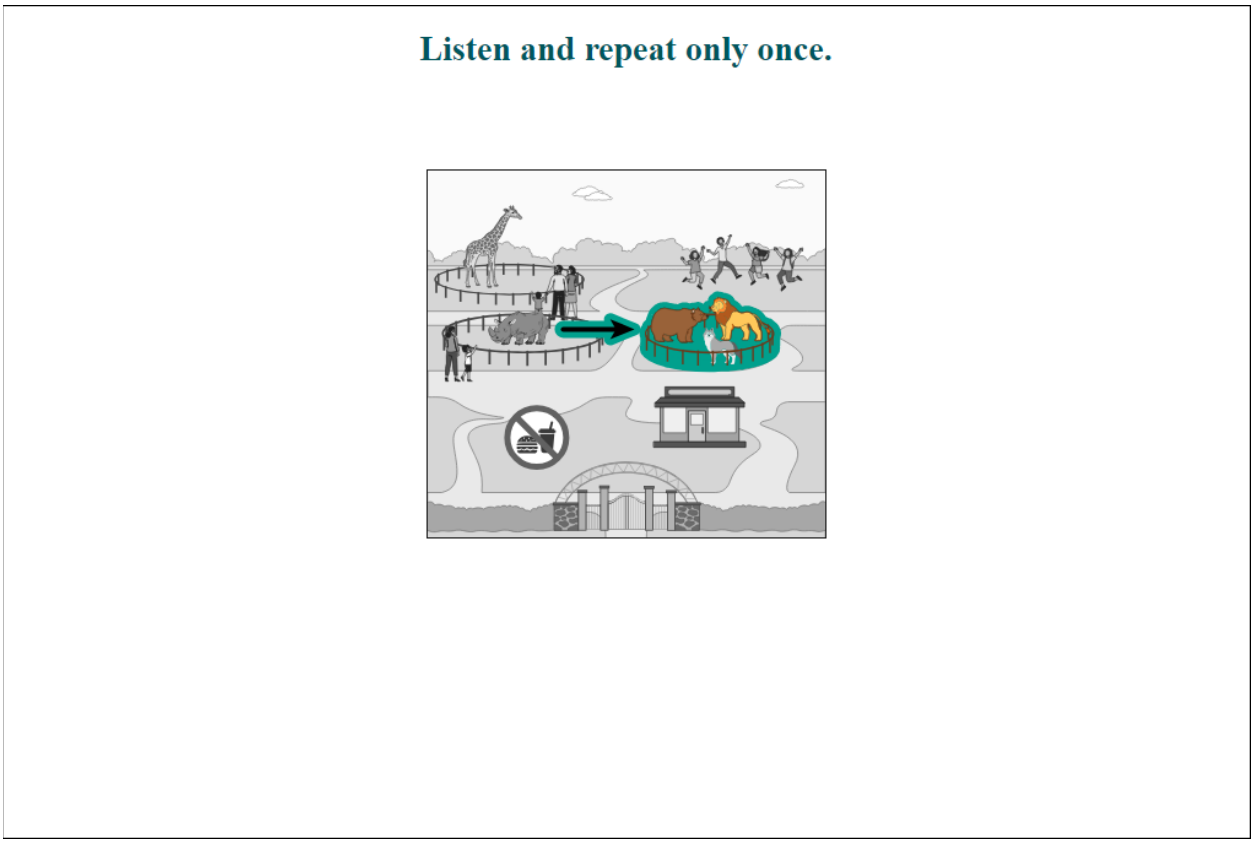
English speaking skills are critical for communicating in multiple ways with other people, including to socialize and to complete a wide range of academic or daily life tasks. The tasks in the Speaking section measure both foundational language skills as well as the ability to communicate. Foundational skills, such as the ability to process language and produce fluent and intelligible speech, are measured when test takers reproduce spoken input. Communication ability is measured when test takers speak about their opinions and experiences in the context of a simulated conversation. Speaking skills are measured with the following task types: *Listen and Repeat* and *Take an Interview*.

Listen and Repeat

The *Listen and Repeat* task measures the test taker's ability to process the sentences they hear and then accurately and intelligibly reproduce these sentences. In the *Listen and Repeat* task, test takers repeat a series of sentences within a scenario in an academic or daily life setting (see Figure 11). The scenario provides a communicative purpose for listening and repeating the sentences. Each series of sentences is associated with a visual representation of the setting, and progress through the sentences corresponds to visual movement through related parts of the illustration on the screen. After each sentence, there is a pause, and then test takers repeat exactly what was said. Sentences get progressively longer and more complex as test takers progress through the scenario. The *Listen and Repeat* task measures the test taker's ability to process the sentences they hear and then produce a spoken response that is

- an accurate repetition and
- clearly intelligible.

Figure 11. Example of *Listen and Repeat* Task Type



Source: TOEFL iBT® test, ETS

Note. Test takers hear audio and then repeat:

We have a variety of wildlife.

Bears, wolves, and large cats are to the right.

You can find sea lions and elephants further down the path.

Please, no outside food or drinks, and do not feed the animals.

Avoid banging or tapping on the displays and enclosures.

For those with children, we offer summer camps and educational opportunities.

The visitor's center, located near the front entrance, can give you more information.

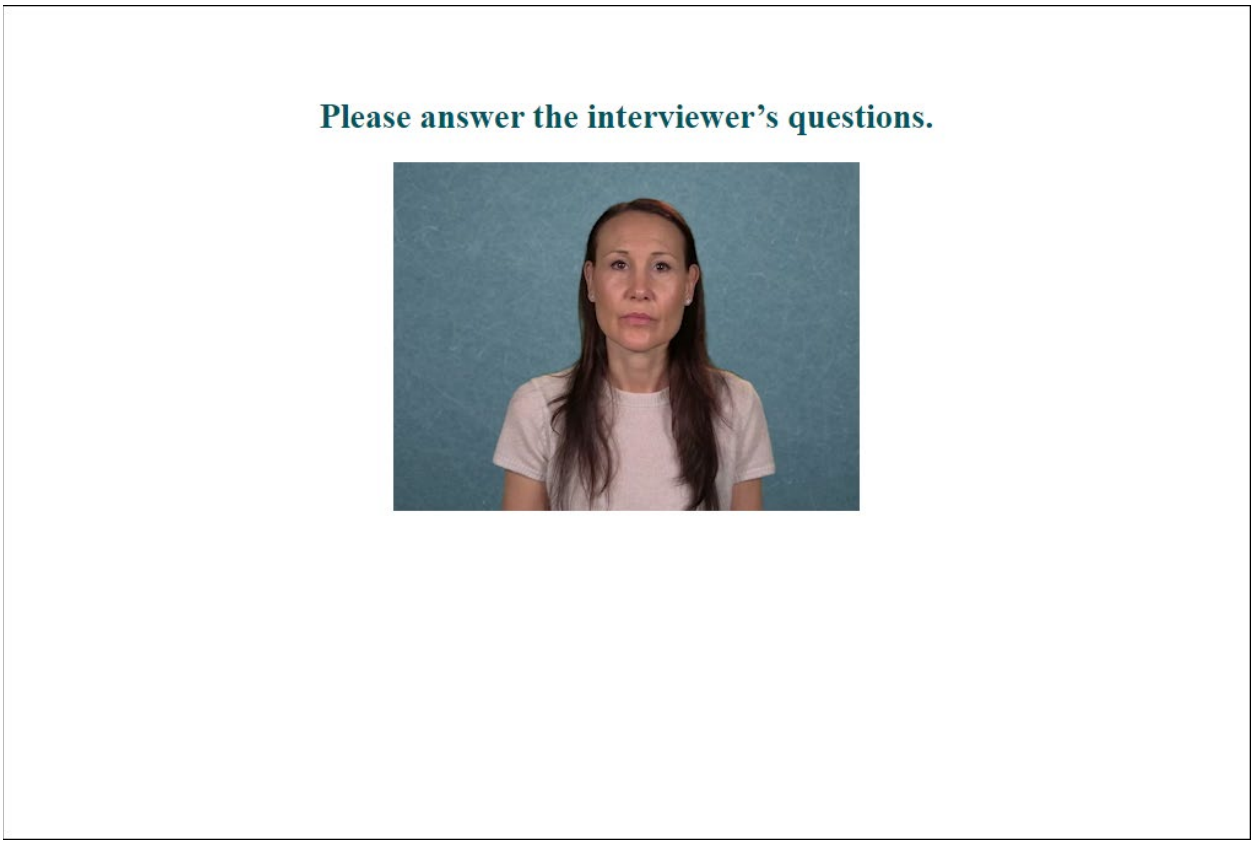
Take an Interview

In the *Take an Interview* task, test takers participate in a simulated conversation with a prerecorded interviewer (see Figure 12). The interview takes place during a variety of situations, such as applying for scholarships or participating in a research study, among others. During the interview, test takers answer questions related to the interview topic, where they describe their experiences and opinions. Initial questions focus on factual information and personal experience, whereas later questions ask test takers to express and support opinions regarding broader issues.

The *Take an Interview* task measures the test taker's ability to respond to a range of questions on general and academic topics, producing a spoken response that

- answers the question with appropriate and coherent elaboration;
- maintains a good conversational speaking pace;
- is intelligible and makes good use of rhythm and intonation to convey meaning; and
- makes effective and accurate use of a range of vocabulary and grammatical structures.

Figure 12. Example of *Take an Interview* Task Type



Source: TOEFL iBT® test, ETS

Note. Test takers hear audio and then answer the question.

Thank you for speaking with me today. I'm conducting a study about people's experiences and perceptions of living in a city. I'd like to ask you some questions. Now, do you currently live in a big city, a small town, or a village?

Great. Cities affect people in different ways. Some people find cities dynamic and exciting. Others find that cities are overwhelming and drain them of energy. What kind of reaction do you have to cities? Why do you think you react in this way?

OK. Next, I'd like to ask your opinion. Some people believe that those who live in cities lead more interesting lives. They would argue, for example, that people who live in cities have more access to professional opportunities and interesting leisure activities. Do you agree that people who live in cities lead more interesting lives? Why or why not?

Good points. Let me ask you one final question. For some time now, researchers have been interested in whether green spaces, such as parks, make people who live in cities happier. Do you think that city governments should create more parks in urban areas to promote a general sense of happiness and life satisfaction? Why or why not?

II-7. TOEFL Test Design

Reading and Listening Multistage Adaptive Test Design

To measure language proficiency efficiently, both the TOEFL Reading and Listening sections are designed as two-stage adaptive tests. The first stage, also known as router module, contains tasks of moderate difficulty (i.e., CEFR Levels B1 or B2). The second stage, or second module, follows, with its difficulty level determined by the test taker's performance on the first module. Content in the second stage of the Reading and Listening sections is classified as lower or upper difficulty modules. Each Reading and Listening section router module may include unscored tryout questions that are used for quality control and other operational purposes. Reading modules can contain 15 unscored questions; while Listening section modules may include 12 unscored questions. Each test taker completes a specific path, which consists of one router module and one second-stage module—either lower or upper. In both Reading and Listening sections, there are two possible paths:

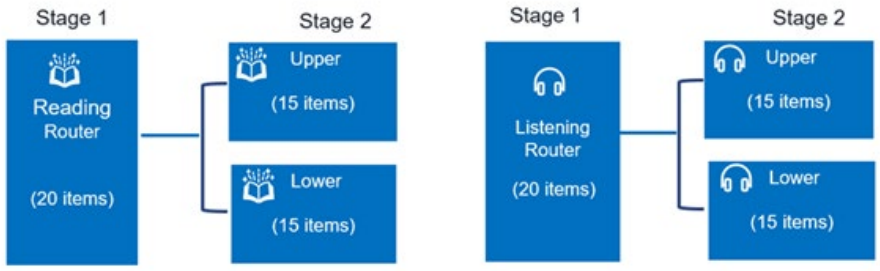
- Router + Lower module
- Router + Upper module

For example, if the student performs well on the first module of the Reading section, the second module received will be at a higher level of difficulty. The scoring for the Reading and Listening sections takes into consideration the total number of questions answered correctly across the two modules as well as the difficulty level of these modules included in a test taker's path. The MST design for the TOEFL test is presented in Figure 13. Table 1 details the content design for both the Reading and Listening sections.

The MST design was the preferred solution for the TOEFL test because it combines the advantages of adaptive and linear test designs (Hendrickson, 2007). MST balances practicality, flexibility, measurement accuracy, and control over test content coverage. When stringent psychometric requirements are met, MST offers practical benefits over question-level adaptive testing, such as better management of question pool usage, more control over test content and greater flexibility in test assembly (Zenisky et al., 2010). By employing MST methodology, the

TOEFL test measures language proficiency efficiently by matching test content to the test taker's ability level. At the same time, because adaptation happens at the section level and not at the individual question level, the test is able to operationalize the task-based approach in test design that underpins the design of other ETS language tests (Papageorgiou et al., 2021). In addition, section-level adaptation allows the test content to be assembled into multitask modules reflecting distinct levels of difficulty with expert assessment specialists' review of test content before administration. In other words, the MST methodology allows the TOEFL test to deliver relevant test content, including robust communication tasks, for its intended purposes in a targeted and efficient way.

Figure 13. TOEFL Reading and Listening Multistage Adaptive Test Methodology



Note. Each Reading and Listening section module may contain unscored questions.

Table 1. TOEFL MST Content Design for Reading and Listening Sections

Section	Task type	Number of scored questions in stages			Number of scored questions in paths	
		Stage 1	Stage 2 lower	Stage 2 upper	Easy path	Hard path
Reading	*Complete the Words*	10	10	10	20	20
	Read in Daily Life	5	5	0	10	5
	Read an Academic Passage	5	0	5	5	10
	Total	20	15	15	35	35
Listening	*Listen and Choose a Response*	8	7	3	15	11
	Listen to a Conversation	4	4	4	8	8
	Listen to an Announcement	4	4	0	8	4
	Listen to Academic Talk	4	0	8	4	12
	Total	20	15	15	35	35

Note. Each Reading and Listening section module may contain extra unscored questions.

TOEFL Writing and Speaking Design

The TOEFL Writing and Speaking sections are linear, where all test takers of a specific form receive the same set of tasks. Tasks in both sections are designed to be accessible across a range of proficiency levels with many opportunities for test takers to demonstrate writing and speaking skills. A range of difficulty combined with multiple measurement opportunities makes it possible to cover the full range of language proficiency without the need for separate stages. Scores for the Writing and Speaking sections are based on overall performance on all tasks.

The Writing section consists of three task types:

- Task 1: *Build a Sentence*
- Task 2: *Write an Email*
- Task 3: *Write for an Academic Discussion*

The Speaking section consists of two task types:

- Task 1: *Listen and Repeat*
- Task 2: *Take an Interview*

For the Writing section, the *Build a Sentence* task type contains 10 sentences. The *Write an Email* and *Write for an Academic Discussion* tasks each require one written response. For the Speaking section, the *Listen and Repeat* task type contains seven questions. The *Take an Interview* has four questions. Table 2 summarizes the task types, the numbers of questions, and the raw score ranges in the Writing and Speaking sections (for an explanation or raw scores, see Section II-8. Test Content Development Process).

Table 2. TOEFL Content Design for Writing and Speaking Sections

Section	Task type	Number of questions	Raw score range
Writing	*Build a Sentence*	10	0-10
	Write an Email	1	0-5
	Write for an Academic Discussion	1	0-5
	Total	12	0-20
Speaking	*Listen and Repeat*	7	0-5
	Take an Interview	4	0-5
	Total	11	0-55

II-8. Test Content Development Process

The development of each new test form (version) involves a complex series of steps. The aim of these steps is to develop new content according to strict quality and fairness standards and to produce test-taking experiences that are similar in content, difficulty, and level of engagement.

Test Development Staff

All ETS test developers, known as assessment specialists, have been trained in language learning or related subjects at the university and graduate level, and the majority of them have taught at K–12 schools, colleges, or universities internationally. Some assessment specialists are themselves English language learners who have achieved graduate-level degrees from universities where English is the language of instruction. These assessment specialists formulate the test stimuli (e.g., reading passages, lectures) and items (test questions and tasks) that the test takers eventually see.

Test Development Process

Assessment specialists follow detailed guidelines when selecting and creating test content (texts, audio, photographs, graphics, and videos) and writing test questions so that test content is construct relevant and comparable across different test administrations. They consider whether the test materials and the questions associated with them

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

ETS assessment specialists review test materials multiple times before using them in tests. Multiple assessment specialists who have not participated in the authoring stage sequentially and independently review each stimulus and its associated questions. They may suggest revising a stimulus or an associated question or rejecting a question or a stimulus entirely. Stimuli and questions 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 the review stages. Additionally, when required for a given test stimulus or question, a subject matter expert checks the accuracy and currency of the content in the stimulus. For some task types, ETS assessment specialists also use proprietary technological capabilities to

facilitate the content development process. These capabilities integrate task content specifications and difficulty parameters specifically developed for the TOEFL test. After the task content is generated through these capabilities, it undergoes the rigorous, multistage review process described previously.

Assessment specialists conduct multiple reviews of stimuli and questions for both language and content, considering questions such as these:

- Is the language in the test materials clear? Is it accessible to second language speakers of English?
- 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 relevance of what is being tested to the question specifications, the uniqueness of the answer or answers (the question keys), the clarity and accessibility of the language used, and the plausibility and attractiveness of the distracters—the incorrect options. For constructed response tasks (writing and speaking), the process is similar but not identical. Reviewers tend to focus on accessibility, clarity in the language used, and how well they believe a task will generate a fair and scorable response. It is also essential that reviewers judge each task to be comparable with others and at the intended level of difficulty. Expert judgment, then, plays a major role in deciding whether a writing or speaking task is acceptable and can be included in an operational test.

All TOEFL test materials receive an editorial review. The purpose of this review is to help ensure that all of the test content is as clear, concise, and consistent as possible. Both assessment specialists and editors use ETS-wide and test program–specific editorial and graphic guides to perform their reviews. In addition, when warranted, editors check facts in stimuli for accuracy or for advances in current knowledge (e.g., in areas such as physics or geography).

The *ETS Standards for Quality and Fairness* (ETS, 2014) mandates fairness reviews. This fairness review must take place before using materials in a test. All assessment specialists undergo fairness training—in addition to question-writing training—soon after their arrival at ETS. As part of their training, question writers become familiar with the *ETS Guidelines for Fair Tests and Communications* (ETS, 2016a) and the *ETS International Principles for Fairness of*

Assessments (ETS, 2016b) and use them when developing and reviewing test stimuli and questions. Fairness issues are thus considered at each stage of the development process.

Reviewers carefully analyze each stimulus or question before signing off. A subsequent reviewer typically consults with the previous reviewer on suggested changes to the stimulus or question. Thus, the test development process for the TOEFL test is collaborative.

After assessment specialists and the psychometric team approve test tasks, the materials enter a database and become available for assembly into a test. Each test form is assembled and reviewed so that it is similar in terms of content and statistical specifications to previous test forms. This similarity, in turn, facilitates score equating, which is the statistical process used to calibrate the results of different forms of the same test to ensure score comparability across forms.

Section III. Scoring and Score Reporting

III-1. Reading and Listening Scoring

As noted above, the TOEFL Reading and Listening sections follow an MST design with two stages in each section. In both the Reading and Listening sections, questions are evaluated as either correct or incorrect, with 1 score point awarded for each correct answer. The total score points that a test taker earns in each section—Reading section and Listening section—known as the *raw score*, are converted to a reported scaled score through a statistical process called *equating*. The item response theory (IRT) method is used for score equating in the TOEFL Reading and Listening sections. All test questions are calibrated using IRT and placed on a common scale. Reading and Listening test forms are assembled by selecting questions from a question pool that is regularly replenished with qualified items, following rigorous question analyses conducted under both classical test theory (CTT) and IRT framework. Question selection is guided by the content and statistical specifications of the TOEFL MST design. IRT true score equating is applied to generate the raw-to-scale conversion tables for each assembled test form. This process converts the raw scores on a new form to equated raw scores that represent corresponding raw scores on a pre-established base form. These equated raw scores are then transformed into scaled scores using the raw-to-scale conversion of the base form.

The application of equating procedures helps to support fairness for all test takers in several ways. First, the equated score for a test section takes into account the differences in difficulty introduced by the multistage adaptation. Second, the equating process accounts for any minor variations in difficulty across different versions of the test. Thus, a given reported score for a particular section reflects the same level of language ability irrespective of the second stage administered and when the test was taken. Note, because the scores are equated and scaled, the reported scores are not equal to the number or percentage of raw score points earned nor a simple common linear transformation of them.

III-2. Writing and Speaking Scoring

In the Writing section, all *Build a Sentence* questions are scored correct or incorrect, with 1 or 0 score points awarded respectively. Responses to the *Write an Email* and *Write for an Academic Discussion* tasks are scored on a scale from 0 to 5 score points according to criteria outlined in the scoring rubric. Responses to all speaking tasks are assigned scores from 0 to 5

score points based on criteria defined in the respective scoring rubric. Responses to the *Write an Email* and *Write for an Academic Discussion* in the Writing section, as well as all speaking tasks, are evaluated using ETS proprietary AI scoring engines as well as human scoring to enhance accuracy and consistency of scores.

The total writing and speaking raw score points are converted to a scaled score through innovative weighted equipercentile linking procedures that account for minor variations in difficulty among the different test versions (Haberman, 2015). This type of linking ensures that a given scaled score reflects the same level of language ability, regardless of when the test was taken, or which specific tasks were completed.

Development of Scoring Materials for Writing and Speaking

Separate scoring rubrics were created for each task type to reflect the fact that each task makes specific demands on the test taker and elicits differing evidence of language ability. Initial rubric development involved outlining the performance features considered relevant for good performance followed by review of sample responses collected in the prototyping study (see Section II-2: Test Design Process). Responses to prototype tasks were placed into quartiles by general proficiency of the test taker, as indicated by a *Complete the Words* measure, and then responses were sampled from each quartile and grouped by overall performance by a group of assessment specialists and research scientists. Specific scoring criteria were written to reflect performance characteristics observed in responses that were more or less successful in accomplishing the task followed by trial scoring of a random sample of responses drawn from each quartile. Revisions were then made to the scoring criteria and trial scoring repeated as needed.

The resulting draft rubrics were then used by a larger group of assessment and research staff to score all prototyping responses, after which additional adjustments were made as needed. Prior to scoring the responses from the pilot study, additional scoring aids were developed, including annotated sets of benchmark samples and sets of responses to be used for practice scoring. Following the pilot study, rubrics underwent further minor revision, primarily to help ensure consistency and clarity in the description of language phenomena. The corpus of sample responses was also greatly expanded using responses collected during the pilot study to meet the needs for large-scale scoring in the field test; this corpus included sets of annotated responses for benchmarks and practice scoring and nonannotated samples for rater calibration (certification of

rater accuracy). These materials were again reviewed following the field test, and minor revisions were made as needed to produce the scoring materials used in the operational test.

Automated Scoring of Writing and Speaking

The ETS proprietary automated scoring engines for the Writing and Speaking sections of the TOEFL test integrate advanced natural language processing (NLP) techniques, combining research with extensive operational expertise for enhanced performance. ETS builds automated scoring engines through an iterative process of response data modeling and rigorous evaluation of system performance. These models are regularly refined to maintain a secure, precise, and up-to-date scoring system (see, for example, McCaffrey et al., 2022; Zechner & Evanini, 2020.).

The automated scoring engine for the Writing section is designed to handle various question types through models tailored to assess different dimensions of writing, ensuring a comprehensive evaluation. It relies on a detailed mapping of writing features, such as relevance and elaboration of explanations (e.g., discourse coherence, prompt similarity metrics), syntactic variety (e.g., sentence variety, word frequency), social conventions (e.g., number of hedge words, use of modals), and the accuracy of content and language (e.g., grammatical errors, word usage errors, mechanical errors). The model is trained using supervised learning, where it learns to map these features to human-assigned scores. This training allows the model to make consistent, accurate assessments of writing quality by recognizing patterns in the features that correspond to all predefined scores. The model is then rigorously tested against established standards, using a variety of cutting-edge analytical methods to assess overall performance, with particular attention to fairness and accuracy for all test takers and subgroups. If a model does not meet the required standards, it undergoes refinement, retraining, and further optimization to enhance its precision.

Tables 3 and 4 provide an overview of the main construct areas (based on the scoring rubrics) for the *Write an Email* and *Write for an Academic Discussion* task types, respectively. For each construct area, examples of writing features or feature categories that are used by the scoring engine are provided.

Table 3. Writing Section–*Write an Email*

Scoring dimensions	Feature examples
Content "elaboration … supports the communication purpose"	• Number of sentences • Discourse coherence • Similarity to question prompt
Syntactic/Lexical variety "syntactic variety … idiomatic word choice"	• Sentence variety • Word frequency • Correctness of collocations
Social Conventions "politeness, register, organization … formulation of actions"	• Use of politeness indicators (e.g., modals, hedge words)
Accuracy/Errors "Almost no lexical or grammatical errors"	• Grammaticality • Grammatical errors • Word or usage errors • Mechanical errors (e.g., spelling or interpunctuation errors)

Table 4. Writing Section–*Write for an Academic Discussion*

Scoring dimensions	Feature examples
Content "Relevant and well-elaborated explanations … details"	• Number of sentences • Discourse coherence • Similarity to question prompt
Syntactic/Lexical variety "variety of syntactic structures and precise, idiomatic word choice"	• Sentence variety • Word frequency • Correctness of collocations

Similarly, the automated scoring engine for speaking tasks evaluates responses by analyzing key speech features that indicate speech fluency (e.g., words spoken per minute), intelligibility (e.g., pronunciation accuracy), grammatical accuracy (e.g., correct phrases or sentences), and coherence (e.g., discourse transition cues). It is trained using supervised learning, leveraging human-scored responses to establish reliable scoring patterns. The model undergoes rigorous testing to ensure it meets accuracy and fairness standards across question types and test taker subgroups.

Tables 5 and 6 provide an overview of the main construct areas (based on the scoring rubrics, see Appendix B) for the *Listen and Repeat* and *Take an Interview* task types,

respectively. For each construct area, examples of speech features or feature categories that are used by the scoring engine are provided.

Table 5. Speaking Section–*Listen and Repeat*

Scoring dimensions	Feature examples
Fluency	• Speaking rate • Length of uninterrupted 'runs' (word sequences without pauses) • Number of pauses • Number of hesitations
Intelligibility	• Correctness of pronunciation • Naturalness of speech rhythm • Naturalness of prosody (e.g., syllable stress)
Repeat accuracy	• Correctly repeated words • Similarity to prompt

Table 6. Speaking Section–*Take an Interview*

Scoring dimensions	Feature examples
Fluency	• Speaking rate • Length of uninterrupted *runs* (word sequences without pauses) • Number of pauses • Number of hesitations
Intelligibility	• Correctness of pronunciation • Naturalness of speech rhythm • Naturalness of prosody (e.g., syllable stress)
Language Use: Vocabulary and Grammar	• Vocabulary diversity (using a wide range of words that are distinct from one another) • Vocabulary richness (use of words which are less common) • Grammaticality • Grammatical accuracy (few grammar errors)
Organization	• Discourse coherence • Use of discourse connectives

Evaluation of Machine Scores for Writing and Speaking

The accuracy of automated machine scoring in the Writing and Speaking sections is crucial for maintaining the validity and reliability of the scores. To evaluate this accuracy, a

random sample of responses across all types of constructed-response tasks for the Writing section (N = 1,914) and the Speaking section (N = 1,521) was scored by two human raters. This scoring allows for the evaluation of the engines' performance in relation to the human raters' scores. Table 7 shows the correlation (Pearson r) between the average human rating and the automated score (i.e., Human–Machine) and between human ratings for single responses (i.e., Human–Human) for the Writing and Speaking sections. The Human–Machine correlations for the Writing and Speaking sections range from 0.86 to 0.89, suggesting a strong agreement between human and machine scores.

Table 7. Correlation of Writing and Speaking Sections by Scoring Method

Section	Human–Machine Correlation	Human–Human Correlation
Writing	0.86	0.85
Speaking	0.89	0.96

Note. Responses were analyzed from the *Write an Email* and *Write for Academic Discussion* tasks. The *Build a Sentence* writing task is key-scored and does not involve human or AI scoring.

Human Rater Training and Monitoring

Human rater training is a critical component of the overall scoring process of tasks in the TOEFL Writing and Speaking sections because the automated scoring engines are trained on human ratings. Human ratings not only set the standard for machine learning but also provide oversight to ensure the accuracy and reliability of automated scoring. The automated scoring is monitored in real time. For responses where the automated scoring lacks confidence or encounters difficulty, human raters step in to provide scores, ensuring reliability across all responses. In addition, a random sample of responses is regularly reviewed by certified human raters to ensure quality and inform model updates.

Human rater scoring quality for the tasks in the TOEFL Writing and Speaking sections is supported in a number of ways, similar to those for other ETS language tests (see Papageorgiou et al., 2021).

- The scoring process is centralized, and it is performed separately from the test administration to help ensure that test data is not compromised. Through centralized, separate scoring, each scoring step is closely monitored to help ensure its security, fairness, and integrity.

- ETS uses its proprietary scoring platform 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, specialists in English as a second/foreign language, or have other relevant experience. In addition to teaching experience, ETS prefers raters who have master's degrees and experience assessing spoken and written language.
- If raters have the formal qualifications, they are then trained using a web-based system. Following their training, raters must pass a certification test to be eligible to score.
- To help ensure reliability of constructed response scoring, scoring leaders monitor raters continuously as they score.
- L2 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 raters who are speakers of English as a first language.

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 proprietary scoring platform sends writing and speaking responses to multiple independent raters for scoring. Responses from each test taker are scored by more than one rater.

III-3. Band Scores and Ranges

Performance on each of the four sections and the overall test are reported in the form of band scores from 1 to 6, in increments of 0.5, rounded to the nearest whole or half band. The overall test score is derived by averaging the individual section band scores. Table 8 presents the raw score and band score ranges for the TOEFL test.

In addition to the section and overall band scores for current test administration, the score report includes MyBest® score (ETS, 2025) report data. These scores are the highest section

scores achieved in any test administration within the last 2 years. The overall band score for MyBest scores reflects the average of the highest section scores.

Table 8. Raw Score and Band Score Ranges for TOEFL Test Sections

Test	Raw score range	Band score range
Reading	0–35	1–6
Listening	0–35	1–6
Writing	0–20	1–6
Speaking	0–55	1–6
Overall	145	1–6

III-4. The Common European Framework of Reference Languages

The TOEFL test measures test takers' English proficiency from A1 to C2 levels on the CEFR. The scale scores and CEFR levels are on the same scale regardless of which test forms are taken. To facilitate the interpretation of section and overall band scores, information about their mapping onto the CEFR levels is provided on the score report and made available on the TOEFL website.

The mapping of TOEFL test scores to the CEFR levels was based on multiple sources of information. First, field test administrations for reading and listening tasks contained test questions previously included in other ETS language tests. Because the scores of these tests had already been mapped to the CEFR levels, it was then possible to also map the TOEFL Reading and Listening scores onto the CEFR levels. Reading and listening questions with a difficulty that fell between two CEFR levels were also inspected by ETS staff to determine if those questions reflected key skills and abilities described in the CEFR levels. In addition, assessment specialists examined relevant CEFR level descriptors to inform decisions about the design of the reading and listening tasks, such as target difficulty, types of stimuli, and comprehension skills to be assessed.

The mapping of the test scores in the TOEFL Writing and Speaking sections was established by combining information from several separate steps. First, task requirements and scoring rubrics were compared to CEFR subscales and level descriptors for different aspects of language to confirm that the content of the test was relevant to language ability as described in the CEFR, and therefore that alignment of test scores to CEFR levels was justified (Davis,

Garcia Gomez, et al., 2023). This step was followed by an ETS-internal standard setting study that identified minimum scores for each CEFR level, using the performance profile method (Fleckenstein et al., 2020). In this exercise, test takers representing different levels of performance (total writing or speaking score) were selected, and then a portfolio was constructed for each individual which contained the written or spoken responses they produced in the test. Language experts then compared the portfolios to performance descriptors from relevant CEFR scales to establish the minimum speaking or writing score for each CEFR level (Davis, Garcia Gomez, et al., 2023). Finally, the score profiles of the test takers in the field test were examined statistically to establish the relationship between the CEFR levels of the test takers across the selected-response sections and the CEFR levels of the same test takers across the constructed response sections of the test.

Table 9 presents the mapping of the TOEFL scores to the CEFR levels. To further facilitate score interpretation, performance descriptors are provided on the TOEFL website to illustrate the knowledge, skills and abilities expected by test takers. These descriptors have been selected from the CEFR (Council of Europe, 2001, 2020) with minor modifications so that they are more relevant to test content. Test takers receiving higher band scores are also expected to be able to demonstrate the performance described at lower band scores.

Table 9. Mapping TOEFL Test Scores to CEFR Levels

CEFR level	Reading	Listening	Writing	Speaking	Overall
C2	6	6	6	6	6
C1	5–5.5	5–5.5	5–5.5	5–5.5	5–5.5
B2	4–4.5	4–4.5	4–4.5	4–4.5	4–4.5
B1	3–3.5	3–3.5	3–3.5	3–3.5	3–3.5
A2	2–2.5	2–2.5	2–2.5	2–2.5	2–2.5
A1	1–1.5	1–1.5	1–1.5	1–1.5	1–1.5

Section IV. Test Administration and Security

IV-1. Test Display Sequence

TOEFL test takers will receive the Reading section first, followed by the Listening section, the Writing section, and the Speaking section. Table 10 provides an overview of the test display sequence for TOEFL.

Table 10. Overview of TOEFL Test Sequence

Test	Display sequence	Number of stages	Task types	Number of scored questions
Reading	1	2	*Complete the Words ; Read in Daily Life; Read an Academic Passage*	35
Listening	2	2	*Listen and Choose a Response; Listen to a Conversation; Listen to an Academic Talk*	35
Writing	3	1	*Build a Sentence; Write an Email; Write for an Academic Discussion*	12
Speaking	4	1	*Listen and Repeat; Take an Interview*	11
Total	--	--	--	93

IV-2. TOEFL Administration and Security Measures

The TOEFL test is delivered both in test centers and over the internet to test takers at their own locations (referred to as TOEFL iBT Home Edition) and at test centers. Test content is delivered using secure transmission protocols, and test forms are assigned through centrally controlled algorithms that consider the location of the test takers and their time zone.

TOEFL iBT Home Edition Security Measures

For at home testing, the test is monitored through a combination of AI and live remote proctoring. The AI-driven technology enhances the proctor capabilities to detect irregularities related to impersonations, assistance, unauthorized software use, and unauthorized use of suspicious objects in real time. The live remote proctoring capability enables a proctor to log into the remote test session and monitor test takers in real time. The remote proctor validates the identity of the test taker and secures the environment before granting access to the test. Each test taker receives a proctoring score that can be used to identify cases that may require additional review or score cancellation.

Prior to test administration, test takers are required to download a TOEFL Test App (TTA) which includes up-to-date security functions to minimize the opportunity to steal test

content and prevent suspicious activity. The test takers are able to run an equipment check and fix any technical issues before the test date.

On the test date, test security is safeguarded throughout the session by using online human proctors and AI security controls. The following main measures are taken prior to starting the test:

- Test takers are required to show a photo ID to their proctor and demonstrate their workspace meets several requirements.
- Test takers are required to integrate a mobile device with their test session to allow a second camera point of view to ensure environmental security.
- The proctor will do two mandatory and one randomly selected security checks before granting access. These checks can/may be related and not limited to scanning the room, checking for earpieces, or validating a clear desk.
- The proctor will request the test taker to use the second camera in their enabled mobile device to show the room and the computer screen, including devices connected to it. The TTA checks for applications that are not part of the TOEFL test administration and ensures that the screen is not shared remotely using unauthorized software. If an unauthorized application is running or the screen is being shared, the TTA will display a notification to inform the test taker of corrective steps that they must take in order to proceed to the test.

During the test, the following major security measures are implemented:

- The proctor monitors the computer screen, observes the examinee via the computer camera, and the mobile camera. The proctor can also cancel the test for security violations in real time.
- The proctor can communicate with the examinee, and examinees can also contact the proctor during the test.
- In addition to synchronous video-based human proctoring of examinees, there are technological innovations for monitoring activity and settings on the test taker's computer, and alerts are sent to proctors about unusual behavior or room conditions (for example, outside noises, communicating with someone other than the proctor, looking away from the screen, and moving away from the screen).

- The TTA locks down the device to prevent test takers from switching to other applications. It also prevents test takers from using short cut keys to cut, copy, and paste text in the Writing section response areas and from copying test content and transferring it to another application.

TOEFL iBT Test Center Security Measures

For tests delivered at an authorized test center, all four skills are delivered via computer under the supervision of trained test center personnel. ETS requires that TOEFL iBT test center administrators (TCAs) be at least 18 years of age and be able to read, write, speak, and understand English. Administrators must also complete certification training and pass an assessment. TCA responsibilities include the following:

- Perform check-in of test takers at the administrative station
- Ensure the security of the test center
- Write supervisor incident reports (SIRs)
- Train and coordinate activities with proctor(s)
- Ensure at least one TCA or proctor is present within every testing room at all times during all test sessions. An additional TCA or proctor must be in the room when more than 25 test takers are present, and two additional TCAs or proctors must be present when there are more than 40 test takers. ETS enforces this policy through unannounced audits of test centers.
- Operate the test center on a non-discriminatory basis
- Administer the test according to prescribed procedures and guidelines
- Use secure check-in procedures for test takers
- Check identification before admitting each test taker into the testing room
- Monitor test takers

In addition to the security measures implemented for both at home and test center testing, as noted earlier, the scoring of TOEFL is controlled centrally to further support security. For example, responses to tasks in the Writing and Speaking sections are evaluated by certified raters, whose scores are recorded and constantly monitored for quality by scoring leaders through

an ETS patented proprietary online platform. The use of the online platform helps ensure that raters will not know the examinees whose responses are being evaluated. Scores are also reviewed and analyzed statistically to identify suspicious patterns of test responses.

Also, after each test administration, comprehensive statistical analyses are carried out on all test takers' response data using advanced techniques to identify test takers with questionable responses. The results are further evaluated and investigated by the Office of Testing Integrity (OTI) at ETS.

Finally, the TOEFL Online Score Verification Service (OSVS) makes it possible for highly trusted organizations to verify the scores sent directly to them by the test taker. OSVS is free of charge, fast, and easy to use. In addition to score results and other personal data, it includes the test taker's original digital image, providing a clearer picture than what can be produced on paper score reports.

Section V. Score Reliability and Standard Error of Measurement

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 across different conditions of administration and/or administration of alternate forms (versions) of a test.

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, which means for some test takers the correct answer for the question depends not on their 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 is.

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. 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 test taker'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). The smaller the value of SEM, the higher the quality of measurement and the more precise the test scores will be.

Table 11 presents the section and overall score reliability estimates and SEMs evaluated based on field test data for a TOEFL form. Reliability estimation for the Reading and Listening sections of the TOEFL test is carried out using a method based on IRT (Kolen et al., 1996). For the Writing section of the test, reliability was estimated using stratified coefficient alpha (Rajaratnam et al., 1965), a measure of internal consistency reliability that offers more accurate than regular coefficient alpha when subsets of questions measure distinct content categories. The reliability estimate for the Speaking section was based on an index known as coefficient alpha (Cronbach, 1951). Cronbach's alpha helps to evaluate internal-consistency reliability, which indicates the consistency of test takers' responses across the questions, as well as whether the questions are measuring the same trait that they are intended to measure. Table 11 indicates that TOEFL section and overall scores are highly reliable, meeting the criteria for high stakes use outlined in the *ETS Standards* (ETS, 2014) as well as *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Table 11. Reliability Estimates and Standard Error of Measurement

Section	Score scale	Reliability estimate	*SEM*
Reading	1–6	0.86	0.37
Listening	1–6	0.88	0.35
Writing	1–6	0.87	0.36
Speaking	1–6	0.94	0.22
Overall	1–6	0.90	0.32

A final note to understand these reliability indices is that for making high-stakes decisions, such as admissions to college or graduate school, the overall score provides the best information—both because it reflects all four language skills and because it is the most reliable measure, as it is based on responses to all test tasks. 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 other information in addition to TOEFL test scores, such as grade point average, scores on other admissions exams, teacher recommendations, or interviews with individuals.

Section VI. Validity and Fairness

VI-1. Validity

Validity refers to the extent to which a test measures what it is intended to measure, as supported by theory and evidence (AERA et al., 2014; ETS, 2014). The construct definition of a test establishes what the test intends to measure. Typically, validity is supported through the collection of evidence from the test development process and subsequent research that shows (a) how test content aligns with the construct to be measured (content validity), (b) whether questions function as expected within the construct framework (internal structure), (c) how test scores correlate with related outcomes or external measures (criterion-related validity), (d) whether test takers' cognitive processes reflect engagement of the targeted skill (response processes), and (e) the extent to which the outcomes of test use are beneficial (consequential validity).

The TOEFL test was designed to provide information about language proficiency that can support important decisions (e.g., admission of international students to higher education institutions). The use of test scores must be supported by a research program that considers relevant aspects of test design and score interpretation, providing evidence that a particular use of the test is appropriate. As is the case with the other ETS language tests (e.g., Chapelle, 2008; Hsieh, 2024a, 2024b; Papageorgiou et al., 2021), the research program for the TOEFL test is organized following an argument-based approach to validation (Kane, 2013). This approach to test validation consists of providing support for core claims about the test score interpretation and use. To provide this support, specific claims about the test (or warrants) are stated, and these claims require backing from theory, test documentation, or empirical evidence. Rebuttals must also be considered, which are alternative claims that can challenge the original warrant. Data are gathered to provide backing for warrants or to evaluate the credibility of potential rebuttals.

The core claims for the score interpretation and use of the TOEFL test are organized into six hierarchical inferences, following those laid out in Chapelle (2008) to support the TOEFL validity argument (see Table 12 at the end of this section). The six inferences cover all aspects of test design and score interpretation and use, from designing test tasks that reflect real-life use of the language (the domain inference) to generating scores that are psychometrically sound (the evaluation, generalization, and explanation inferences) and are useful for making important decisions related to English language proficiency (the extrapolation and utilization inferences).

Each inference is associated with a core claim accompanied by related warrants and examples of empirical evidence that might be used to support (or counter) each warrant.

The warrants in the TOEFL validity argument reflect what Chapelle (2008) described as a "design validity argument" (p. 320). Given that this iteration of the TOEFL test has not launched at the time of writing, the inferences in the validity argument have so far been investigated as part of the test development process. The research conducted during the development of the TOEFL test collected initial evidence to justify the interpretation and intended use of the test scores. After the test is operational, the research program for the TOEFL test will continue to investigate the various claims in the validity argument as test scores are actually interpreted and used by stakeholders. This staged approach to test validation is in keeping with the notion that distinct questions can and should be prioritized for investigation at distinct stages in the development and use of language assessments (Norris, 2008). During the test development stage, validity questions addressed primarily the concerns with domain definition and evaluation as listed in Table 12, including questions about the constellation of tasks that comprise the assessment, the extent to which they reflect a targeted language proficiency construct, how test takers interact with and navigate through test content, whether test-taker responses can be scored reliably, and whether scores on the test can be expected to reveal the intended language proficiency differences. Subsequent planned investigations will address other claims related to generalization, explanation, extrapolation, and utilization (see Table 12).

VI-2. Fairness

Fairness is a central component of all ETS products and services. All materials undergo rigorous reviews for fairness by trained staff who apply, in compliance with *Standards for Educational and Psychological Testing* (APA et al., 2014).

Throughout all stages of design, development, and delivery, the TOEFL program implements quality control measures to ensure the test and test scores are fair, or, in other words, equally valid for all test takers, regardless of nationality, age, or gender. The *Test Development* Section describes how test questions are reviewed systematically and thoroughly to ensure fairness across all aspects. Preliminary studies have been conducted to the extent possible to evaluate fairness of test questions. For example, the comprehensive field test data conducted in 2021 (Papageorgiou et al., 2021) showed comparable performance on Reading, Listening,

Writing, and Speaking task types, which were eventually used in the TOEFL test, across test takers grouped by gender, age, employment status, time spent studying English, and having lived in a country where English is the main language; in addition, these subgroups reacted similarly to AI and human voice rendering of the same task types (Choi & Zu, 2022). Wang (2021) performed a differential item functioning (DIF) study for male and female test takers who took 1,454 tasks (624 Reading tasks, 593 Listening tasks, 48 Writing tasks, and 189 Speaking tasks), whose design formed the basis for the test tasks in the TOEFL test. DIF is a statistical methodology investigating the extent to which groups of test takers with similar levels of language proficiency perform differently on the same test tasks. Of the 1,454 test tasks, only 21 tasks were flagged (5 for Reading, 5 for Listening, 1 for Writing, and 10 for Speaking). However, the test developers who then reviewed these tasks concluded that there was no content bias based on gender. Note that if tasks flagged for DIF are deemed to be biased in terms of their content, then they are removed from further usage. Lu (2025) found that the automated scoring systems of the TOEFL Writing and Speaking sections did not unfairly disadvantage major L1 subgroups with sufficient sample sizes in the field test data.

As the TOEFL test is administered under operational conditions, new evidence regarding fairness will be collected to support relevant claims in a validity argument.

Table 12. Overview of Inferences in the Validity Argument for the TOEFL Test

Inferences	Core claim	Warrant (supporting claim)	Potential backing (supporting evidence)
Domain definition	Observations of performance on the TOEFL test reveal knowledge, skills, and abilities relevant to the domains of academic and general language use.	Test tasks measure foundational aspects of language proficiency	• Review of literature from second language acquisition documents: (1) developmental sequences (e.g., acquisition of word order rules), and (2) the theoretical and empirical linkages between acquisition and specific performance measures (e.g., elicited imitation) • Construct definition proposing a model language ability consisting of foundational skills plus communicative abilities.
		Test tasks reflect language use in academic and general (daily-life) English contexts	• Review of relevant literature and other sources documents the essential language required for academic and general contexts. • Specifications for test tasks document that they capture language skills relevant to communication in academic and general English situations. • Key stakeholders, e.g., students and teachers, believe that the test tasks measure relevant language abilities.
		The test is free of content that might unfairly influence test taker performance	• Procedures are in place to review test content to avoid material that might be objectionable, confusing, or otherwise influence test-taker behavior in construct-irrelevant ways.
Evaluation	Observations of performance on the TOEFL test tasks are evaluated to produce scores reflective of targeted language abilities.	Task administration conditions are appropriate for providing evidence of targeted language abilities.	• Usability data show that test takers successfully navigate test tasks. • System reliability data show minimal technical interruptions; procedures exist for recovering from disruptions during the test, and re-testing is available if needed.
		Task features impact performance in expected ways.	• Comparisons of performance on tasks with differing features show that design features affect performance (or not) as expected.

Inferences	Core claim	Warrant (supporting claim)	Potential backing (supporting evidence)
		Scores for constructed-response tasks reflect the targeted language abilities and skills.	• Correspondence is seen between performance features of constructed responses and corresponding scores awarded. • Rubric development is based on both construct considerations and sampling of test taker responses; scoring rubrics are iteratively revised to help ensure that criteria are appropriate to both the targeted construct and the test taker population. • Procedures are in place to ensure raters are well-trained. Analyses of scores show raters apply the scoring materials consistently (e.g., rater agreement and reliability). • Rater perceptions confirm the scoring criteria are appropriate. • Automated scores are similar to human scores; language phenomena evaluated in automated scores is consistent with scoring criteria used by human raters. Procedures are in place for resolving human-human and human-machine disagreements.
		Scores are free from bias or other types of unfairness.	• Procedures are developed for consistent scoring of all responses. • Scores awarded to defined sub-groups of test takers do not differ.
		Test tasks distinguish among examinees with varying degrees of proficiency.	• Discrimination of items and reliability of sections/test meet acceptable standards.
		Examinees are routed to items of appropriate difficulty (i.e., the MST design functions as planned).	• The difficulty of the second part of each test section increases (or decreases) depending on whether the examinee did well (or poorly) on the first part. Thus, the distribution of scores on each level of the second part of the test is consistent with the expected distribution of test taker proficiency.
		Item responses are scored with high accuracy and combined consistently into total scores.	• Procedures for scoring and rules for combining scores are well-defined.

Inferences	Core claim	Warrant (supporting claim)	Potential backing (supporting evidence)
Generalization	Observed scores are estimates of expected scores over the relevant parallel versions of the test tasks and test forms and across raters.	A sufficient number of tasks are included on the test to provide stable estimates of test takers' performances.	• Reliability and generalizability studies show that scores meet requirements for consistency and precision.
		Appropriate scaling and equating procedures for test scores are used.	• Description of equating procedures that account for minor variations in difficulty among the different test versions (forms) as well as the differences in difficulty introduced by the section-level MST adaptation.
		Task and test specifications are well-defined so that parallel tasks and test forms are created.	• Description of task specifications and task development processes help ensure consistency in creation of test content.
Explanation	Expected scores are attributed to the relevant construct of academic language proficiency in academic and daily-life contexts.	The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components.	• Factor analysis of the test confirms expected internal structure.
		The linguistic knowledge, processes, and strategies required to successfully complete tasks vary in keeping with theoretical expectations.	• Cognitive processing investigations show that tasks elicit expected strategies and abilities. • Higher- and lower-scoring constructed responses show expected differences in performance characteristics.

Inferences	Core claim	Warrant (supporting claim)	Potential backing (supporting evidence)
		Performance on the test measures relates to performance on other test-based measures of language proficiency as expected theoretically.	• Scores show expected relationship to other tests in the TOEFL family. • Scores show expected relationships to other measures of general language proficiency (e.g., C-Test)
Extrapolation	The construct of academic language proficiency as assessed by the TOEFL test accounts for the quality of linguistic performance in English-medium institutions of higher education and other relevant academic and daily life contexts.	Performance on the test is related to real-life measures of language proficiency within the context of use.	• Test scores are associated with indicators of real-life performance such as grades, samples of academic work, teachers' judgements, or other measures of academic success. • Test scores are also associated with performance in general English contexts as appropriate, such as evaluations of language use in job performance.
Utilization	Scores from the TOEFL test are useful for making important decisions, such as those related to educational admissions and instruction.	The meaning of test scores is clearly interpretable by stakeholders.	• Test scores are mapped to external language proficiency levels (CEFR). • The relationship of the test scores with the scores of other tests in the TOEFL family is established empirically through vertical scaling research. • Usability studies show stakeholders correctly interpret information contained in the score report. • Information about the interpretation of the band scores is publicly available.
		The test will have a positive influence on learning and instruction.	• Score users find the section scores, the availability of speaking and writing responses useful for making educational decisions. • Admissions and placement decisions are perceived by learners and teachers to be accurate.

References

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

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Chapelle, C. A. (2008). The TOEFL® validity argument. In C. Chapelle, M. Enright, & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). Routledge.

Choi, I., & Zu, J. (2022). *The impact of using synthetically generated listening stimuli on test taker performance: A case study with multiple-choice, single-selection items* (TOEFL Research Report No. RR-98). ETS. https://doi.org/10.1002/ets2.12347

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press. https://rm.coe.int/1680459f97

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume.* https://rm.coe.int/common-europeanframework-of-reference-for-languages-learning-teaching/16809ea0d

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

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic microfeatures to predict L2 writing proficiency: A case study in automated writing evaluation. *Journal of Writing Assessment*, *7*(1). http://www.journalofwritingassessment.org/article.php?article=74

Davis, L., Garcia Gomez, P., Li, S., & Manna, V. F. (2023). Mapping TOEFL Essentials® speaking and writing scores to the CEFR levels. In S. Papageorgiou & V. F. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation* (pp. 120–140). John Benjamins.

Davis, L., Norris, J., Papageorgiou, S., & Sasayama, S. (2023). Balancing construct coverage and efficiency: Test design, security, and validation considerations for a remotely proctored

online language test. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology mediated language assessment* (pp. 49–63). Routledge.

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

ETS. (2016a). *ETS guidelines for fair tests and communications.* https://www.ets.org/pdfs/about/fair-tests-and-communications.pdf

ETS. (2016b). *ETS international principles for the fairness of assessments.* https://www.ets.org/pdfs/about/fairness-review-international.pdf

ETS. (2025). *MyBest® Scores.* https://www.ets.org/india/toefl/institutions/ibt/set-score-requirements.html Fleckenstein, J., Keller, S., Kruger, M., Tannenbaum, R.J., & Koller, O. (2020). Linking TOEFL iBT® writing rubrics to CEFR levels: Cut scores and validity evidence from a standard setting study. *Assessing Writing*, *43*, 1–15.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273. https://doi.org/10.3102/1076998615574772

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

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice*, *26*(2), 44–52. https://doi.org/10.1111/j.1745-3992.2007.00093.x

Hsieh, C.-N. (2024a). *Building a validity argument for the TOEFL Junior® tests* (TOEFL Research Report No. RR-102). ETS. https://doi.org/10.1002/ets2.12379

Hsieh, C.-N. (2024b). *Building a validity argument for the TOEFL Primary® tests*. (Research Report No. RR-24-16). ETS. https://www.ets.org/Media/Research/pdf/RR-24-16.pdf

Hulstijn, J. H. (2015). *Language learning & language teaching: Vol. 41. Language proficiency in native and non-native speakers: Theory and research.* John Benjamins. https://doi.org/10.1075/lllt.41

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*, 129–140.

Lu, R. (2025). Psychometric evaluation of the quality of the automated scoring for TOEFL Writing and Speaking [Internal memo]. ETS.

McCaffrey, D. F., Casabianca, J., Ricker-Pedley, K. L., Lawless, R., & Wendler, C. (2022). *Best practices for constructed-response scoring* (Research Report No. RR-22-17). ETS. https://doi.org/10.1002/ets2.12358

Norris, J. M. (2005). Using developmental sequences to estimate ability with English grammar: Preliminary design and investigation of a web-based test. *Second Language Studies*, *24*(1), 24–128. https://core.ac.uk/download/pdf/77238726.pdf

Norris, J. M. (2008). *Validity evaluation in language assessment.* Peter Lang. https://doi.org/10.3726/978-3-653-01171-5

Norris, J. M. (2018). Task-based language assessment: Aligning designs with intended uses and consequences. *Journal of Technology, Learning and Assessment*, *21*, 3–20. https://doi.org/10.20622/jltajournal.21.0_3

Norris, J. M., & Ortega, L. (2012). Assessing learner knowledge. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 573–589). Routledge.

Papageorgiou, S., Davis, L., Norris, J. M., Garcia Gomez, P., Manna, V. F., & Monfils, L. (2021). *Design framework for the TOEFL Essentials® test 2021* (Research Memorandum No. RM-21-03). ETS.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*(3), 513–536. https://doi.org/10.1111/1467-9922.00193

Qian, D. D., & Lin, L. H. F. (2020). The relationship between vocabulary knowledge and language proficiency. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 66–80). Routledge. https://doi.org/10.4324/9780429291586-5

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, *29*(3), 325–344. https://doi.org/10.1177/0265532211424478

Wang, L. (2021). Groups for DIF analysis for TOEFL assessments. [Internal memo]. ETS.

Xi, X., & Norris, J. M. (Eds.). (2021). *Assessing academic English for higher education admissions.* Routledge. https://doi.org/10.4324/9781351142403

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*(4), 497–528. https://doi.org/10.1177/0265532215594643

Zechner, K., & Evanini, K. (Eds.). (2020.). *Automated speaking assessment: Using language technologies to score spontaneous speech.* Routledge.

Zenisky, A. L., Hambleton, R. K., & Luecht, R. M. (2010). Multistage testing: Issues, designs, and research. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 355–372). Springer.

Appendix A: Scoring Rubrics

TOEFL Writing Rubrics

For two of three TOEFL Writing task types, test takers produce written responses: the *Write an Email* task and the *Write for an Academic Discussion* task. Separate rubrics—or guidelines for scoring—are used to evaluate test taker responses.

Write an Email

In the *Write an Email* task, test takers are presented with a scenario in text regarding either an academic or social setting. Scores for this task type range from 0 to 5 (see Figure A1).

Write for An Academic Discussion

In the *Write for an Academic Discussion* task, test takers are asked to state and support an opinion within the context of an online class discussion forum. Scores for this task type range from 0 to 5 (see Figure A2).

TOEFL Speaking Rubrics

For both of the TOEFL Speaking task types—*Listen and Repeat* task and the *Take an Interview* task—test takers produce spoken responses. Separate rubrics—or guidelines for scoring—are used to evaluate test taker responses.

Listen and Repeat

In the *Listen and Repeat* task, test takers repeat a series of sentences within a scenario in an academic or daily life setting. Scores for this task type range from 0 to 5 (see Figure A3).

Take an Interview

In the *Take an Interview* task, test takers participate in a simulated conversation with a prerecorded interviewer. Scores for this task type range from 0 to 5 (see Figure A4).

Figure A1. Rubric for the *Write an E-mail* Task Type

Score	General Description
5	**A fully successful response** The response is effective, is clearly expressed, and shows consistent facility in the use of language. A typical response displays the following: • Elaboration that effectively supports the communicative purpose • Effective syntactic variety and precise, idiomatic word choice • Consistent use of appropriate social conventions (e.g., politeness, register, organization of information and formulation of actions such as requests, refusals, criticisms, etc.) • Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there/their*)
4	**A generally successful response** The response is mostly effective and easily understood. Language facility is adequate to the task. A typical response displays the following: • Adequate elaboration to support the communicative purpose • Syntactic variety and appropriate word choice • Mostly appropriate social conventions • Few lexical or grammatical errors
3	**A partially successful response** The response generally accomplishes the task. Limitations in language facility may prevent parts of the message from being fully clear and effective. A typical response displays the following: • Elaboration that partially supports the communicative purpose • A moderate range of syntax and vocabulary • Some noticeable errors in structure, word forms, use of idiomatic language and/or social conventions
2	**A mostly unsuccessful response** The response reflects an attempt to address the task, but it is mostly ineffective. The message may be limited or difficult to interpret. A typical response exhibits one or more of the following: • Limited or irrelevant elaboration • Some connected sentence-level language, with a limited range of syntax and vocabulary • An accumulation of errors in sentence structure and/or language use
1	**An unsuccessful response** The response reflects an ineffective attempt to address the task. The message may be limited to the point of being unintelligible. A typical response exhibits one or more of the following: • Very little elaboration, if any • Telegraphic language (i.e., short and/or disconnected phrases and sentences) with a very limited range of vocabulary • Serious and frequent errors in the use of language • Minimal original language; any coherent language is mostly borrowed from the stimulus
0	**The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt or consists of arbitrary keystrokes.**

Figure A2. Rubric for the *Write for an Academic Discussion* Task Type

Score	Description
5	**A fully successful response** The response is a relevant and very clearly expressed contribution to the online discussion, and it demonstrates consistent facility in the use of language. A typical response displays the following: • Relevant and well-elaborated explanations, exemplifications and/or details • Effective use of a variety of syntactic structures and precise, idiomatic word choice • Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there/their*)
4	**A generally successful response** The response is a relevant contribution to the online discussion, and facility in the use of language allows the writer's ideas to be easily understood. A typical response displays the following: • Relevant and adequately elaborated explanations, exemplifications and/or details • A variety of syntactic structures and appropriate word choice • Few lexical or grammatical errors
3	**A partially successful response** The response is a mostly relevant and mostly understandable contribution to the online discussion, and there is some facility in the use of language. A typical response displays the following: • Elaboration in which part of an explanation, example or detail may be missing, unclear or irrelevant • Some variety in syntactic structures and a range of vocabulary • Some noticeable lexical and grammatical errors in sentence structure, word form or use of idiomatic language
2	**A mostly unsuccessful response** The response reflects an attempt to contribute to the online discussion, but limitations in the use of language may make ideas hard to follow. A typical response displays the following: • Ideas that may be poorly elaborated or only partially relevant • A limited range of syntactic structures and vocabulary • An accumulation of errors in sentence structure, word forms or use
1	**An unsuccessful response** The response reflects an ineffective attempt to contribute to the online discussion, and limitations in the use of language may prevent the expression of ideas. A typical response displays the following: • Words and phrases that indicate an attempt to address the task, but with few or no coherent ideas • Severely limited range of syntactic structures and vocabulary • Serious and frequent errors in the use of language • Minimal original language; any coherent language is mostly borrowed from the stimulus
0	**The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt or consists of arbitrary keystrokes.**

Figure A3. Rubric for the *Listen and Repeat* Task Type

Score	Description
5	**The response exactly repeats the prompt.** A typical response exhibits the following: • The response is fully intelligible and is an exact repetition of the prompt.
4	**The response captures the meaning expressed in the prompt, but it is not an exact repetition.** A typical response exhibits the following: • Minor changes in words or grammar are present that do not substantially change the meaning of the prompt. For example: • one or two function words may be missing or changed, • a content word may be missing (in longer stimuli) or replaced with a related word, • markers of tense/aspect/number may be missing or incorrect, or • two words may be transposed. • One or two content words may be ambiguous because of imprecise pronunciation. The speaker may self-correct, but successfully completes the response.
3	**The response is essentially full, but it does not accurately capture the original meaning.** A typical response exhibits the following: • The response contains a majority of the content words or ideas in the prompt. • Multiple function words may be changed or missing; one or more content words may be missing or substantively changed. • The response is a full sentence. • In some cases, intelligibility issues cause occasional difficulty in understanding meaning. The speaker may struggle over a word or phrase or run words together, reducing intelligibility.
2	**The response is missing a significant part of the prompt and/or is highly inaccurate.** A typical response exhibits the following: • A large portion of the prompt is missing, and important original meaning is left out. • The speaker may repeat the first part of the sentence. Then the speaker may stop or fill with inaccurate content and/or include the last few words. • The response is not a self-standing sentence; meaning is fragmentary. • Intelligibility is low; the response would be difficult to understand for a listener unfamiliar with the prompt.
1	**The response captures very little of the prompt or is largely unintelligible.** A typical response exhibits the following: • A minimal response of a few words is made; most of the prompt is missing. • The response is recognizable as an attempt to repeat the prompt, but it is mostly unintelligible.
0	**No response OR the response is entirely unintelligible OR there is no English in the response OR the content is entirely unconnected to the prompt (or consists only of phrases such as "I don't know").**

Figure A4. Rubric for the *Take an Interview* Task Type

Score	Description
5	**A fully successful response** The response fully addresses the question, and it is clear and fluent. A typical response exhibits the following: • The response is on topic and well elaborated. • Good conversational speaking pace is maintained with appropriate and natural use of pauses. • Pronunciation is easily intelligible; rhythm and intonation effectively convey meaning. • A range of accurate grammar and vocabulary allows clear expression of precise meanings.
4	**A generally successful response** The response addresses the question, and it is reasonably clear. A typical response exhibits the following: • The response is on topic and elaborated, but it may lack effective sentence-level connectors. • Good speaking pace is generally maintained, with some pausing that may minimally affect flow. • Intelligibility and meaning are not impeded by pronunciation, rhythm and intonation, although occasional words/phrases may require minor effort to understand. • Grammar and vocabulary are adequate to express general meanings most of the time.
3	**A partially successful response** The response addresses the question but with limited elaboration and/or clarity. A typical response exhibits the following: • The response is generally on topic, but elaboration may be relatively limited. • Frequent or lengthy pauses result in a choppy pace; filler words are frequent. • Intelligibility is sometimes affected by inaccuracies in word-level pronunciation or stress/rhythm. • Limited range and accuracy of grammar and vocabulary noticeably restrict the precision and clarity of meanings.
2	**A mostly unsuccessful response** The response reflects an attempt to address the question, but it is not supported in a meaningful and/or intelligible way. A typical response exhibits the following: • The response is minimally connected to the interviewer's question, but it has little or no relevant elaboration or consists mainly of language from the question. • Intelligibility is limited; the speaker's intended meaning is often difficult to discern. • The response shows a very limited range of grammar and vocabulary.
1	**An unsuccessful response** The response minimally addresses the question, and it may demonstrate very limited control of language. A typical response exhibits the following: • The response is only vaguely connected to language in the interviewer's question. • The response is mostly unintelligible. • The response consists mainly of isolated words or phrases
0	**No response OR the response is entirely unintelligible OR there is no English in the response OR the content is entirely unconnected to the prompt (or consists only of phrases such as "I don't know").**

Appendix B. Research Related to Test Design and Score Interpretation

Comparing *Write for Academic Discussion* and *Independent Writing* tasks

Davis, L., & Norris, J. M. (2023). *A comparison of two TOEFL® writing tasks* (Research Memorandum No. RM-23-06). ETS. https://www.ets.org/Media/Research/pdf/RM-23-06.pdf

Writing section reliability

Gu, L., Li, S., Li, T., & Norris, J. M. (2023). Maintaining score quality on the enhanced TOEFL iBT® test (Research Memorandum No. RM-23-05). ETS. https://www.ets.org/Media/Research/pdf/RM-23-05.pdf

Distinguishing proficiency levels in English language programs

Norris, J. M., & Lee, J. (2023). *The effectiveness of the TOEFL® Essentials™ test for distinguishing English proficiency levels* (Research Memorandum No. RM-23-07). ETS. https://www.ets.org/Media/Research/pdf/RM-23-07.pdf

Balancing construct coverage and efficiency

Davis, L., Norris, J., Papageorgiou, S., & Sasayama, S. (2023). Balancing construct coverage and efficiency: Test design and validation considerations for a remote-proctored online language test. In K. Sadeghi & D. Douglas (Eds.), *Fundamental Considerations in Technology Mediated Language Assessment* (pp. 49–63). Routledge.

Writing tasks for lower proficiency levels

Sasayama, S., Garcia Gomez, P., & Norris, J. M. (2021). *Designing efficient L2 writing assessment tasks for low-proficiency learners of English* (TOEFL Research Report No. 97). ETS. https://doi.org/10.1002/ets2.12341

Mapping to CEFR levels (speaking and writing)

Davis, L., Garcia Gomez, P., Li, S., Manna, V. (2023). Mapping TOEFL Essentials Speaking and Writing Scores to the CEFR Levels. In S. Papageorgiou & V. Manna (Eds.), *Meaningful language test scores: Research to enhance score interpretation* (pp. 120–140). John Benjamins. https://doi.org/10.1075/illa.1.07dav

Mapping to Canadian Language Benchmarks

Papageorgiou, S., Davis, L., Ohta, R., & Gomez, G. G. (2022). *Mapping TOEFL® Essentials™ test scores to the Canadian Language Benchmarks* (TOEFL Research Report No. RR-100). ETS. https://doi.org/10.1002/ets2.12357

Synthetically generated speech

Choi, I., & Zu, J. (2022). *The Impact of using synthetically generated listening stimuli on test taker performance: A case study with multiple-choice, single-selection items* (TOEFL Research Report No. RR-98). ETS. https://doi.org/10.1002/ets2.12347

Development of elicited imitation task

Davis, L., & Norris, J. (2021). *Developing an innovative elicited imitation task for efficient English proficiency assessmen*t (TOEFL Research Report No. 96). ETS. https://doi.org/10.1002/ets2.12338

Suggested Citation

Manna, V. F., Li, S., Papageorgiou, S., & Gu, L. (2025). *TOEFL iBT® technical manual* (TOEFL Research Report No. RR-106). ETS.

Action Editor: Jonathan Schmidgall

Reviewers: Ching-Ni Hsieh and Jiyun Zu

Find other ETS-published reports by searching the ETS ReSEARCHER database.