Research Memorandum
ETS RM–21-03

Design Framework for the *TOEFL® Essentials™* Test 2021

Spiros Papageorgiou
Larry Davis
John M. Norris
Pablo Garcia Gomez
Venessa F. Manna
Lora Monfils

May 2021

ETS Research Memorandum 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 Memorandum 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 Memorandum 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.

Design Framework for the *TOEFL® Essentials™* Test 2021

Spiros Papageorgiou, Larry Davis, John M. Norris, Pablo Garcia Gomez, Venessa F. Manna, and Lora Monfils
ETS, Princeton, New Jersey, United States

May 2021

Corresponding author: S. Papageorgiou, Email: spapageorgiou@ets.org

Suggested citation: 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.

Action Editor: Brent Bridgeman

Reviewers: Jonathan Schmidgall and Rick Tannenbaum

Abstract

The *TOEFL® Essentials*™ test is a new English language proficiency test in the *TOEFL®* family of assessments. It measures foundational language skills and communication abilities in academic and general (daily life) contexts. The test covers the four language skills of reading, listening, writing, and speaking and is intended to provide academic programs and other users with reliable information regarding the test taker's ability to understand and use English. This report presents the theoretical and empirical basis underlying the development of the TOEFL Essentials test. 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 report 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.

Keywords: English language proficiency, language assessment, language test design, language test validation, online testing, *TOEFL® Essentials*™ test

Acknowledgments

We thank Phil Everson and Danielle Guzman-Orth for their help with earlier versions of the Test Design Process section. We also thank Srikant Gopal, Jennifer Brown, and Lydia Liu for reviewing earlier versions of the manuscript. Special thanks to our ETS colleagues Brent Bridgeman, Rick Tannenbaum, and Jonathan Schmidgall for their careful review of an earlier version of the manuscript and their helpful comments.

Table of Contents

General Description of the *TOEFL® Essentials™* Test

The *TOEFL® Essentials™* test is a new English language proficiency test in the *TOEFL®* family of assessments. It measures foundational language skills and communication abilities in academic and general (daily life) contexts. The test covers the four language skills of reading, listening, writing, and speaking and is intended to provide academic programs and other users with reliable information regarding the test taker's ability to understand and use English.

An optimal combination of convenience and quality is a major goal of the TOEFL Essentials test. It can be taken at home, requires approximately 90 minutes to complete, and unofficial scores for the TOEFL Essentials Listening and Reading sections are available at the end of the test session with official scores available in 6 days. Test security during the administration of the test is provided by trained human proctors who monitor the entire testing session. Proctors are assisted by artificial intelligence (AI) technology, which 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.

The TOEFL Essentials test is designed for efficient measurement of both foundational aspects of language proficiency (lexical and grammatical competence) as well as the ability to communicate in English through a range of language knowledge activities and communicative language tasks. 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 skills. Some examples of communicative language tasks represented in the test include

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

The TOEFL Essentials test is designed to be suitable for language learners across a wide range of abilities. It uses multistage adaptive test (MST) methodology to help ensure the most accurate measure of the test taker's language ability in an efficient way. Based on this methodology, test-taker performance on the first part of a test section is used to select the content for the second part of the section so that the difficulty of the test tasks matches the ability level of the test taker. Tailoring test content to a test taker's ability level supports the accuracy of the scores with reduced administration time.

Overall, the TOEFL Essentials test is designed to provide valid and reliable information about someone's ability to use English in a relatively brief test-taking time and at an affordable price using a format that is easy to access and engaging.

The purpose and intended uses of the test, its target test-taker population, and relevant language use domains are described first in this report. 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 paper 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. It will also be updated periodically to include reference to research studies currently in preparation for publication in various outlets.

Target Population, Language Domain, and Intended Uses

The TOEFL Essentials 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 contexts. The 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. Through the use of MST methodology, the TOEFL Essentials 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 Essentials test is designed to cover the full range from A1 to C2 (see the Scoring section).

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 Essentials 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.

Based on input from extensive market research (nearly 250 score users from institutions in the United States, Canada, and the United Kingdom and 7,200 test takers around the world), a need was identified for a language proficiency test that is affordable and convenient to access. Accordingly, the TOEFL Essentials test is designed to provide academic programs and other scores users with valid and reliable information about someone's ability to use English in a relatively brief test-taking time and at an affordable price using a format that is intended to be test-taker friendly and engaging. Recommended uses of the TOEFL Essentials 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.

Construct Definition

In light of the intended uses and administration requirements for the TOEFL Essentials test discussed 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 settings. 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 Essentials 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 TOEFL Essentials 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 solving a range of communication challenges that represent English as it is actually 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 test measures the ability to listen to and comprehend both conversational and extended monologic (e.g., lecture) speech. It measures the ability to read and comprehend information presented in a variety of formats, including short informational graphics as well as extended passages. It measures the ability to write effectively in common genres such as describing a scene, writing an email, and responding to an academic discussion. It 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 Essentials test focuses on predicting overall English ability and discerning the likelihood that learners can accomplish real-life English communication tasks.

Test Design Process

The design of the test was the result of collaboration among researchers, content developers, psychometricians, and business directors of the TOEFL program. The process of designing test tasks for the TOEFL Essentials test began with discussions of the requirements that were necessary to make the final product useful to score users and language learners based on feedback from the multiple market research studies with institutions and test takers around the world, as mentioned previously. Requirements that influenced the design of the test included

- measure and report scores for all four language skills: reading, listening, speaking, and writing;
- measure a wide range of abilities, from novice to advanced users of English (CEFR A1 to C2 levels);
- measure language ability in the academic and general (daily life) contexts;
- offer content that reflects use of the English language beyond North American contexts;
- time required to complete the full test should last no more than 90 minutes; and
- can be completed online, at home, from the test taker's own computer, with administration in test centers being a possibility in the future.

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. The test was designed to balance these demands by employing an efficient test administration model (MST methodology) as well as by combining task types addressing both foundational language abilities and communication skills. Tasks measuring foundational abilities, such as knowledge of sentence word order 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 combined with tasks that require the test taker to understand spoken or written input or produce spoken or written responses. The combination of these task types represents the hybrid approach to construct operationalization mentioned previously, 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.

Taking these requirements into account, the designers of the TOEFL Essentials test first created prototype speaking and writing tasks. Initial efforts 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 speaking and writing 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.

A pilot administration was organized next. It included the refined speaking and writing tasks and listening and reading tasks adapted for rapid assessment of language proficiency. The pilot administration included a population of English learners from diverse regions of the world (*N* = 700). Both the prototype administration and the pilot administration included more task types than were needed for the final test design. Based on the results of the pilot administration, a subset of the best performing task types was selected for the operational test design and specifications for those tasks were refined.

The final step in operational test design was the field testing of a pool of items on a population that was similar to the expected operational population and of sufficient size to produce stable item statistics (*N* ≈ 5,000). The field test pool was intended to support the first administrations of the operational test.

A core design principle of the TOEFL Essentials 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, relatively affordable cost and at-home delivery is expected to increase access to the test compared to traditional test delivery through test centers. 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 towards specific test-taker groups identified on the basis of gender and first language.

Test takers also have open access to the TOEFL Essentials official practice tests. Using the practice tests, test takers have the opportunity to become familiar with test navigation as well as the listening, reading, writing, and speaking tasks prior to test administration. Additionally, test takers with documented disabilities or health-related needs, who may need reasonable accommodations to demonstrate their English skills in reading, listening, writing, and speaking, can confidentially request and select accommodations prior to registration. If approved, test takers can register for select accommodations from their ETS account, including extended time, extra breaks, screen magnification, and selectable colors. If there is a need to request other accommodations for disability or health-related needs, test takers must register through ETS Disability Services.

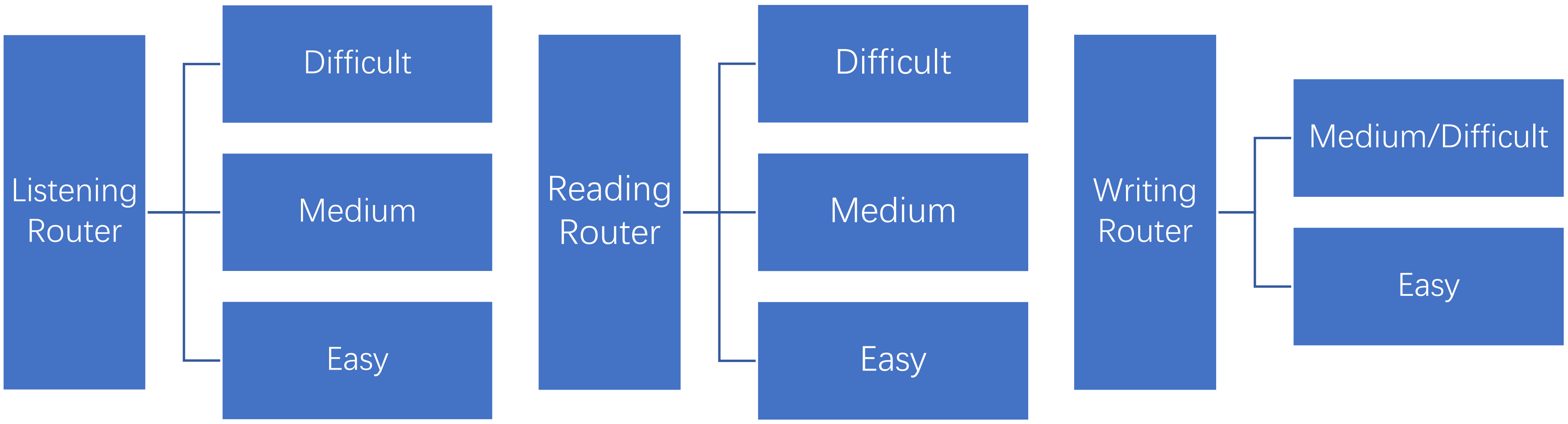
Multistage Adaptive Test Design

To provide for efficient measurement of language proficiency, the TOEFL Essentials Listening, Reading, and Writing sections are designed as section-level MSTs. The first part (stage) of a test section contains tasks of average difficulty. A second part, with a difficulty level dependent on the test taker's performance on the first part, follows. For example, if the student does very well on the first part of the listening section, the second part of the listening section will be at a higher level of difficulty. The scoring for the listening, reading, and writing sections takes into consideration the total number of questions answered correctly across the two parts as well as the difficulty level of these parts. The TOEFL Essentials Speaking section is linear. Speaking tasks are designed to be accessible across a range of proficiency levels with many opportunities for the test taker to speak. 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 speaking section are based on overall performance on all tasks.

The MST design for the TOEFL Essentials test is presented in Figure 1. The listening, reading, and writing sections of the test consist of two stages. Test takers first respond to the questions in the first stage (often referred to as a *router*). Based on how well they respond to these questions, test takers then encounter content appropriate to their ability in the second stage of the section. Content in the second stage of the reading and listening sections is classified as low, medium, or high difficulty. Content in the second stage of the writing section is classified as low or medium/high difficulty. It was expected that items in the medium/high difficulty second stage would be accessible to individuals across a broad range of proficiency with the scoring rubric providing for differentiation between medium and high proficiency levels. However, there was concern that these tasks might be overly challenging for test takers at beginning levels. Accordingly, low-difficulty content for the second stage of the writing section was created where the linguistic complexity of the input was reduced and communicative demands were simplified. For listening and reading, the first and second stages include all task types as described in next section. For writing, the first stage is composed of dichotomously scored tasks, whereas the second stage is composed of constructed response tasks that require rater scoring.

The MST design was the preferred solution for the TOEFL Essentials test because it combines the advantages of adaptive and linear test designs (Hendrickson, 2007). By employing MST methodology, the TOEFL Essentials 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 the individual item level, the test is able to operationalize the task-based approach in test design that underpins the design of all tests in the TOEFL family of assessments. In addition, section-level adaptation allows the test content to be assembled into multitask panels reflecting distinct levels of difficulty with expert assessment specialists' review of test content before administration. In other words, the MST methodology allows the TOEFL Essentials test to deliver relevant test content, including robust communication tasks, for its intended purposes in a targeted and efficient way.

Figure 1. TOEFL Essentials Multistage Adaptive Test Methodology



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 level, and the majority of them have taught at K–12 schools, colleges, or universities internationally. Many 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. ETS also carefully selects and trains outside item writers (who have experience teaching English as a second or foreign language or other academic content areas) to develop an initial draft of test questions that are then reviewed by assessment specialists. ETS considers item writers' experience and backgrounds so that the pool of item writers reflects, to the greatest degree possible, the diversity of the international test-taking population.

Content Writing and Reviewing

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. Three or more assessment specialists who have not participated in the authoring stage sequentially and independently review each stimulus and its associated items. They may suggest revising a stimulus or an associated item or rejecting an item or a stimulus entirely. Stimuli and items only become eligible for use in a test if all reviewers judge them to be acceptable. This linear peer review process includes discussion between and among reviewers at each of the review stages. Additionally, when required for a given test stimulus or item, a subject matter expert checks the accuracy and currency of the content in the stimulus. For some task types, ETS assessment specialists also use a proprietary technological capability, called Technology Assisted Item Creation (TAIC), to facilitate the content development process. TAIC integrates task content specifications and difficulty parameters specifically developed for the TOEFL Essentials test. After the task content is generated through TAIC, it undergoes the rigorous, multistage review process described previously.

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

- Is the language in the test materials clear? Is it accessible to 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 item specifications, the uniqueness of the answer or answers (the item keys), the clarity and accessibility of the language used, and the plausibility and attractiveness of the distracters—the incorrect options. For constructed response tasks (speaking and writing), 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 speaking or writing task is acceptable and can be included in an operational test (see also discussion of tryouts for constructed response tasks in the Typical Test Review Chronology section).

The *ETS Standards for Quality and Fairness* (ETS, 2014) mandate fairness reviews. This fairness review must take place before using materials in a test. All assessment specialists undergo fairness training—in addition to item-writing training—soon after their arrival at ETS. As part of their training, item writers become familiar with the *ETS Guidelines for 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 items. Fairness issues are thus considered at each stage of the development process.

All TOEFL Essentials 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).

Typical Test Review Chronology

The chronology of a typical review chain is as follows:

1. First content review
2. Second content review
3. Editorial review
4. Fairness review
5. Final content review

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

Pretest reading, listening, sentence construction, and vocabulary questions are included in operational test forms, and data are collected on real TOEFL Essentials test takers' ability to answer the questions. Test takers cannot identify pretest questions because they do not differ in any distinguishable way from the operational (scored) questions on the test. Pretesting items allows assessment specialists to identify poorly functioning questions and revise them or exclude them from the operational pool.

For the constructed response sections, ETS conducts small-scale tryouts of selected speaking and writing prompts (the questions defining the tasks for the test takers) among members of the target population. Assessment specialists review and evaluate spoken or written responses to these tryout questions. These specialists use expert judgment to determine which prompts are likely to elicit scorable responses from test takers across the range of proficiency levels; these viable prompts are the ones that appear in operational test forms.

After assessment specialists approve test tasks that have been pretested (in the case of reading and listening sections) or successfully tried out (in the case of speaking and writing sections), 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.

Test Tasks

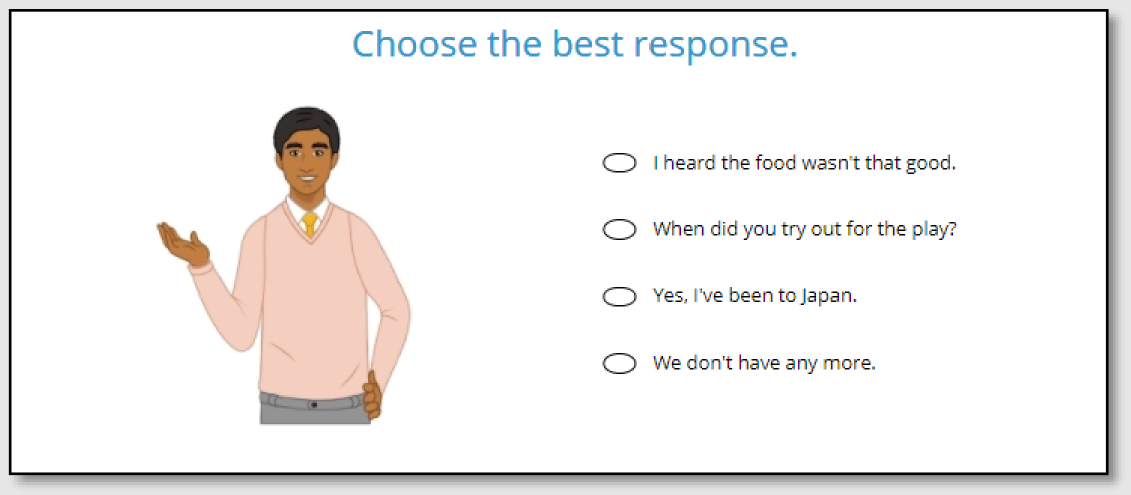
Listening Section

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 Reply, Listen to a Conversation, Academic Listening: Announcements,* and *Academic Listening: Talks.*

The *Listen and Reply* 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 2). 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 2. Example of *Listen and Reply* Task Type



Choose the best response.

○ I heard the food wasn't that good.

○ When did you try out for the play?

○ Yes, I've been to Japan.

○ We don't have any more.

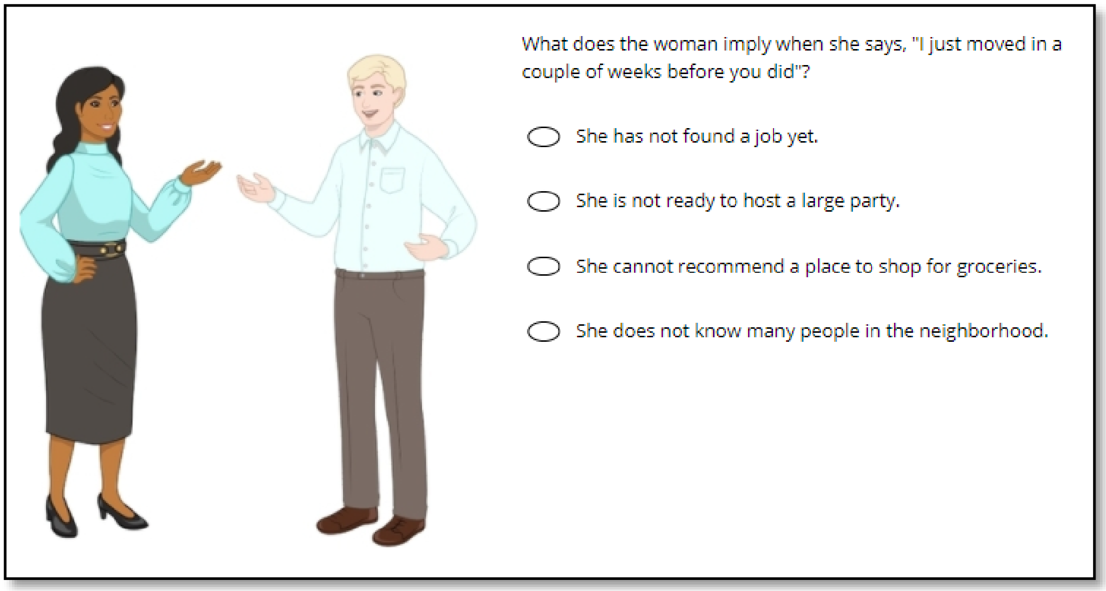
Note. Test takers hear: "How about trying out that new Japanese restaurant?"

The *Listen to a Conversation* task (see Figure 3) 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 two 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 3. Example of *Listen to a Conversation* Task Type



What does the woman imply when she says, "I just moved in a couple of weeks before you did"?

○ She has not found a job yet.

○ She is not ready to host a large party.

○ She cannot recommend a place to shop for groceries.

○ She does not know many people in the neighborhood.

Note. Test takers hear:

Woman: Thanks for inviting me to your barbecue this weekend. Should I bring anything? A salad? Dessert?

Man: Thanks for the offer, Janet, but everything is taken care of. However, there is one thing you might be able to help me with.

Woman: Sure. What's up?

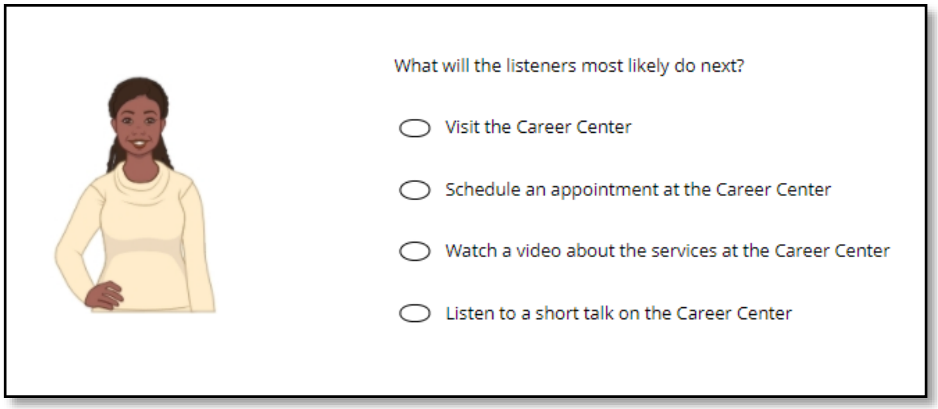
Man: Well, I've only been in the neighborhood for a few months, and I don't know that many people. Maybe you could help me with the guest list?

Woman: I don't think I'm the best person to ask. I just moved in a couple of weeks before you did!

The *Academic Listening: Announcements* 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 4). 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 4. Example of *Academic Listening: Announcements* Task Type



What will the listeners most likely do next?

○ Visit the Career Center

○ Schedule an appointment at the Career Center

○ Watch a video about the services at the Career Center

○ Listen to a short talk on the Career Center

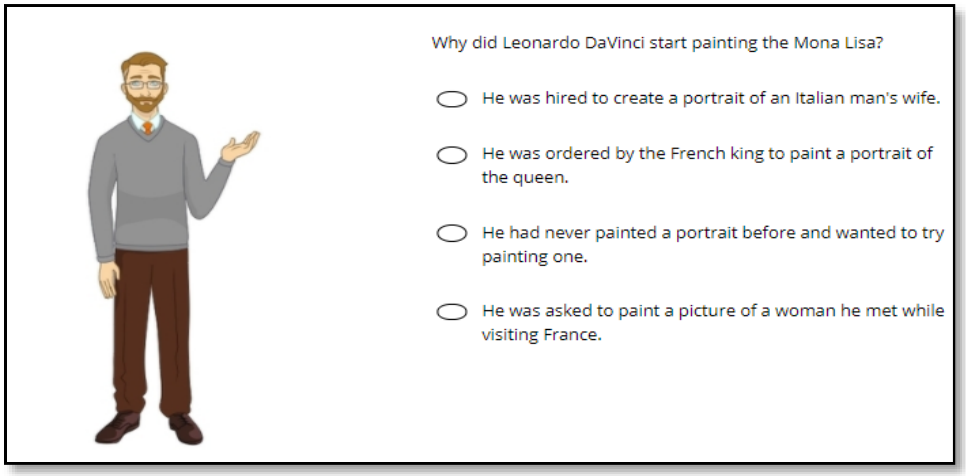
Note. Test takers hear:

Hello, everyone. I'm Jennifer Wilson from the Career Center. Thank you all for attending this résumé-building workshop. Today, my colleagues and I will cover several useful strategies on how to make your résumé stand out to potential employers. First, I want to give the floor to my colleague Pierre Moreau, who will go over the Career Center's services, such as career advising, help with internships, and one-on-one appointments.

The *Academic Listening: Talks* task is designed to simulate academic talks given by educators (see Figure 5). The test taker listens to a short (100–250 words) academic-related talk and answers two to four 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 5. Example of *Academic Listening: Talks* Task Type



Why did Leonardo DaVinci start painting the Mona Lisa?

○ He was hired to create a portrait of an Italian man's wife.

○ He was ordered by the French king to paint a portrait of the queen.

○ He had never painted a portrait before and wanted to try painting one.

○ He was asked to paint a picture of a woman he met while visiting France.

Note. Test takers hear:

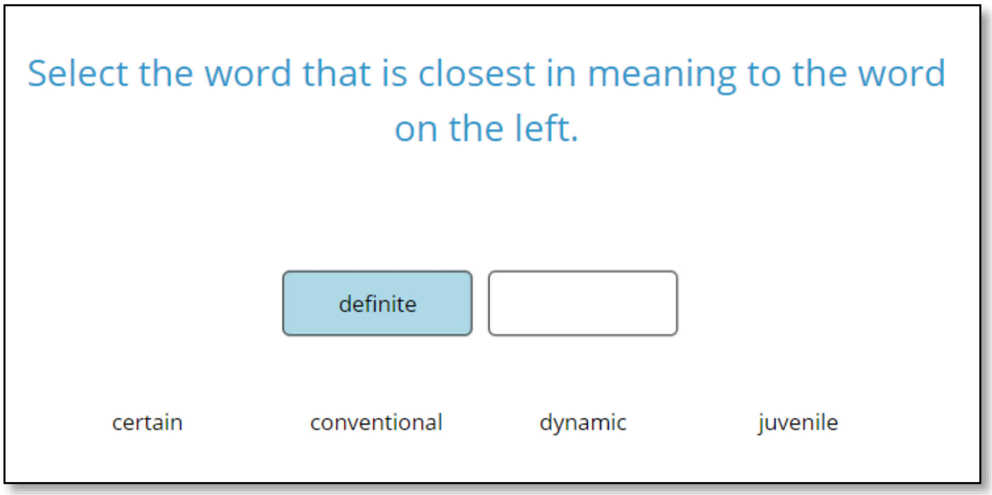
You've probably heard of the portrait called Mona Lisa by Leonardo DaVinci. It's one of the most famous paintings in the world. What you may not know is the interesting story behind the painting. Back in the 1500s, a wealthy Italian man hired DaVinci to create a painting of his wife. But instead of giving the portrait to the man as promised, DaVinci moved to France before the painting was finished, and he took it with him. DaVinci started working for the French king, who decided he wanted to buy the painting, and likely for more money. DaVinci agreed. French rulers kept the painting for centuries, until it was moved to the Louvre Museum in Paris. Then in 1911, the painting was stolen. A museum employee put the Mona Lisa under his shirt one night and just walked out! Why? He wanted to return the painting to Italy, the country of its origin. During the two years the painting was missing, newspapers around the world reprinted pictures of it, along with articles about the search. All the publicity created global interest in the painting. The Mona Lisa was eventually returned to the museum, but a number of conspiracy theories cropped up. Was the theft planned as an attempt to draw interest in the painting? Did the employee make a copy of the painting and keep the original for himself? Whatever the truth is, the Mona Lisa remains one of the most famous paintings in the world.

Reading Section

Around the world people learn from academic texts and other academic materials in English. In their daily lives, people also need to navigate the variety of everyday reading material they encounter: compressed, informational texts (such as receipts, schedules, signs, menus) as well as more expanded informal texts (webpages, magazine and news articles, text messages, emails). Reading questions measure the test taker's ability to understand both academic and nonacademic texts from English language contexts around the world. Reading skills are measured with the following task types: *Vocabulary, General Reading: Daily Life, Academic Reading: Tables.* and *Academic Reading: Passages.*

The *Vocabulary* task serves as a simple, efficient, fast-paced indicator of reading ability. The task contains several questions, for which the test taker sees a word and chooses the most similar word from four options (see Figure 6). These questions measure the test taker's knowledge of English vocabulary by their ability to correctly choose the closest synonym. Tested words range from fairly common to less common words.

Figure 6. Example of *Vocabulary* Task Type



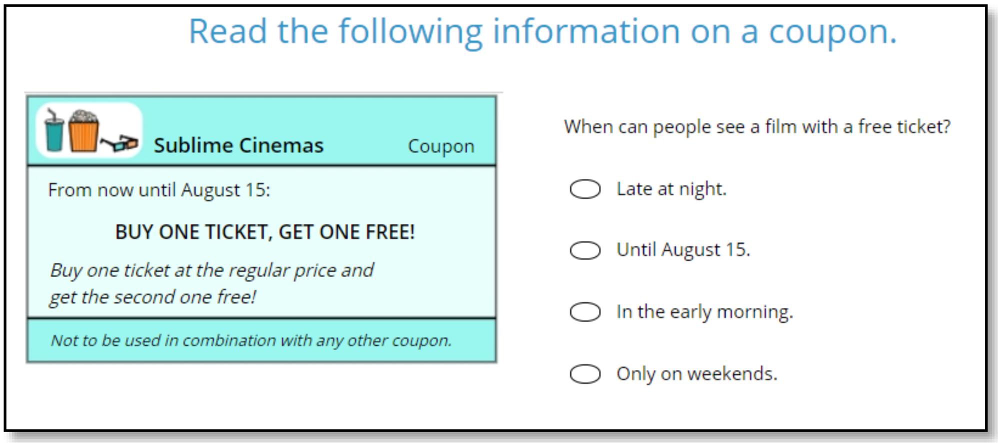
Select the word that is closest in meaning to the word on the left.

The *General Reading: Daily Life* task includes short, nonacademic texts commonly encountered in daily life around the world (see Figure 7). 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, three, or four 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 7. Examples of the *General Reading: Daily Life* Task Type



Read the following information on a coupon.

Sublime Cinemas Coupon

From now until August 15:

BUY ONE TICKET, GET ONE FREE!

Buy one ticket at the regular price and get the second one free!

Not to be used in combination with any other coupon.

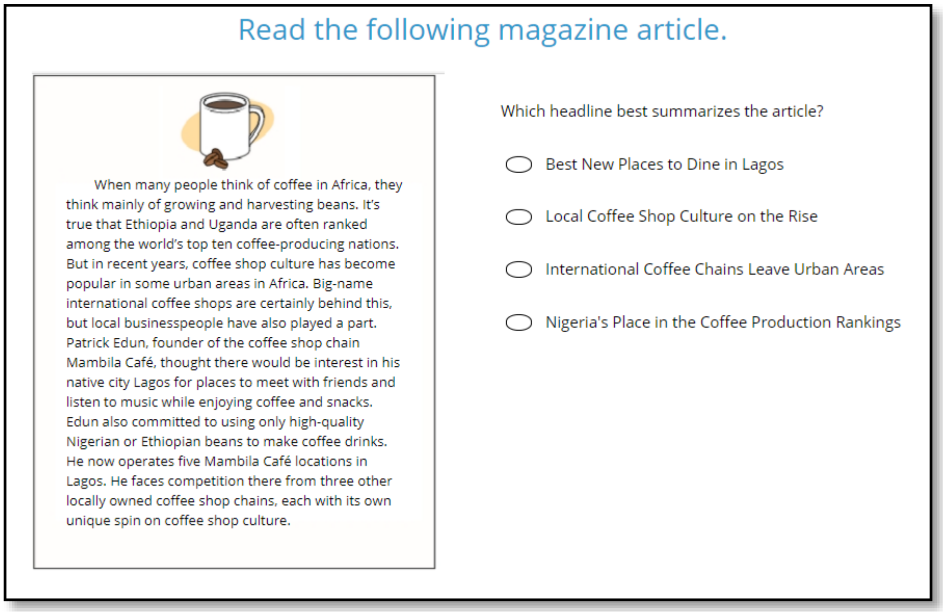
When can people see a film with a free ticket?

○ Late at night.

○ Until August 15.

○ In the early morning.

○ Only on weekends.



Read the following magazine article.

When many people think of coffee in Africa, they think mainly of growing and harvesting beans. It's true that Ethiopia and Uganda are often ranked among the world's top ten coffee-producing nations. But in recent years, coffee shop culture has become popular in some urban areas in Africa. Big-name international coffee shops are certainly behind this, but local businesspeople have also played a part. Patrick Edun, founder of the coffee shop chain Mambila Café, thought there would be interest in his native city Lagos for places to meet with friends and listen to music while enjoying coffee and snacks. Edun also committed to using only high-quality Nigerian or Ethiopian beans to make coffee drinks. He now operates five Mambila Café locations in Lagos. He faces competition there from three other locally owned coffee shop chains, each with its own unique spin on coffee shop culture.

Which headline best summarizes the article?

○ Best New Places to Dine in Lagos

○ Local Coffee Shop Culture on the Rise

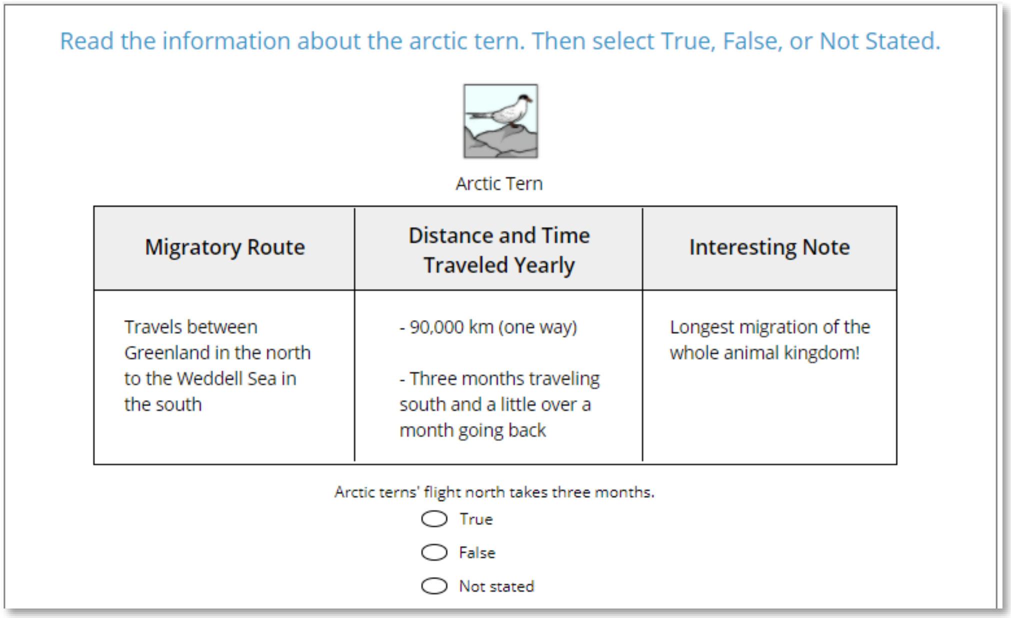
○ International Coffee Chains Leave Urban Areas

○ Nigeria's Place in the Coffee Production Rankings

The *Academic Reading: Tables* task includes tables with two or three categories of information about a wide range of academic topics for readers to learn about (see Figure 8). These tables serve as a summary of key pieces of information typically displayed in different academic sources such as textbooks, the science sections of newspapers and magazines, or websites. The task is designed so that background knowledge is not required. Topics include history, art and music, business and economics, life science, physical science, and social science. The texts in the tables can be anywhere from 50 to 85 words and are accompanied by four multiple-choice questions. Each question is followed by three options that are always the same for all the questions: true, false, and not stated. “Not stated” means that the information is not present in the text. The questions require test takers to

- understand information presented in nonlinear texts,
- understand important facts and details,
- infer meaning from information that is not explicitly stated,
- understand a range of academic vocabulary,
- connect information,
- scan to pick up details, and
- skim for main ideas.

Figure 8. Example of *Academic Reading: Tables* Task Type



Read the information about the arctic tern. Then select True, False, or Not Stated.

Arctic Tern

Migratory Route	Distance and Time Traveled Yearly	Interesting Note
Travels between Greenland in the north to the Weddell Sea in the south	- 90,000 km (one way) - Three months traveling south and a little over a month going back	Longest migration of the whole animal kingdom!

Arctic terns' flight north takes three months.

○ True

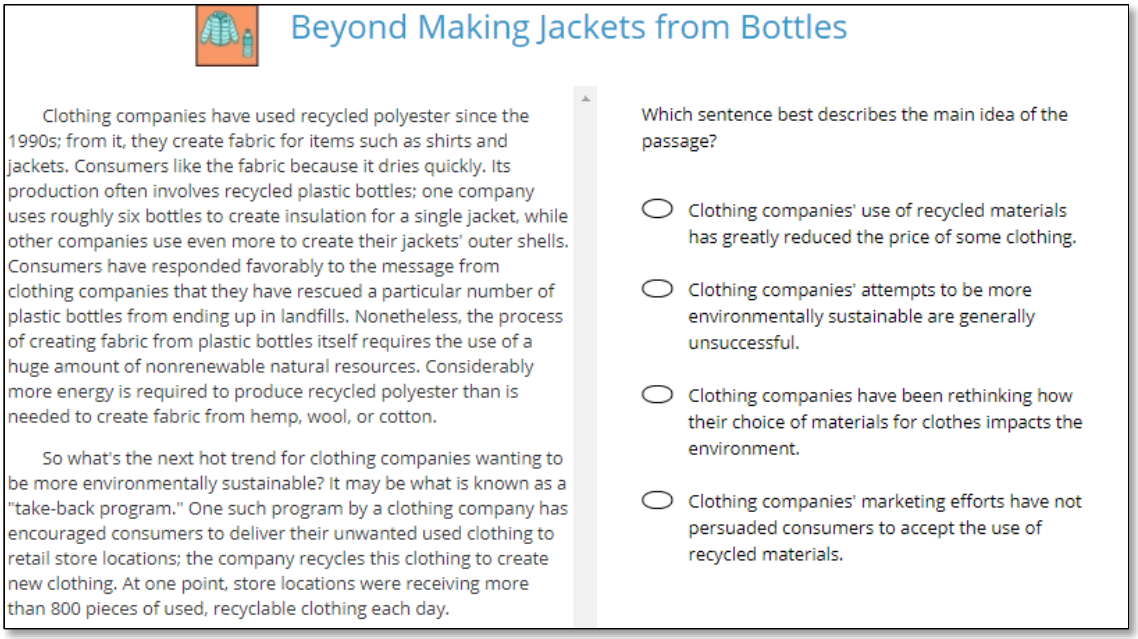
○ False

○ Not stated

The *Academic Reading: Passages* task includes short expository passages typical of those in secondary and higher education (see Figure 9). 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 followed by six questions, which may ask about factual information, vocabulary in context, inferences, relationship between ideas, and 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 9. Example of *Academic Reading: Passages* Task Type



Beyond Making Jackets from Bottles

Clothing companies have used recycled polyester since the 1990s; from it, they create fabric for items such as shirts and jackets. Consumers like the fabric because it dries quickly. Its production often involves recycled plastic bottles; one company uses roughly six bottles to create insulation for a single jacket, while other companies use even more to create their jackets' outer shells. Consumers have responded favorably to the message from clothing companies that they have rescued a particular number of plastic bottles from ending up in landfills. Nonetheless, the process of creating fabric from plastic bottles itself requires the use of a huge amount of nonrenewable natural resources. Considerably more energy is required to produce recycled polyester than is needed to create fabric from hemp, wool, or cotton.

So what's the next hot trend for clothing companies wanting to be more environmentally sustainable? It may be what is known as a "take-back program." One such program by a clothing company has encouraged consumers to deliver their unwanted used clothing to retail store locations; the company recycles this clothing to create new clothing. At one point, store locations were receiving more than 800 pieces of used, recyclable clothing each day.

Which sentence best describes the main idea of the passage?

○ Clothing companies' use of recycled materials has greatly reduced the price of some clothing.

○ Clothing companies' attempts to be more environmentally sustainable are generally unsuccessful.

○ Clothing companies have been rethinking how their choice of materials for clothes impacts the environment.

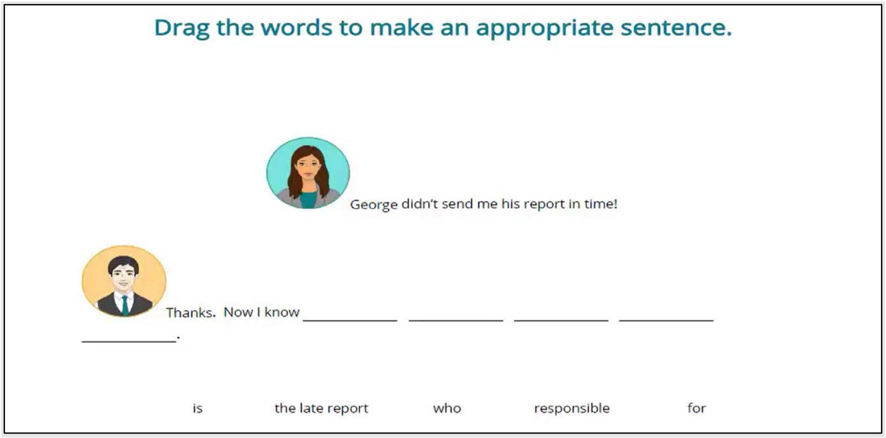
○ Clothing companies' marketing efforts have not persuaded consumers to accept the use of recycled materials.

Writing Section

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, Describe a Photo, Write an Email,* and *Write for an Academic Discussion.*

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 10). This task measures the test taker's command of sentence structures, a skill that is essential for all written communication.

Figure 10. Example of *Build a Sentence* Task Type



Drag the words to make an appropriate sentence.

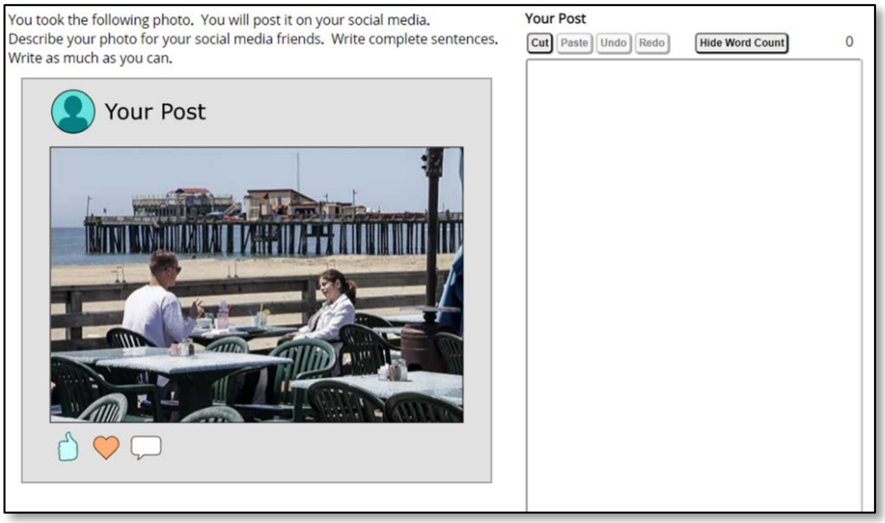
George didn't send me his report in time!

Thanks. Now I know ___________ ___________ ___________ ___________ ___________.

In the *Describe a Photo* writing task, test takers write a social media post about a photo (see Figure 11). Test takers are asked to describe the photo to their social media friends, and they have 7 minutes to prepare and write their post. This writing task measures the test taker's ability to produce a multisentence description that

- is adequately elaborated, 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 11. Example of *Describe a Photo* Task Type

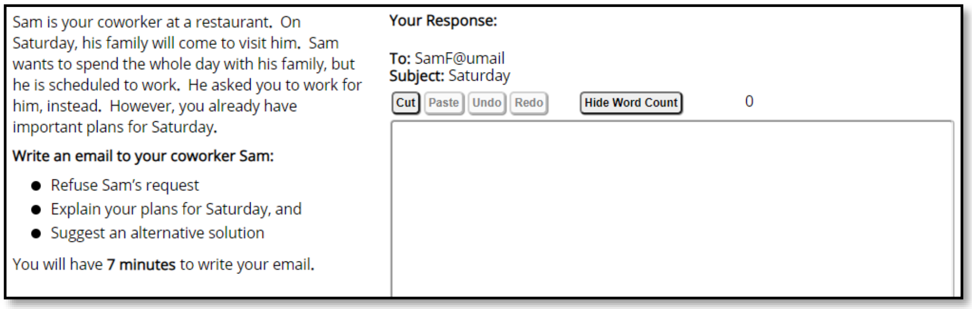


You took the following photo. You will post it on your social media. Describe your photo for your social media friends. Write complete sentences. Write as much as you can.

In the *Write an Email* task, test takers are presented with a scenario in text regarding either an academic or social setting (see Figure 12). 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 12. Example of *Write an Email* Task Type



Sam is your coworker at a restaurant. On Saturday, his family will come to visit him. Sam wants to spend the whole day with his family, but he is scheduled to work. He asked you to work for him, instead. However, you already have important plans for Saturday.

Write an email to your coworker Sam:

- Refuse Sam's request
- Explain your plans for Saturday, and
- Suggest an alternative solution

You will have **7 minutes** to write your email.

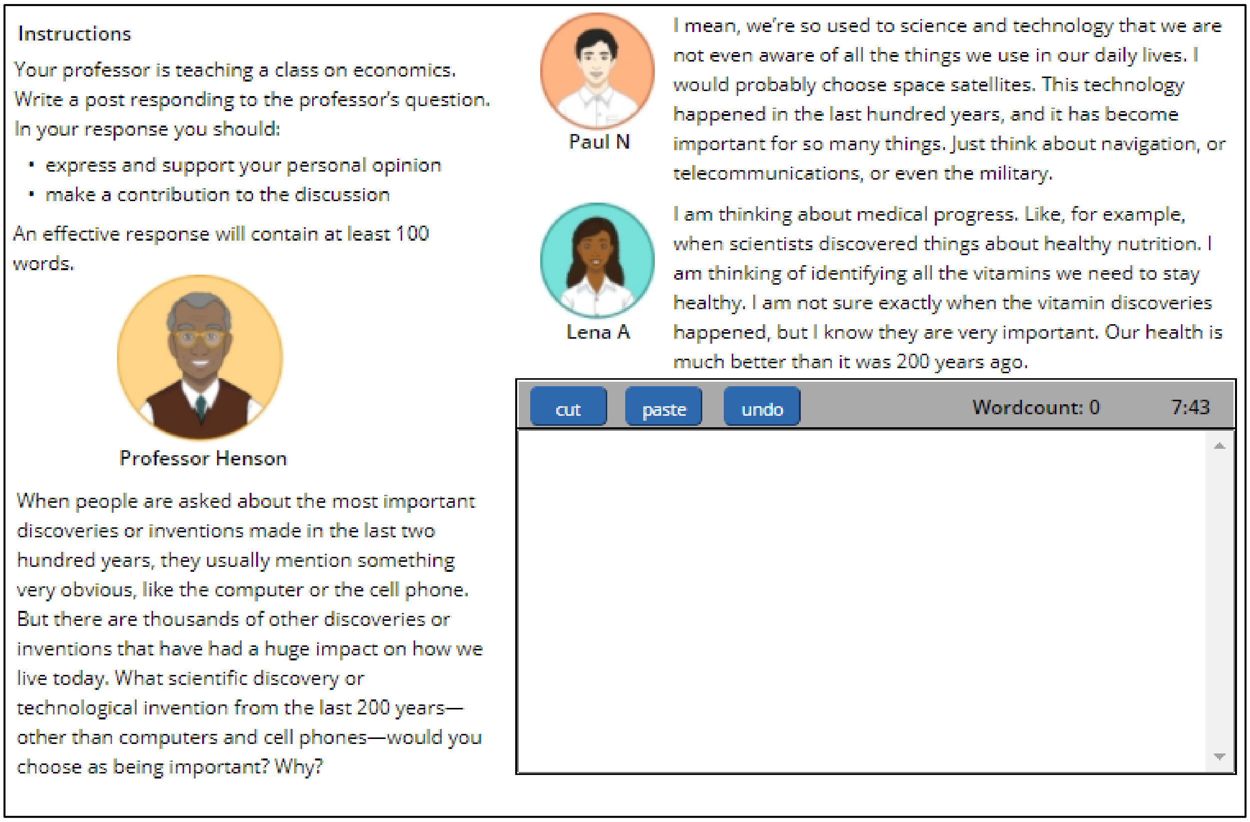
Your Response:

To: SamF@umail
Subject: Saturday

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 13). A post from the professor briefly frames the topic and poses an opinion question related to the topic for the class to discuss. 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, 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 13. Example of *Write for an Academic Discussion* Task Type



Instructions

Your professor is teaching a class on economics. Write a post responding to the professor's question.

In your response you should:

- express and support your personal opinion
- make a contribution to the discussion

An effective response will contain at least 100 words.

Professor Henson

When people are asked about the most important discoveries or inventions made in the last two hundred years, they usually mention something very obvious, like the computer or the cell phone. But there are thousands of other discoveries or inventions that have had a huge impact on how we live today. What scientific discovery or technological invention from the last 200 years—other than computers and cell phones—would you choose as being important? Why?

Paul N

I mean, we're so used to science and technology that we are not even aware of all the things we use in our daily lives. I would probably choose space satellites. This technology happened in the last hundred years, and it has become important for so many things. Just think about navigation, or telecommunications, or even the military.

Lena A

I am thinking about medical progress. Like, for example, when scientists discovered things about healthy nutrition. I am thinking of identifying all the vitamins we need to stay healthy. I am not sure exactly when the vitamin discoveries happened, but I know they are very important. Our health is much better than it was 200 years ago.

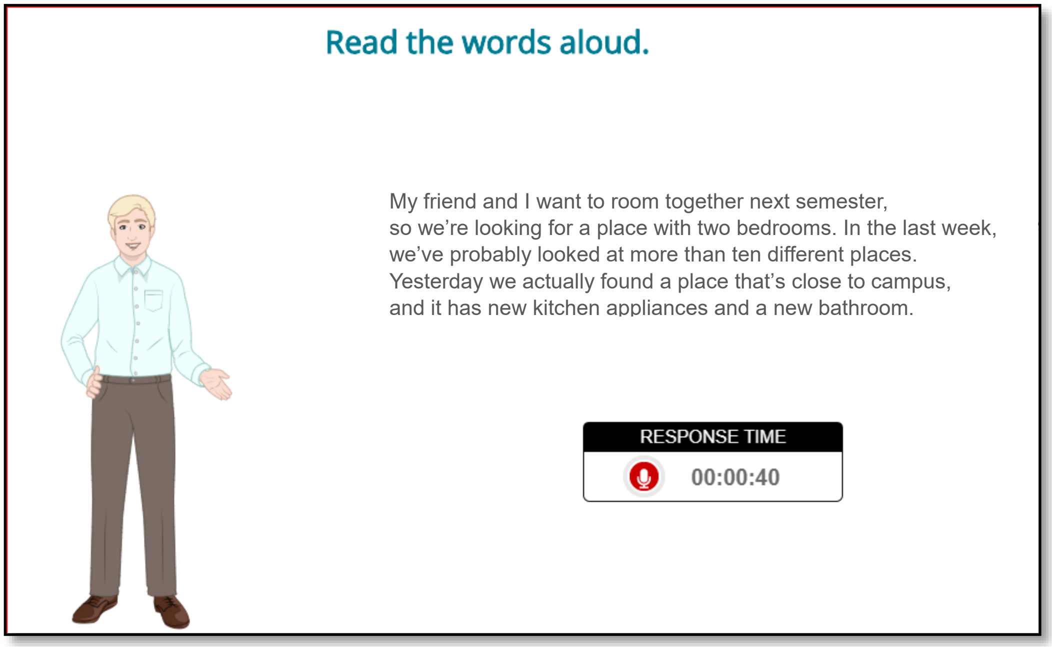
Speaking Section

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 by tasks where test takers reproduce written or spoken input. Communication ability is measured through items where test takers speak about their opinions and experiences in the context of a simulated conversation. Speaking skills are measured with the following task types: *Read Aloud, Listen and Repeat,* and *Virtual Interview.*

The *Read Aloud* task is a directed speaking task in which test takers read aloud one part of a dialogue that takes place in a daily life or campus situation (see Figure 14). Test takers hear a person make statements or ask questions within their part of a conversation. Test takers respond by reading aloud their part of the conversation, which is written on the screen. Each test-taker "turn" in the conversation is around one to three sentences long. Unlike a traditional read aloud task where a stand-alone passage is read in isolation, the simulated interaction provides an overall setting for communication along with a logical sequence for each text, supporting the reading process and providing a context to guide test takers in making appropriate use of emphasis or grouping ideas. This task measures the test taker's ability to process the written text into speech to produce a response that shows

- appropriate pacing with minimal hesitation;
- appropriate use of pausing, sentence stress, and intonation to mark ideas;
- intelligible pronunciation; and
- accurate reproduction of the source text.

Figure 14. Example of *Read Aloud* Task Type



Read the words aloud.

My friend and I want to room together next semester,
so we're looking for a place with two bedrooms. In the last week,
we've probably looked at more than ten different places.
Yesterday we actually found a place that's close to campus,
and it has new kitchen appliances and a new bathroom.

Note. Test takers hear audio and then read the text:

Friend (audio): Hi! How have you been? It's been a while.

Test Taker (text): Hi! I've been doing OK. How are you?

Friend (audio): I'm fine. Last time we met, you were looking for a new apartment. Did you find one?

Test Taker (text): My friend and I want to room together next semester, so we're looking for a place with two bedrooms. In the last week, we've probably looked at more than ten different places. Yesterday we actually found a place that's close to campus, and it has new kitchen appliances and a new bathroom.

Friend (audio): Great! Are you thinking about signing a lease?

Test Taker (text): We are still making up our minds. The problem is that the two bedrooms in this apartment are very different from one another. One is really big and has a beautiful view of a park with lots of trees, and the other one is much smaller and doesn't have much of a view at all.

Friend (audio): Oh wow, that doesn't seem like it'd be fair then. How would you decide who gets which one?

Test Taker (text): That's why we are still undecided about whether this apartment is the right place for us, although we do have some ideas about how to make it work. One idea that we talked about was switching rooms halfway through the year. Another idea is that we just pay different amounts of rent. The person with the big bedroom should pay more than the one with the smaller bedroom.

Friend (audio): One of those arrangements might be a good solution.

Test Taker (text): Yeah, I think I am leaning toward the arrangement of making different contributions to the rent. Trying to switch bedrooms in the middle of the year seems like a lot of trouble. And this is by far the best place we have looked at. All the others were either too far from campus or not as nice as this one.

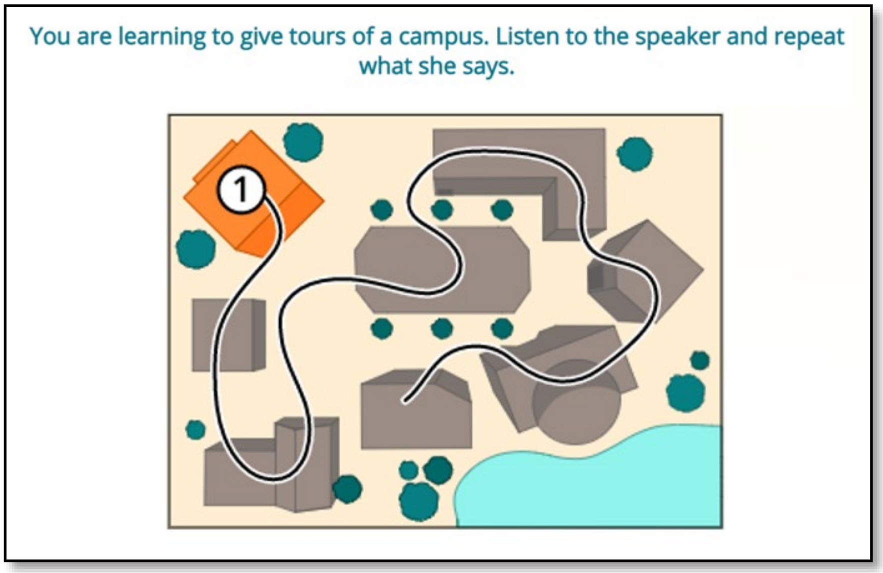
Friend (audio): Good luck getting it sorted out.

Test Taker (text): Thanks! I'm sure we will! You know, now that I've discussed it with you, I think I know what we should do. I'm going to text my friend and say that we should take the apartment and sign a lease. We can work out the details of what's fair later. After all, the semester is about to start."

The *Listen and Repeat* task measures the test taker's ability to process the sentences they hear and then to 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 15). 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 15. Example of *Listen and Repeat* Task Type



You are learning to give tours of a campus. Listen to the speaker and repeat what she says.

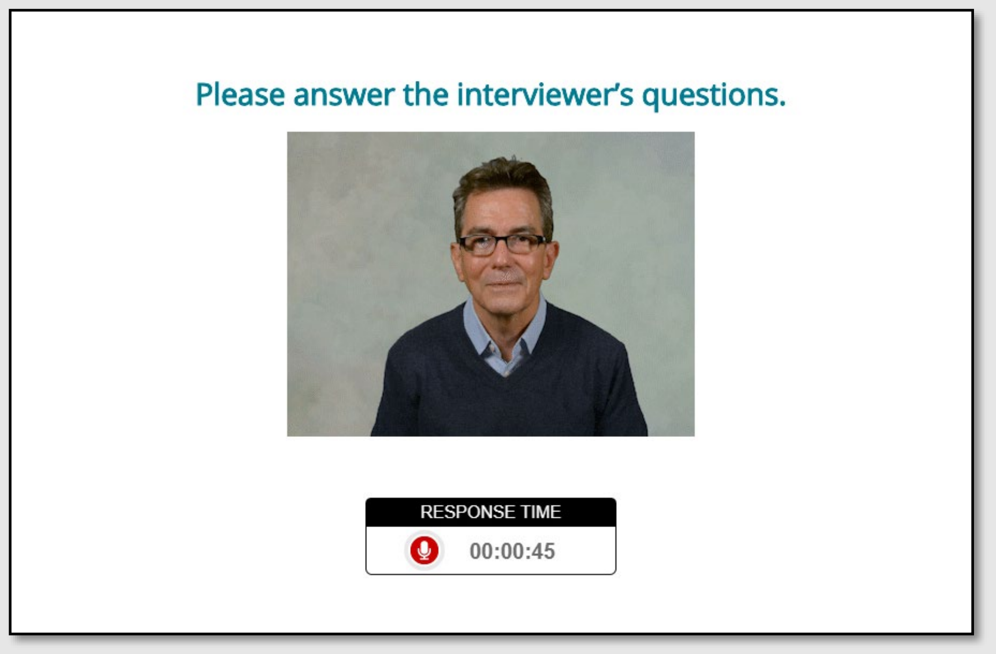
Note. Test takers hear audio and then repeat:

Welcome to our university.
Living on campus is really fun.
The café is a great place to meet friends.
This is where researchers are creating new technology.
Some students gain work experience here as lab assistants.
With a student ID card, you can open a free savings account.
Many of our courses, taught by excellent professors, can be taken online.
Paying tuition and other school fees can also be taken care of right here.

In the *Virtual Interview* task, test takers participate in a simulated conversation with a prerecorded interviewer (see Figure 16). 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 a total of five 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 *Virtual 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 16. Example of Virtual Interview Task Type



Please answer the interviewer's questions.

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

Thank you for your interest. Today, I would like to ask you some questions to see if you are a good fit for the program. First, are you currently a student?

Now, tell me about your travel experiences. When was the last time you went on a trip? Where did you go and what did you do?

And, what is one foreign country that you might like to visit? Tell me about why you would like to visit this country.

Very interesting! Here is a question about studying abroad. In your opinion, what would you recommend that students do to prepare for studying in another country?

Good points. Now, a final question. Sometimes, students studying abroad end up spending most of their time with other students from their own country. When studying in a foreign country, is it important to interact with that country's students and the local people? Why or why not?

Scoring

Calculation of Section Scores

As noted earlier, the TOEFL Essentials Listening, Reading, and Writing sections follow an MST design with two parts (stages) in each test section. All questions presented to the test taker in both parts contribute to the final score. Questions in both parts of the reading and listening sections are scored as correct or incorrect. For questions answered correctly, 1 score point is awarded. In the writing section, all questions in the first part are also scored as correct or incorrect, with 1 or 0 score points awarded, respectively. In the second part of the writing section, responses to each writing task are scored on a scale from 0 to 5 score points according to criteria listed in a scoring rubric. Speaking responses are scored distinctly from the other test sections as described later in this section of the report. In the future, responses to the speaking and writing tasks will be evaluated not only by certified raters, as is the case with the launch of the test, but also augmented with proprietary AI scoring engines. Combining human scoring with AI scoring is expected to further increase the accuracy and consistency of scores.

The total of the score points a test taker receives in each of the three sections (listening, reading, and writing), called the *raw score*, is converted to the reporting score scale through a statistical process known as *equating*. For listening and reading, equating is conducted within an item response theory (IRT) framework, whereas for writing, a hybrid IRT/equipercentile linking approach is used (Kolen & Brennan, 2004; Lord, 1980). The application of equating procedures helps 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 do not reflect the number or percentage of raw score points earned.

The TOEFL Essentials Speaking section is not adaptive but linear, which means that all test takers will encounter the same set of test task types. Tasks in the speaking section span the full range of difficulty, and raw scores are based on overall performance on all tasks. Responses to speaking tasks are scored on a multipoint scoring rubric with the score points varying from 0 to 4 or 0 to 5 depending on the task. The speaking raw score is 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). Thus, a given speaking scaled score reflects the same level of language ability regardless of when the test was taken or what specific tasks the test taker performed.

Score Reporting

Performance on each of the four test sections are reported in the form of band scores from 1 to 12. At the total test level, an overall band score is calculated as the average of the four section band scores. The overall band score ranges from 1 to 12 in increments of 0.5, rounded to the nearest whole or half band. In addition to the section and overall band scores for current test administration, the score report includes *MyBest®* score 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.

The score report also provides information about two foundational skills: vocabulary knowledge and sentence construction. These foundational skills underlie broad areas of language ability, and information about test-taker performance is reported in the form of a percentile value to help test takers understand how they performed on these skills in relation to other test takers. The percentile indicates that the test taker performed better than that percentage of all those who took the test. For example, a percentile value of 75 means that the test taker performed better than 75% of all test takers. In the future, the addition of AI scoring to human scoring for the constructed response tasks is expected to allow for the reporting of additional foundational skills (e.g., pronunciation, fluency, grammatical accuracy).

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 the Test Design Process section). Responses to prototype tasks were placed into quartiles by general proficiency of the test taker, as indicated by a C-test 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.

Mapping Test Scores to CEFR Levels

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 Essentials website. The mapping of TOEFL Essentials test scores to the CEFR levels was based on multiple sources of information. First, the TOEFL Essentials Reading and Listening sections in the field test administrations contained test questions previously included in other tests in the TOEFL family of assessments. Because the scores of these tests had already been mapped to the CEFR levels, it was then possible to also map the TOEFL Essentials Reading and Listening scores onto the CEFR levels. Reading and listening items with a difficulty that fell between two CEFR levels were also inspected by ETS Research and Development staff to determine if those items reflected key skills and abilities described in the CEFR levels. Assessment specialists also 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 TOEFL Essentials Speaking and Writing section test scores was established by combining information from three separate steps. First, scoring rubrics were designed to reflect wording from relevant CEFR level descriptors with higher ratings on speaking and writing task rating scales reflecting descriptions of language use at higher levels of the CEFR scale. Second, assessment specialists examined exemplar responses from the field test in relation to these CEFR descriptors. The purpose of this step was to confirm that the performance described in the CEFR levels was also reflected in the performance of the test takers in the field test. 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 students across the selected-response sections and the CEFR levels of the same students across the constructed response sections of the test.

Rater Training and Monitoring

Scoring quality for the TOEFL Essentials Speaking and Writing tasks is supported in a number of ways, similar to those for other tests in the TOEFL family of assessments (see ETS, 2020):

- The scoring process is centralized, and it is performed separately from the test administration to help ensure that test data are not compromised. Through centralized, separate scoring, each scoring step is closely monitored to help ensure its security, fairness, and integrity.
- ETS uses its patented Online Network for Evaluation (ONE) 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 in order 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 ONE sends speaking and writing responses to multiple independent raters for scoring. Responses from each test taker are scored by more than one rater.

Personal Video Statement

After completing the TOEFL Essentials test, all test takers record a video of themselves speaking in English to share additional insights about themselves and their interests with institutions and score users. The personal video statement is modeled on personal statements used in various academic contexts and is intended to provide score users with an impression of the goals, motivations, and communication abilities of the test taker. The personal statement also supplements examples of test-taker responses to scored writing and speaking tasks, which are available to score users. In the video, test takers respond to two questions about themselves and their opinions. The recorded video responses are not scored but are shared with the score users to whom the test taker chooses to send their scores.

In the first question, test takers talk about themselves. This question is always the same: *"What would you like to tell people about yourself?"* Test takers have considerable latitude in deciding how to present themselves. For example, they can talk about their background or their plans for the future. In the second question, test takers select a topic and give their opinions. Unlike the first question, the second question is different each time. Test takers choose one of two topics provided to them. For each question, test takers have up to 2 minutes to respond. After 1 minute, if they are done they can stop the recording. If test takers are not happy with a response, they have the option to record it one additional time, and only the second try is saved. Completing the personal video statement takes up to 5 minutes.

Test Administration and Security

The TOEFL Essentials test is delivered over the Internet to test takers at their own locations. Prior to test administration, examinees are required to download a secure browser on the computer they will use to take the test, run a system check, and fix any issues before the test date. Test content is delivered using secure transmission protocols, and test forms are assigned through centrally controlled algorithms that consider the location of the examinees and their time zone. On the test date, test security is safeguarded throughout the session by use of online human proctors and AI measures. Prior to starting the test, examinees are required to show a photo ID to their proctor and demonstrate their workspace meets several requirements. The proctor then reviews the test-taking rules and requests access to the computer screen for monitoring purposes. Examinees are also asked to use either a handheld mirror or a cell phone to show the proctor their computer screen. The proctor then instructs the examinee to launch the secure browser and provides an ID and password to access the test. Throughout the test, the proctor monitors the computer screen, observes the examinee via the computer camera, and can cancel the test for security violations. 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 examinee'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).

Scoring is also controlled centrally to further support security. For example, responses to the speaking and writing tasks are evaluated by certified raters, whose scores are recorded and constantly monitored for quality by scoring leaders through a 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.

Research and Validation

The TOEFL Essentials 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 tests in the TOEFL family of assessments (e.g., Chapelle, 2008; So et al., 2015), the research program for the TOEFL Essentials 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 Essentials test are organized into six hierarchical inferences, following those laid out in Chapelle (2008) to support the validity argument for the *TOEFL iBT*® test (Table 1). The six inferences in the TOEFL Essentials validity argument 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 Essentials validity argument reflect what Chapelle (2008) described as a "design validity argument" (p. 320). Given that the TOEFL Essentials 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 Essentials 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 Essentials 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 1, 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 1).

Ongoing Oversight

Ongoing oversight is a key feature of the TOEFL family of assessments. The TOEFL Essentials test undergoes regular internal audits every 3 years. The auditors evaluate compliance with *ETS's Standards for Quality and Fairness* (ETS, 2014), which are aligned with current measurement industry standards as reflected by the *Standards for Educational and Psychological Testing*, published jointly by the American Educational Research Association et al. (2014). Auditors report directly to the ETS Board of Trustees on any issues they may find.

The TOEFL Committee of Examiners (COE) provides guidance and oversight for research and development related to all tests in the TOEFL family of assessments, including the TOEFL Essentials test. The TOEFL COE is a panel of 11 experts from around the world, each of whom has achieved professional recognition in an academic field related to the teaching, learning, and testing of English as a second or foreign language (see https://www.ets.org/toefl/score-users/about/board/).

Table 1. Overview of Inferences in the Validity Argument for the TOEFL Essentials Test

Inferences (Chapelle, 2008)	Core claim (Chapelle, 2008)	Warrant (supporting claim)	Potential backing (supporting evidence)
Domain definition	Observations of performance on the TOEFL Essentials 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: (a) developmental sequences (e.g., acquisition of word order rules), and (b) 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.
Domain definition		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.
Domain definition		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 Essentials 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.
Evaluation		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 (Chapelle, 2008)	Core claim (Chapelle, 2008)	Warrant (supporting claim)	Potential backing (supporting evidence)
Evaluation		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.
Evaluation		Scores are free from bias or other types of unfairness.	• Procedures are developed for consistent scoring of all responses. • Scores awarded to defined subgroups of test takers do not differ.
Evaluation		Test tasks distinguish among examinees with varying degrees of proficiency.	• Discrimination of items and reliability of sections/test meet acceptable standards.
Evaluation		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.
Evaluation		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 (Chapelle, 2008)	Core claim (Chapelle, 2008)	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.
Generalization		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.
Generalization		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.
Explanation		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.
Explanation		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).

Inferences (Chapelle, 2008)	Core claim (Chapelle, 2008)	Warrant (supporting claim)	Potential backing (supporting evidence)
Extrapolation	The construct of academic language proficiency as assessed by the TOEFL Essentials 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 Essentials 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.
Utilization		The test will have a positive influence on learning and instruction.	• Score users find the section scores, the information about the foundational skills, and the availability of speaking and writing responses useful for making educational decisions.
Utilization			• Admissions and placement decisions are perceived by learners and teachers to be accurate.
Utilization			• Admissions staff indicate that the personal video statement provides useful information for decision-making; analyses of admissions decisions indicate that use of video does not contribute to bias in decisions.

References

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

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.

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-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

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. *The Journal of Writing Assessment*, *7*(1). http://www.journalofwritingassessment.org/article.php?article=74

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

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

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

ETS. (2020). *TOEFL® research insight series: Vol. 1. TOEFL iBT® test framework and test development.* https://www.ets.org/s/toefl/pdf/toefl_ibt_research_insight.pdf

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

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

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., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer-Verlag. https://doi.org/10.1007/978-1-4757-4310-4

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

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. *JLTA Journal*, *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.

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

So, Y., Wolf, M. K., Hauck, M. C., Mollaun, P., Rybinski, P., Tumposky, D., & Wang, L. (2015). *TOEFL Junior® Design Framework* (Research Report No. RR-15-13). ETS. https://doi.org/10.1002/ets2.12058

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

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