ets research institute

JUNE
2026
RM–26-04

RESEARCH MEMORANDUM

Mapping TOEIC® Link™ Scores to the Common European Framework of Reference for Languages

AUTHORS

Kathryn Hille, Jonathan Schmidgall, Saerhim Oh, and Renka Ohta

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 Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

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

Mapping TOEIC® Link™ Scores to the Common European Framework of Reference for Languages

Kathryn Hille, Jonathan Schmidgall, Saerhim Oh, & Renka Ohta

ETS Research Institute, ETS, Princeton, New Jersey, United States

June 2026

Corresponding author: Kathryn Hille, E-mail: khille@ets.org

Suggested citation: Hille, K., Schmidgall, J., Oh, S., & Ohta, R. (2026). *Mapping TOEIC® Link™ scores to the Common European Framework of Reference for Languages* (Research Memorandum No. RM-26-04). ETS. https://doi.org/10.64634/4amdns26

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

To obtain a copy of an ETS research report, please visit
https://www.ets.org/contact/additional/research.html

Action Editor: Larry Davis

Reviewers: Spiros Papageorgiou and Richard Tannenbaum

Table of Contents

Abstract

This study maps TOEIC® Link™ assessment scores to levels of the Common European Framework of Reference for Languages based on the judgments of a panel of experts. The Bookmark method was used to map the TOEIC Link assessments containing multiple-choice items (listening and reading), whereas the Expected Task Score method was used to map the assessments containing constructed-response items (speaking and writing). The panelists reported high levels of agreement, comfort with the panelist-recommended cut scores, and satisfaction with the standard setting process. A few small poststudy adjustments were made in the interest of congruence among ETS English proficiency assessments with common items or overlapping task types.

Keywords: standard setting, score mapping, TOEIC®, TOEIC® Link™, Common European Framework of Reference for Languages, CEFR, score interpretation

Introduction

This standard setting study maps test-taker performance on each of the four newly developed TOEIC® Link™ assessments to the Common European Framework of Reference for Languages (CEFR) proficiency levels A1, A2, B1, B2, and C1. The purpose is to provide information that will aid in score interpretation and use, which is especially important given that these are new assessments using a new score scale, with which score users are expected to be initially unfamiliar. TOEIC Link score reports include test-taker CEFR levels, both for each assessment taken and overall, based on the results of this study.

Standard setting or score-mapping studies seek to establish a correspondence between scores on a given assessment and classification into a given performance category (Cizek, 2012). The standard setting process first requires the articulation of the performance standards to be used for classification, including the knowledge, skills, and abilities expected from members of each performance category. Specifically, it is crucial to describe the characteristics of performance at the lower boundary of each category (i.e., the absolute minimum requirements for classification into that category). It is the operationalization of these minimum requirements via panelist judgments that yields the cut scores on the assessment for each performance category.

For this study, the targeted performance categories belong to a pre-existing framework of language proficiency standards: the CEFR levels A1, A2, B1, B2, and C1. The CEFR, published in 2001 (Council of Europe, 2001), contains performance descriptors across a range of language competences and sub-competences classified into six main levels of proficiency for language learners, with additional descriptors introduced in a subsequent companion volume (Council of Europe, 2020). The CEFR is intended to be broad enough to apply to the learning of any language across a range of pedagogical approaches and assessment methods. It provides a robust array of descriptive scales that target various combinations of language skills (including listening, speaking, reading, writing, grammar, and vocabulary) within communicative contexts (including in-person and electronic modalities, across various social settings) and using communicative strategies (including clarification, mediation, turn-taking, collaboration, and simplification). Not all of the descriptive scales will be relevant to every communicative situation, neither in real-life interactions nor in assessment contexts. Therefore, one prerequisite of mapping test scores to CEFR levels is the identification of the CEFR descriptive scales that are the most relevant to the knowledge, skills, and abilities that the assessment is designed to assess.

This study begins with a description of the new TOEIC Link assessment, including its design, construct definition, task types, and scoring model. Then, a construct congruence analysis is presented that identifies the CEFR scales most relevant to each of the TOEIC Link assessments (listening, reading, speaking, and writing). The procedures of the standard setting study are then described, beginning with the selection of test forms and items, and the selection and preparation of panelists prior to the panel meeting. The procedures during the panel meeting are described next, including the creation of descriptors for the just qualified candidate (JQC) at each relevant CEFR level and the procedures for panelist judgments. The standard setting panel results are presented, and a few small poststudy adjustments are recommended in the interest of congruence among the CEFR mappings of assessments with overlapping task types. Finally, the recommended cut scores are supported with procedural, internal, and external validity evidence.

Development of TOEIC Link Assessments

The TOEIC Link assessments were developed to meet market needs for efficient assessments of English language proficiency in the everyday and workplace contexts. The design requirements included the following:

- assessment of four skills: listening, reading, speaking, and writing
- modular assessment so that the assessment for each of the four skills can be administered alone or together in any combination
- assessment across a range of proficiency levels (i.e., CEFR A1–C1)
- a fully digital assessment experience
- machine scoring, with scores reported within 48 hours
- a maximum of 90 minutes of combined testing time across the four assessments
- score use recommended for medium- to high-stakes purposes
- one reporting scale for all four skills

Building on these design requirements, blueprints were created for each assessment, using task types known from other ETS tests where possible, in order to most successfully anticipate the functionality and performance of the items. For the TOEIC Link Listening and Reading assessments, task types were drawn from the listening and reading sections of the TOEIC® test. For the TOEIC Link Speaking test, task types were drawn from the speaking sections of the TOEIC test and of the TOEFL® Essentials™. For the TOEIC Link Writing test, task types were drawn from the writing sections of the TOEIC Bridge® and the TOEFL Essentials.

TOEIC Link Construct Definition

The purpose of the TOEIC Link assessments is to evaluate the English language proficiency of individuals whose first language is not English in the everyday and workplace contexts. Each assessment is administered by computer, alone or with other TOEIC Link assessments (listening, reading, speaking, and/or writing). Test scores are intended to be used for the purposes of hiring (e.g., for positions within an organization where workplace/everyday

English is a valued job skill), selection (e.g., for vocational or training institutes), placement (e.g., into training or courses of study), promotion, or measurement of individuals' progress in workplace or everyday English proficiency levels over time. Test takers are intended to be young adults (secondary school students) and adults for whom English is a foreign language; collectively, their nationalities and native languages vary. Test takers' educational backgrounds and purposes for learning English (e.g., academic purposes, occupational purposes) may also vary.

Listening Construct Definition

The TOEIC Link Listening assessment measures the ability of beginning to advanced English language learners to understand spoken English texts in everyday and workplace contexts. Test tasks are designed to elicit evidence of test takers' abilities with respect to the claims that the test taker can understand the following:

- gist in short spoken text
- gist in extended spoken text
- details in short spoken text
- details in extended spoken text
- purpose or implied meaning/pragmatic understanding

These abilities are assessed in task types that incorporate a variety of everyday and workplace settings, including corporate development, dining out, entertainment, finance and budgeting, general business, health, housing/corporate property, manufacturing, offices, personnel, purchasing, technical areas, and travel (see ETS, 2022, pp. 3–4). These settings provide context for test questions, but test takers are not required to have business experience or to understand specialized vocabulary (see Schmidgall et al., 2024).

The design of the TOEIC Link Listening assessment is based on the TOEIC Listening test section and consequently intends to measure the same listening skills using the same test tasks. Therefore, the construct definition and its operationalization in item and test specifications are comparable and draw upon the same conceptual framework.

Reading Construct Definition

The TOEIC Link Reading assessment measures the ability of beginning to advanced English language learners to understand written English texts in everyday and workplace contexts. Test tasks are designed to elicit evidence of test takers' abilities with respect to the following claims:

- making inferences based on information that is explicitly stated in texts (within text inferences and across text gist)
- understanding specific (factual) information in tables and passages, including negative facts (i.e., factual information expressed through clause-level negation)
- connecting information across multiple sentences in single texts and across two texts
- understanding vocabulary (i.e., in context)
- understanding grammar

These abilities are assessed in task types that incorporate a variety of everyday and workplace settings, including corporate development, dining out, entertainment, finance and budgeting, general business, health, housing/corporate property, manufacturing, offices, personnel, purchasing, technical areas, and travel (see ETS, 2022, pp. 3–4). These settings provide context for test questions, but test takers are not required to have business experience or to understand specialized vocabulary (see Schmidgall et al., 2024).

The design of the TOEIC Link Reading assessment is based on the TOEIC Reading test section and consequently intends to measure the same reading skills using the same test tasks. Therefore, the construct definition and its operationalization in item and test specifications are comparable and draw upon the same conceptual framework.

Speaking Construct Definition

The TOEIC Link Speaking assessment measures the general speaking proficiency of beginning to advanced English language learners in everyday and workplace contexts. Measurement of general speaking proficiency is achieved using test tasks that elicit evidence of

(a) foundational skills that have been shown to be predictive of overall speaking proficiency and (b) communicative skills that are typical of those needed in everyday and workplace contexts.

The TOEIC Link Speaking assessment was developed utilizing a hybrid approach to test design that balances the measurement of foundational and communicative skills (see Davis, Norris, et al., 2023; Papageorgiou et al., 2021). In this approach, some test tasks are designed to measure foundational language skills (e.g., fluency, pronunciation, accurate reproduction of language) that have shown to be strong predictors of speaking proficiency. These test tasks are complemented by tasks designed to provide a more direct measurement of communicative skills in an authentic or real-world setting. This hybrid approach to assessment design is intended to achieve efficiencies through targeted assessment of foundational skills while retaining measurement of communicative skills.

The TOEIC Link Speaking test tasks that measure foundational language skills are based on task types that were previously developed for TOEFL Essentials, with contextual elements modified for use in the TOEIC Link. The TOEIC Link Speaking test task that measures communicative language skills is based on a task type from the TOEIC Speaking test.

Writing Construct Definition

The TOEIC Link Writing assessment measures the general writing proficiency of beginning to advanced English language learners in everyday and workplace contexts. Measurement of general writing proficiency is achieved using test tasks that elicit evidence of (a) foundational skills that have been shown to be predictive of overall writing proficiency and (b) communicative skills that are typical of those needed in everyday and workplace contexts.

The TOEIC Link Writing assessment was developed using the same hybrid approach to assessment design described above for the TOEIC Link Speaking assessment (see also Davis, Norris, et al., 2023; Papageorgiou et al., 2021). The TOEIC Link Writing test task that measures foundational language skills is based on a task type utilizing key scoring that was previously developed for TOEIC Bridge. The TOEIC Link Writing test tasks that measure communicative language skills are based on task types utilizing automated scoring that were previously developed for TOEFL Essentials, with contextual elements modified for use in the TOEIC Link.

TOEIC Link Task Types

The TOEIC Link task types vary by assessment, with the listening and reading assessments containing multiple-choice items, whereas the speaking and writing assessments contain constructed-response items.

Listening Task Types

The TOEIC Link Listening assessment follows a multistage test (MST) design, whereby test content is administered in two stages: Stage 1 (Router Module) and Stage 2 (Easy Module or Hard Module). Test takers are assigned to a Stage 2 module based on their performance on Stage 1. Scoring is based on test-taker performance on both the Stage 1 Router Module, which is of medium difficulty, and the Stage 2 Easy Module or Hard Module as assigned.

Stage 1 (Router Module) contains four task types, Stage 2 (Easy Module) contains two task types, and Stage 2 (Hard Module) contains two task types. Each task type contains items designed to provide evidence with respect to the multiple claims described earlier. Conversations and Talks items are grouped into sets of three items each, where the set of items addresses a single conversation or talk. Table 1 presents the task types and their arrangement for listening.

Table 1. Listening Task Types and Arrangement

Stage	Module	Task type	Number of items	Claims addressed
1	Router	Photos	2	1 and 3
		Question Response	3	1, 3, and 5
		Conversations	6 (in 2 sets)	2, 4, and 5
		Talks	6 (in 2 sets)	2, 4, and 5
2	Easy	Question Response	7	1, 3, and 5
		Conversations	6 (in 2 sets)	2, 4, and 5
2	Hard	Question Response	7	1, 3, and 5
		Talks	6 (in 2 sets)	2, 4, and 5

Photos. Test takers hear four statements about a picture that is printed in their test book. Test takers must select the one statement that best describes the picture by marking the appropriate option (A–D). These questions are designed to provide evidence of test-taker ability with respect to Claim 1, understanding of the gist or central idea of the listening input, and Claim 3, matching this understanding to the obvious content of an image.

Question Response. Test takers hear a question or statement followed by three potential responses spoken in English. The question and response involve two speakers. Neither the question/statement nor responses are printed in the test book. Test takers must select the best response to the question/statement by marking the appropriate option on their answer sheet (A–C). These questions are designed to provide evidence of test-taker ability with respect to Claim 1, understanding gist, purpose, and basic context in short spoken texts; Claim 3, understanding details in short spoken texts; and Claim 5, understanding implied meaning in short spoken texts (pragmatics).

Conversations. Test takers hear a conversation between two or three speakers. The length of each conversation varies between 70 and 115 words. Some sets also contain a graphic (visual stimulus). After listening to each conversation, test takers answer three questions about what speakers say. Each question is spoken aloud, and questions and response options are printed in the test book. Test takers select the best response on their answer sheet (A–D). These questions evaluate test-taker ability with respect to Claim 2, inferring gist (main idea, context); Claim 4, understanding details; and Claim 5, understanding a speaker's purpose or implied meaning in a phrase or sentence (pragmatics).

Talks. Test takers listen to talks given by a single speaker. The length of talks varies between 70 and 115 words. Some talks also contain a graphic (visual stimulus). After hearing each talk, test takers answer three questions about what the speaker says. Each question is spoken aloud and printed in the test book, and response options are printed in the test book. Test takers select the best response option on their answer sheet (A–D). Similar to Conversations task type, these questions evaluate test-taker ability with respect to Claim 2, inferring gist (main idea, context); Claim 4, understanding details; and Claim 5, understanding a speaker's purpose or implied meaning in a phrase or sentence (pragmatics).

Reading Task Types

The reading assessment follows an MST design, similar to that of the listening assessment. Stage 1 (Router Module) contains four task types, Stage 2 (Easy Module) contains three task types, and Stage 2 (Hard Module) contains three task types. Reading Comprehension Passages items are grouped into sets of three to five items each, where the set of items

assesses understanding of one reading passage (Single Passage) or group of two or three passages (Multiple Passages). Table 2 presents the task types and their arrangement for reading.

Table 2. Reading Task Types and Arrangement

Stage	Module	Task type	Number of items
1	Router	Incomplete Sentences	4
		Text Completion	4
		Single Passage (based on one text)	4 (in one set)
		Multiple Passages (based on two texts)	5 (in one set)
2	Easy	Incomplete Sentences	5
		Single Passage (based on one text)	3 (in one set)
		Multiple Passages (based on two texts)	5 (in one set)
2	Hard	Incomplete Sentences	5
		Single Passage (based on one text)	3 (in one set)
		Multiple Passages (based on three texts)	5 (in one set)

Incomplete Sentences. Test takers are presented with a sentence that has a missing word or phrase. Test takers must then select the word or phrase, from among four options, that best completes the sentence. These questions specifically assess either grammar or vocabulary at the sentence level.

Text Completion. Test takers read short texts in a variety of formats. Each short text is missing four elements such as words, phrases, or key sentences. Test takers must correctly identify each missing element by selecting the appropriate word, phrase, or sentence from among four options. These questions assess grammar, vocabulary, and the ability to connect information within a short text.

Reading Comprehension. For this task type, which includes both single and multiple passages, test takers must read everyday texts (e.g., notices, letters, forms, advertisements) and answer three to five questions about each text or set of texts. The questions accompanying each text or set of texts may require the test taker to identify the main idea, identify stated details, infer implied meanings such as the context or the writer's purpose, or connect information within or across texts.

Speaking Task Types

The speaking assessment follows a linear design, whereby the same test content is administered to all test takers for a given form. Table 3 presents the task types and their arrangement for speaking. The speaking assessment consists of two foundational task types, the Read Aloud task and the Listen and Repeat task, as well as a communicative task type, Express an Opinion.

Table 3. Speaking Task Types and Arrangement

Task type	Number of items
Read Aloud	3
Listen and Repeat	7
Express an Opinion	1

Read Aloud. Test takers read aloud one part of a dialogue that takes place in a daily life or workplace context. Test takers are first provided with information on who the interlocutor is (e.g., friend), and then they hear the interlocutor make a statement or ask a question. Test takers then respond by reading aloud their part of the conversation, which is written on the screen. Each test taker's turn in the conversation consists of one to three sentences. The topic of the conversation includes everyday and workplace contexts. Test takers are given 40 seconds to complete each item.

This task measures a test taker's ability to accurately process written text into speech in a way that is intelligible to proficient speakers of English. This task aims to measure foundational speaking skills such as fluency, pronunciation, pacing, and prosodic features such as stress and intonation.

Listen and Repeat. Test takers repeat a series of sentences within a scenario in an everyday or workplace context. Test takers are first presented with the scenario which provides a communicative purpose for listening and repeating the sentences. Each sentence test takers listen to is associated with a visual/graphic that corresponds to the content being heard. After each sentence, there is a pause, and test takers repeat exactly what was said. Sentences get progressively longer and more complex as test takers progress through the scenario.

This task measures a test taker's ability to process linguistic information and reproduce it accurately and intelligibly—foundational skills that are essential elements of spoken language production.

Express an Opinion. Test takers are presented with a topic and are asked to give their opinion about an everyday or workplace topic. Test takers are given 45 seconds to prepare and 60 seconds to respond.

This communicative task measures a test taker's ability to create connected, sustained, and intelligible discourse appropriate to the typical workplace or casual social situation. The ability to state and support an opinion, as measured by this task, also includes foundational language skills related to delivery, intelligibility, effective and accurate use of grammar and vocabulary, and topic development and coherence.

Writing Task Types

The writing assessment follows a linear design, whereby—as defined previously—the same test content is administered to all test takers for a given form. Table 4 presents the task types and their arrangement for writing. The writing assessment consists of a foundational task type, Build a Sentence, as well as three communicative task types, Write a Brief Email, Write an Extended Email, and Write for an Online Discussion.

Table 4. Writing Task Types and Arrangement

Task type	Number of items
Build a Sentence	3
Write a Brief Email	1
Write an Extended Email	1
Write for an Online Discussion	1

Build a Sentence. Test takers read a sentence or question with words or phrases in the wrong order, and they drag and drop words or phrases to form a grammatical sentence or question. The maximum length of each sentence or question is 15 words, and the topics include everyday personal, public, and familiar workplace topics. Test takers have 1 minute to complete each item.

This task measures a test taker's ability to produce grammatically correct sentences—an important foundational skill for writing proficiency.

Write a Brief Email. Test takers are presented with a scenario in either a workplace or a social setting along with a table with information regarding context. They are tasked to write an email and are provided with two specific tasks to accomplish in the email. The context of the scenario includes settings such as dining, health, purchasing, transportation, and work. The table presented to the test takers also includes content such as information about a product or service. The two tasks which test takers need to accomplish in the email include those such as providing information, making a recommendation, asking for help, and responding to a request. Test takers are given 7 minutes to respond and are instructed to write as much as they can without a word count recommendation.

This communicative task measures a test taker's ability to write a brief message that achieves a designated communication goal with pragmatic appropriateness (e.g., share information, extend an invitation, make a recommendation) and provides some elaboration. The ability to write a brief text for a communication goal involves foundational skills that include grammatical accuracy and range as well as vocabulary accuracy and range.

Write an Extended Email. Test takers are presented with a scenario in either a workplace or a social setting. They are tasked to write an email and are provided three specific tasks to accomplish in the email. The context of the scenario includes settings such as dining, health, purchasing, transportation, and work. The three tasks test takers need to accomplish in the email include providing information, making a recommendation, describing a problem, expressing complaints, and refusing plans. The relationship between the sender (test taker) and the recipient of the email includes those such as friend–friend, student–professor, employee–customer, and customer–business. Test takers are given 7 minutes to respond and are instructed to write as much as they can without a word count recommendation.

This communicative task measures a test taker's ability to write a message that achieves a designated communication goal with pragmatic appropriateness (e.g., describe a problem, make plans, request or suggest a solution, make a complaint) and provides some elaboration. The ability to write a text for a communication goal also involves foundational skills that include grammatical accuracy and range as well as vocabulary accuracy and range.

Write for an Online Discussion. Test takers are presented with an online discussion forum that includes a moderator's post (the initial post in the discussion that introduces the topic of the discussion) and two posts, each from a different participant, which provide ideas and information to stimulate test takers' responses. Test takers are tasked to read the posts by the moderator and two participants and write a post responding to the moderator's question. They are told to express and support their opinion and make a contribution to the discussion. The topics of the forum are related to everyday life and the workplace and are accessible to a general audience. Test takers are given 10 minutes to complete the task and are recommended to write at least 100 words.

This task measures a test taker's ability to produce a written text that clearly elaborates an argument for a position while responding to arguments and/or using information provided in short texts. This includes foundational skills such as grammatical accuracy and range, vocabulary accuracy and range, and communicative aspects including elaboration and content relevance.

TOEIC Link Scoring Model

The TOEIC Link comprises four assessments—listening, reading, speaking, and writing—which can be taken in any combination during a given testing session; in other words, any one, two, three, or all four assessments may be selected and taken by the test taker. A separate score is reported for each assessment taken, and an overall score is reported that is the average of the assessment scores for that testing session. The assessment scores and the overall score are each reported on a scale of 0 to 25 with whole numbers only. CEFR levels are also reported for each assessment and overall for the testing session, based on the results of this study.

The method of calculating scores varies by assessment. For the TOEIC Link Listening and Reading assessments, a scoring model based on item response theory (IRT) generates the assessment scores for each form as well as the raw-to-scale-score conversions. Equating methodology ensures fairness and comparability among forms. For the TOEIC Link Speaking and Writing assessments, test-taker performances are scored using ETS's automated scoring engines for constructed-response items (except for the Build a Sentence task in the TOEIC Link Writing assessment, which is scored using a fixed answer key). The total score points from all

questions in the speaking or writing assessment are transformed into scale scores. For speaking and writing, form comparability and scoring fairness are ensured through the item development process and the automatic scoring engines.

Construct Congruence

An important step in mapping language assessment scores to CEFR levels is the specification stage or evaluation of construct congruence (Council of Europe, 2001). In their guide to mapping language assessment scores to CEFR levels, the Council of Europe (2001; see also Council of Europe, 2009) described a specification stage which involves defining how test content and tasks are expected to align with language ability as described in the CEFR. Part of this process involves identifying the CEFR descriptor scales (and descriptors within relevant scales) that align with the knowledge and skills targeted by test tasks. This step of establishing construct congruence is crucial for score-mapping studies, as it establishes the expected overlap between the construct(s) measured by the test and the construct(s) specified in the language proficiency descriptors (Tannenbaum & Cho, 2014).

In the case of the CEFR, dozens of descriptor scales have been developed to elaborate many aspects of language knowledge, skills, and abilities; construct congruence (or specification) also serves to make the limits of the intended mapping between assessment scores and CEFR levels clearer. For example, assessment scores from an assessment of grammatical knowledge may be mapped to CEFR levels, but in doing so may only draw upon two or three descriptor scales. Conversely, assessment scores from a four-skills assessment of communicative competence may be mapped to CEFR levels while drawing upon dozens of descriptor scales. In both examples, language assessment scores may facilitate inferences about a test taker's CEFR level, but an assessment assessing a narrower construct (e.g., grammatical knowledge) facilitates a much narrower inference about a test taker's CEFR level (i.e., for grammar) when compared to the inferences facilitated by the assessment assessing a broader construct (e.g., communicative competence).

Additionally, the outcome of the construct congruence step supports the standard setting meetings, as the relevant CEFR descriptors identified through this process are provided to the panelists so that they can become familiar with the CEFR levels during training and can

refer to the descriptors when creating JQC level descriptors. Panelists use the relevant CEFR descriptors to complete premeeting activities that help to deepen their understanding of, and facility of working with, the CEFR levels, with a focus on the aspects that will be most relevant to their judgments regarding the assessment. Thus, this step provides support for another important aspect of the CEFR mapping process: familiarization (Council of Europe, 2001).

For the TOEIC Link Listening and Reading assessments, a previous standard setting study (Tannenbaum & Wylie, 2008) included a construct congruence procedure in which the CEFR descriptor scales most relevant to the TOEIC Listening and Reading test were identified. These CEFR descriptor scales were deemed relevant to the TOEIC Link Listening and Reading assessments, given the construct overlap between the TOEIC Link and TOEIC Listening and Reading tests.

For the TOEIC Link Speaking and Writing assessments, two researchers familiar with the construct definitions and item specifications first reviewed the CEFR descriptor scales and identified the scales that were potentially relevant to the constructs assessed by the assessments. Then, researchers independently reviewed the descriptors from these scales associated with CEFR levels A1–C1 to evaluate whether each descriptor was related to the constructs directly assessed by the tests, coding each descriptor Yes, No, or Maybe. The descriptors coded as Maybe, and those that were not in agreement between the two researchers, were discussed and resolved.

Tables 5, 6, 7, and 8 include the CEFR descriptor scales from the previous standard setting study for listening and reading as well as the scales included in the review for speaking and writing. Those that were determined to be most relevant to TOEIC Link are denoted by a check mark (e.g., the CEFR descriptor scales related to speaking included in the review related to oral production, oral interaction, online interaction, linguistic competence, sociolinguistic competence, and pragmatic competence). However, scales related to oral interaction, online interaction, and sociolinguistic competence were not determined to be relevant (and thus, are not denoted by a check mark). See also Davis, Garcia Gomez, et al. (2023) for additional construct congruence evidence for the speaking and writing assessment tasks drawn from TOEFL Essentials.

Table 5. Congruence Between Common European Framework of Reference for Languages (CEFR) Descriptor Scale and Descriptors for TOEIC Link: Reception Activities

CEFR descriptor scale	Listening	Reading	Speaking	Writing
Oral comprehension				
Overall oral comprehension				
Understanding conversation between other speakers	✓			
Listening to announcements and instructions				
Listening to audio media and recordings				
Reading comprehension				
Overall reading comprehension				
Reading correspondence				
Reading for orientation		✓		
Reading for information and arguments				
Reading instructions				
Reading as a leisure activity				

Note. A check mark denotes CEFR descriptor scales determined to be most relevant to the TOEIC Link assessments.

Table 6. Congruence Between Common European Framework of Reference for Languages (CEFR) Descriptor Scale and Descriptors for TOEIC Link: Production Activities

CEFR descriptor scale	Listening	Reading	Speaking	Writing
Oral production				
Overall oral production				
Sustained monologue: describing experience				
Sustained monologue: giving information			✓	
Sustained monologue: putting a case				
Public announcements				
Addressing audience				
Written production				
Overall written production				✓
Creative writing				
Reports and essays				

Note. A check mark denotes CEFR descriptor scales determined to be most relevant to the TOEIC Link assessments.

Table 7. Congruence Between Common European Framework of Reference for Languages (CEFR) Descriptor Scale and Descriptors for TOEIC Link: Interaction Activities

CEFR descriptor scale	Listening	Reading	Speaking	Writing
Oral interaction				
Overall oral interaction				
Understanding an interlocutor				
Conversation				
Informal discussion				
Formal discussion				
Goal-oriented co-operation				
Obtaining goals and services				
Information exchange				
Interviewing and being interviewed				
Using telecommunication				
Written interaction				
Overall written interaction				✓
Correspondence				
Notes, messages, and forms				
Online interaction				
Online conversation and discussion				
Goal-oriented online transactions and collaboration				

Note. A check mark denotes CEFR descriptor scales determined to be most relevant to the TOEIC Link assessments.

Table 8. Congruence Between Common European Framework of Reference for Languages (CEFR) Descriptor Scale and Descriptors for TOEIC Link: Communicative Language Competencies

CEFR descriptor scale	Listening	Reading	Speaking	Writing
Linguistic competence				
General linguistic range				
Vocabulary range				
Grammatical accuracy				
Vocabulary control		✓	✓	✓
Overall phonological control				
Sound articulation				
Prosodic features				
Orthographic control				
Sociolinguistic competence				✓
Sociolinguistic appropriateness				
Pragmatic competence				
Flexibility				
Turn-taking				
Thematic development			✓	✓
Coherence and cohesion				
Propositional precision				
Fluency				

Note. A check mark denotes CEFR descriptor scales determined to be most relevant to the TOEIC Link assessments.

The descriptor scales identified as relevant to the TOEIC Link assessments in Tables 5, 6, 7, and 8 reflect the initial construct comparability claim between TOEIC Link assessment scores and CEFR descriptor scales. These descriptor scales were consequently included in preparation material provided to the panelists (see the Preparation of Panelists section). As a result, the familiarization process for panelists emphasized the CEFR descriptor scales most relevant to TOEIC Link assessment score interpretations, and the judgment-based procedure used in standard setting emphasized the most construct-relevant subset of CEFR descriptors.

Standard Setting Study

This study conducted the score mapping using the Bookmark method (Cizek & Bunch, 2007; Reckase, 2023) for the TOEIC Link Listening and Reading assessments and the Expected Task Score method (Plake & Cizek, 2012) for the TOEIC Link Speaking and Writing assessments.

The Bookmark method was considered to be well suited for the listening and reading assessments, due to the fact that both it and the scoring for these two assessments are based on IRT parameters. This method involves assembling an ordered item booklet (OIB) of test items in ascending order of difficulty, as measured by either the IRT *b*-parameter or the theta value at a selected response probability (RP; most commonly, RP = 0.67). The panelists then identify, or bookmark, the first item they judge that test takers at the very bottom of a proficiency level will no longer have a certain chance (again, most commonly 0.67) of answering correctly.

The Bookmark method has several advantages. Its basis in IRT means that cut scores can be calculated precisely from panelist judgments, with no need to factor in additional calculations to address possible test-taker guessing behavior on multiple-choice items. Another advantage is that Bookmark panelists make judgments regarding items that are already objectively in order of difficulty, as calculated using IRT, based on actual test-taker performance data. This eliminates the possibility of panelists mis-ordering the difficulty of items, which may occur in standard setting methods where panelists make item-level judgments without item-level difficulty data. However, we also encouraged panelists to check items above and below their intended bookmark to make sure that the cut score was the best place overall, in their judgment. Finally, the Bookmark method is efficient because panelists only need to make one

judgment per cut score, versus the many item-level judgments per cut score that many other methods require. This efficiency not only saves time, but also lowers panelist fatigue and allows panelists to focus their attention on making just one very careful judgment per cut score. So, for assessments with an IRT-based scoring system and available IRT parameters, the Bookmark method is advantageous, and it was selected for use for mapping the TOEIC Link Listening and Reading assessments.

For the speaking and writing assessments, the Expected Task Score method was used. These assessments do not use IRT for scoring, and operational test-taker responses were not available for them. So, it was necessary to have panelists make judgments based on task type information, sample item content, and especially scoring rubrics for each task type. The Expected Task Score method is an extension of the Angoff method (Plake & Cizek, 2012) that works with these available materials by having panelists judge what scores test takers at the very bottom of each proficiency level would be likely to receive on each item type.

The Expected Task Score method has the advantage of being generally clear and intuitive for panelists, as it easily connects with their typical professional experiences of using a rubric to score test-taker responses, and then uses those responses and scores to make interpretations about the language proficiency of learners. Unlike with the Bookmark method, there are no item-level difficulty values to incorporate into their judgments; panelists are working exclusively with actual test materials. However, a disadvantage of the Expected Task Score method is that panelists do not have the opportunity to review any actual test-taker responses. Instead, the panelists need to make inferences about how test takers at a certain proficiency level would be likely to respond to certain task types, and then make further inferences about how those test-taker responses would likely be scored. For this study, this disadvantage was unavoidable because no actual test-taker responses for speaking and writing were available. We therefore used this method so as to utilize all of the materials that we did have available; and during the poststudy adjustment process, additional consideration was given to operational data from adjacent tests with task types that overlap with those in the TOEIC Link Speaking and Writing assessments.

Prior to Panel Meeting

Selection of Assessment Forms and Items

Because the Bookmark method was used to map scores for the TOEIC Link Listening and Reading assessments, it was necessary to select listening and reading assessment forms to be used for the OIB for each assessment. Three assembled TOEIC Link forms were available for this purpose. After consideration of the statistical values across these forms, Forms 2 and 3 were selected for both listening and reading to serve as sources of items for the OIBs.

A key goal in assembling the OIBs was to select items that, once reordered according to theta value at RP = 0.67, would allow for these theta values to be as evenly spaced as possible. This would maximize the precision of panelist-selected cut scores by eliminating the possibility of panelists setting a bookmark that would place the cut score in between two theta values that were some distance from each other. Meanwhile, it was also desirable to select item sets in their entirety, and to maintain original item sequence numbers from the test form. With these considerations in mind, listening and reading OIBs were assembled from items in Forms 2 and 3. Exactly one item was selected for each sequence number in the test form, and each item set was selected in its entirety from one form or the other. Meanwhile, the evenness of spacing of the theta values between items (when ordered by theta) was prioritized during this OIB item selection and assembly process.

For the TOEIC Link Speaking and Writing assessments, the considerations in form selection were different. The primary role of the forms would be simply to provide sample items for panelists to use as reference as they made their judgments within the Expected Task Score approach. However, it was desirable to also select sample items that were as similar as possible to corresponding items in the other English language proficiency tests consulted in the development of the TOEIC Link. This selection would allow operational data from those corresponding items to be considered during the poststudy adjustments phase; the TOEIC, TOEIC Bridge, and TOEFL Essentials tests have all been mapped to the CEFR, which could serve as additional reference points. Because no operational data were available for the TOEIC Link Speaking and Writing assessments, the data from corresponding items in the other tests were expected to be an important consideration alongside the panelists' judgments. Therefore, Form

1 was selected to provide the speaking and writing sample items for the standard setting process because it contained items that were all identical or very similar to their corresponding items in the other tests.

Selection of Panelists

The expert panelists were recruited based on the following selection criteria:

(a) familiarity with needs and characteristics of young adult and adult English language learners, who study English for everyday and workplace purposes, typically through English language teaching

(b) experience developing learning or assessment materials for English language learners

(c) familiarity with CEFR levels, ranging from beginner to advanced proficiency levels

(d) high proficiency in English

Due to the global nature of expert recruitment for this study, ETS Princeton staff worked closely with the ETS Europe, Middle East, and Africa (EMEA) office in Paris and an organization in the ETS Preferred Provider Network, Review Quality, in Mexico, both of which assisted with identifying potential candidates in their local contexts. The following procedure was undertaken at the end of July 2024. First, a recruitment email was sent directly by ETS Princeton staff to relevant individuals with whom we had already established working relationships. The individuals were those whom we knew through our professional and personal connections, and who met the selection criteria. Second, the recruitment email was distributed widely through the ETS EMEA office and Review Quality to reach out to potential candidates in France, Italy, Türkiye, and Mexico. ETS EMEA contacted those who had participated in a previous ETS standard setting study, as well as those who might meet the selection criteria. Review Quality, which is a distributor of TOEIC tests, also identified candidates based on their professional connections.

Interested candidates were asked to respond to a background survey by clicking on a link included in the recruitment email. The online background survey included a series of questions regarding their demographic information (e.g., affiliation, country of residence, job

title) and their experience teaching and developing learning/assessment materials for English language learners, as well as their familiarity with CEFR levels. They were also asked to upload a curriculum vitae (CV) or résumé if available. Lastly, they were asked to indicate their availability for three days of standard setting meetings in August and September 2024.

By early August 2024, we received survey responses from 36 candidates around the world. Using an Excel spreadsheet, the candidate information was tabulated to allow for comparison among candidates. A simple rubric was used to sort the candidates based on the aforementioned criteria, allowing us to compare and identify candidates with more experience and readiness for standard setting activities. After rank-ordering candidates, we considered demographic balance (i.e., gender, country of origin, institution type, job title). This process was critical to ensure the recruitment of diverse panelists and enable broader perspectives and professional judgments to be reflected in the cut score judgments.

A total of 14 candidates—eight females and six males—were selected, were offered a panelist position, and agreed to participate in the standard setting study; a few of them were replaced well in advance of the meeting because they withdrew from the study for scheduling conflicts and personal reasons. Panelists were located in five different countries: France (*n* = 2), Italy (2), Türkiye (4), Mexico (3), and the USA (3). Of the 14 panelists, eight were currently affiliated with a university, three were employed in an English language school, and two were teaching at a high school and a business school. One panelist was an ETS employee who was involved in test development. The panel had good diversity in terms of age; four were 31–40 years old, seven were 41–50 years old, and three were 51–60 years old. Their English language teaching experience also varied from 6–10 years (*n* = 4), 11–15 years (*n* = 3), 16–20 years (*n* = 2), and 21–25 years (*n* = 2), to above 26 years (*n* = 3). In terms of their familiarity with the CEFR levels, particularly A1–C1, 11 panelists indicated that they were Very Familiar with them, and three were Familiar. In addition, almost all panelists (*n* = 13) reported that they were Very Familiar with the needs and characteristics of young adult and adult English language learners studying English for everyday and workplace purposes, and one indicated that they were Familiar. Panelist name, affiliation, institution type, and country are provided in Appendix A.

Preparation of Panelists

One week before the panel meetings began, panelists received familiarization manuals containing information about the TOEIC Link assessments and the CEFR, as well as preparation activities for panelists to complete online prior to the first panel meeting. The preparation activities included a CEFR sorting activity in which panelists classified CEFR descriptors from relevant scales into their corresponding levels: A1, A2, B1, B2, or C1. Another preparation activity involved panelists taking sample TOEIC Link assessments (listening, reading, speaking, and writing). All panelists successfully completed both preparation activities. Screenshots of the first panelist preparation activity are presented in Appendix B.

During Panel Meeting

The panel meetings took place online, via the Zoom videoconferencing platform. Zoom was selected due to its ease of use and stable functionality across global locations. The panel meetings took place on 3 consecutive days, lasting 7 hours per day. Additional meeting time each day was not possible due to the already extreme starting or ending times in time zones spanning disparate panelist locations.

The first day of panel meetings began with introductions and an overview of the plan for the 3 days. Then, the majority of the first day was spent on standard setting for listening and reading. First, panelists created descriptors for the JQC at CEFR levels A2, B1, B2, and C1 for both listening and reading, as described in detail in the Defining the Just Qualified Candidate section. Then, panelists made cut score judgments (first for listening, then for reading) for each of these CEFR levels using the Bookmark method. On the second day of the meetings, panelists created the JQC descriptors for speaking and then made cut score judgments using the Expected Task Score method. The third day followed the same process as the second day, but for writing.

Defining the Just Qualified Candidate

A key concept in standard setting is that of the just qualified candidate (JQC). The JQC represents a test taker with the absolute minimum knowledge, skills, and abilities necessary to

be classified within a given proficiency category. Thus, the test performance of the JQC functions as a working definition of the cut score at the bottom of that category of proficiency.

Panelists were introduced to the concept of the JQC and were encouraged to keep in mind the difference between a typical test taker in a given proficiency category (scoring near the middle of the score range for that category) and the JQC for that category (scoring at the very bottom of the score range for that category). Then, panelists were asked to write descriptors for the JQC for CEFR levels A2, B1, B2, and C1, focusing on the skill for that day of the study (listening and reading descriptors were drafted in tandem on Day 1, and speaking and writing descriptors were drafted separately on Days 2 and 3, respectively). Sample JQC descriptors were provided to panelists from previous relevant standard setting studies that mapped TOEIC or TOEFL® tests to CEFR levels, to make the descriptor-writing process more efficient and to support congruent conceptualizations of JQCs for CEFR levels across mappings of related tests.

Panelists revised the existing JQC descriptors from the previous studies and wrote new descriptors first as a whole group, for the B2 JQC. Then, panelists were divided into two breakout rooms, according to prearranged assignments that helped ensure balance between the groups in terms of demographics, location, and other background characteristics. The panelists in one breakout room revised and wrote JQC descriptors for CEFR levels A2 and B1, while those in the other breakout room did the same for CEFR level C1. Both groups then returned to the main room to review, discuss, and make final revisions to the JQC descriptors for all four CEFR levels. The final version of the JQC descriptors was emailed to panelists for their use while making judgments and is also presented in Appendix C.

Procedures for Panelist Judgments

The procedures for cut score judgments varied according to the standard setting methods used for each skill. For the TOEIC Link Listening and Reading assessments, panelists followed procedures pertaining to the Bookmark method. First, panelists received an OIB containing the listening or reading items to be used for this study, as detailed earlier in the Selection of Assessment Forms and Items section. Within each OIB, items were ordered by ascending value of theta at RP = 0.67. The rationale for using theta values rather than *b*-

parameters for ordering the OIB was the fact that items may have some differences in sequence between the two methods; therefore, panelists working with *b*-parameters may inadvertently select a lower cut score in some cases by choosing to set a stricter requirement in terms of *b*-parameter, or vice versa. This threat to the valid implementation of panelist judgments is avoided by having panelists work directly with ordering in terms of theta values; this is because it is the theta values of bookmarked items that ultimately establish cut scores within the Bookmark method (Reckase, 2023).

The decision to use the theta value at RP = 0.67 was based on both psychometric and practical considerations; for a 2-PL model such as the one used in this study, this theta value maximizes the information yielded by a correct response, and the concept of a "two-thirds chance of answering an item correctly" is also a straightforward one for panelists to consider (Cizek & Bunch, 2007; Reckase, 2023). However, theta values are by nature very small numbers and frequently negative, often in the range of −2 to 2. Because panelists were chosen for their content expertise and not their familiarity with IRT nor their facility in working with very small numbers, it was decided to provide panelists with ability values that comprised a rounded linear transformation of the theta values at RP = 0.67, for panelist convenience and easier conceptualization. These ability values ranged from 2.5 to 38.6 for listening and from 5.6 to 41.4 for reading, which allowed panelists to work with numbers that had only one decimal place and no negative numbers. Panelists were also provided with *p*-values for items that had comparable values available, in the following format: "For reference, xx% of test takers answered this item correctly." Because some listening and reading items are in sets of two to five items that share a common stimulus (e.g., audio conversation, reading passage), panelists were also provided with reference documents that contained these stimuli and indicated which items relate to which stimuli.

Panelists were then led through the process of making judgments. Beginning with the B2 cut score for listening, the panelists as a group considered an item near the middle of the OIB and judged whether a JQC B2 would have at least a two-thirds chance of answering the item correctly. If they judged that a JQC B2 would have less than a two-thirds chance of answering correctly, then they were encouraged to consider items requiring less ability (i.e.,

previous items in the OIB). If they judged that a JQC B2 would have more than a two-thirds chance of answering correctly, then they were encouraged to consider items requiring more ability (i.e., subsequent items in the OIB). Once they identified the first item that, in their judgment, a JQC B2 would no longer have a two-thirds chance of answering correctly, they were instructed to enter the ability value for this item into a spreadsheet for listening under the column for JQC B2.

During the initial whole-group portion of introducing this judgment procedure, the audio for the listening items being considered and discussed was played for the group of panelists, followed by discussion about what panelists should do if their judgment is that the item requires more/less ability than that of a JQC B2, for it to have a two-thirds chance of being answered correctly. After each item was discussed, panelists could request to next listen to the audio for an item with either a lower or a higher ability value, depending on their judgments of which direction they needed to move in the OIB to approach the ability value of the first item that a JQC B2 would no longer have a two-thirds chance of answering correctly.

After the audio of several items had been played, and after the group had discussed those items together, the panelists expressed readiness to continue the process individually, using audio for items that they could listen to independently via a secure website that was provided to them. So, at this point panelists were asked to submit a Ready to Proceed form, to ensure all panelists were clear on the procedures and ready to continue listening to items and making judgments individually.

Panelists then completed their judgments for listening individually, first finalizing their selection for JQC B2, and then applying the same process to make judgments for JQC B1, A2, and C1. Upon completion of these Round 1 judgments, panelists sent their spreadsheets to the research team, who then compiled and presented this data to the panel for discussion. This included individual judgments (anonymous by panelist number) in addition to descriptive statistics (minimum, maximum, mean, median, mode, and standard deviation) for each of the four judgments (JQC A2, B1, B2, and C1). Panelists then discussed their rationales for their judgments and made notes of items or judgments they might want to revisit in Round 2.

In Round 2, panelists individually revisited judgments in light of the discussion after Round 1, and then sent Round 2 judgments to the research team. Panelists were welcome to keep or change any or all judgments from Round 1 as they saw fit. Panelists were then presented the same individual and summary statistics as in Round 1, in addition to the scale score cuts (out of total scale score of 25) that would ensue from the Round 2 average panel judgments, for consideration in terms of reasonableness and impact.

In Round 3, panelists submitted their final judgments, again with full freedom to keep or change any or all judgments from Round 2. Panelists were then presented with the final individual and summary statistics for listening panel judgments, including final panelist-recommended scale score cuts for listening.

Panelists then repeated this judgment process for reading, starting with an item near the middle of the reading OIB and making judgments first for JQC B2 and then for JQC B1, A2, and C1. The reading judgment process also took place over three rounds, with the same presentation of data and summary statistics after each round, and time for discussion between rounds. Panelists did not require much instruction at the beginning of the process for reading, as it was very similar to the process they had just completed for listening. All rounds of judgments for both listening and reading were completed on the first day of the panel meetings.

For speaking (second day) and writing (third day), the panelists used the Expected Task Score method to make judgments. This method involves judgments about how JQCs would be expected to score on constructed-response items that are scored polytomously using a rubric. After completion of the corresponding JQC descriptors for that day's assessment (speaking or writing), panelists received materials, including, for each task type, a description of the task type, sample item(s), a scoring guide, and a rubric. Panelists were then led through the process of making judgments. For each task type, they were first asked to imagine how 100 JQC B2s would perform on that task type, and to enter the expected number of JQC B2s at each score point (with the requirement that these values must add up to 100). For example, for the Read Aloud item type (scored on a scale of 0 to 4) on the TOEIC Link Speaking assessment, panelists were asked to expect how many of 100 JQC B2s would receive a score of 0 for this task type,

how many would receive a score of 1, how many a 2, how many a 3, and how many a 4. In cases where the assessment design includes multiple items of the same task type with varying difficulty in each assessment form, panelists were asked to consider expectations with respect to overall performance across items of that task type. Each day, panelists completed a Ready to Proceed form after instruction, discussion, and practice of this method, and prior to embarking on Round 1 individual judgments. Appendix D presents the speaking and writing rubrics provided to panelists for their use in making judgments.

Panelists completed three rounds of judgment for speaking and for writing, similar to the process for listening and reading. For Round 1, panelists made individual judgments, first for each task type with respect to 100 JQC B2s and then for each task type with respect to 100 JQC B1s, 100 JQC A2s, and 100 JQC C1s. Panelists then submitted the spreadsheets containing their judgments to the research team. Summary statistics of these panelist judgments were then presented to panelists for the discussion after Round 1, which focused on panelists' rationales for how many JQCs they had placed at each score point for each task type. The summary statistics were the result of a two-step process. First, panelist judgments were simplified by averaging the scores assigned by a given panelist to the 100 JQCs for each task type at each CEFR level. These average score values for each panelist at each level were then fed into the calculations of mean, median, mode, and so forth, across the panel for each task type.

For Round 2, panelists submitted revised judgments, free as always to keep or change any judgments from the previous round. For the Round 2 discussion, summary statistics were presented, in addition to overall scale score cuts if average panelist judgments were to be implemented. The scale score cuts were calculated by feeding the average task-level panelist judgments into the scoring formula for the speaking or the writing test, to identify the scale score that would be earned by a test taker who scored exactly at the average panelist judgment for that task type on every item of the test. The discussion then focused on task-level average scores (not individual score point breakdowns) and their implications for scale-score cuts.

For Round 3, panelists submitted their final judgments, and summary information was presented, including the final panelist-recommended scale score cuts for speaking (second day) and for writing (third day).

At the end of each day, panelists completed a Meeting Evaluation survey.

Standard Setting Results

Tables 9–16 show the results of three rounds of panelist judgments for the TOEIC Link Listening, Reading, Speaking, and Writing assessments respectively. For each assessment, the results are first shown in the units that the panelists worked with directly (ability values for listening and reading; expected task-level scores in rubric-defined score points for speaking and writing). Then, for each assessment, the results from Round 3 are displayed in scale score units (0–25 for each assessment) to provide the final panelist-recommended cut scores for each assessment.

The tables include information about central tendency (i.e., mean, median, mode), maximum and minimum ratings among the 14 panelists, range, standard deviation (*SD*), and standard error of judgment (*SEJ*). The *SEJ* is calculated by dividing the *SD* by the square root of the number of panelists. The *SEJ* estimates the likely variation in average panel judgments among potential panels with similar background and training to the actual panel, indicating consistency in cut score judgments across possible participating panelists (Tannenbaum & Katz, 2013). A small *SEJ* is desirable, and it is hoped that the *SEJ* will decrease across rounds of judgment as a result of increasing panelist agreement (Figueras et al., 2013). The *SEJ* values for this study are discussed following each table with examples from the standard setting results.

Table 9 presents the standard setting results from the three rounds of judgments for listening, all reported in the ability units used by the panelists, except for the rows presenting scale score cuts for Round 2 and Round 3.

Table 9. Standard Setting Results From Three Rounds of Judgments for Listening (Ability Units: 2.5–38.6)

Round	Statistic	A2	B1	B2	C1
1	Mean	8.8	12.4	16.9	20.5
	Median	8.6	12.5	17.2	20.0
	Mode	8.6	11.8	14.3	20.0
	Maximum	13.4	18.4	21.9	25.4
	Minimum	3.6	7.3	7.5	12.8
	Range	9.8	11.1	14.4	12.6
	SD	3.13	2.85	3.68	3.32
	SEJ	0.78	0.71	0.92	0.83
2	Mean	8.5	12.3	17.6	21.6
	Median	8.6	12.2	17.2	21.1
	Mode	8.6	12.2	17.2	20.0
	Maximum	13.4	18.8	21.9	25.4
	Minimum	3.6	8.6	14.3	19.5
	Range	9.8	10.2	7.6	5.9
	SD	2.22	2.36	1.84	1.87
	SEJ	0.55	0.59	0.46	0.47
	Scale score cut	9	14	18	20
3	Mean	8.0	12.3	18.2	22.4
	Median	8.6	12.2	17.2	21.9
	Mode	8.6	12.2	17.2	21.9
	Maximum	8.6	14.3	23.7	27.5
	Minimum	3.6	9.6	17.2	20.2
	Range	5.0	4.7	6.5	7.3
	SD	1.39	1.04	1.75	1.89
	SEJ	0.35	0.26	0.44	0.47
	Scale score cut	8	14	18	21

Agreement among panelists (as measured by *SD* and *SEJ*) increased for each of the four judgments across the three rounds, with the only exception being the C1 judgment, for which the *SD* increased from 1.87 in Round 2 to 1.89 in Round 3. This increase may have been related to a shift toward somewhat higher values being chosen by some panelists for this judgment in Round 3. After Round 2, there was some discussion around the C1 scale score cut of 20 resulting from the Round 2 panelist judgments, and how this cut score was relatively close to the B2 cut (18) while still being some distance from the top of the score scale (25). After observing the spacing of these proposed cut scores, some panelists reconsidered their judgments for B2 and C1 and ultimately chose to select a higher ability value for their C1 judgment, which in turn resulted in a scale score cut of 21 for C1 in Round 3.

The other listening judgment that had a shift in the scale-score cut between Round 2 and Round 3 was for A2. In this case, there was no issue with proximity to another level's cut score, but there was a wide range of panelists' judgments in ability units during the first two rounds. In particular, the maximum score represented an outlier judgment of 13.4, when the mean judgments across panelists was just 8.8 in Round 1 and 8.5 in Round 2. After the Round 2 discussion, the maximum panelist judgment for A2 in Round 3 was 8.6, which greatly reduced the range of panelist judgments for A2 and ultimately resulted in an A2 scale score cut of 8 in Round 3, where it had been 9 in Round 2. Meanwhile, the B1 and B2 average ability judgments shifted slightly over the three rounds, but to a small enough degree that the corresponding scale score cuts did not change at all between Round 2 and Round 3.

Table 10 presents information about the listening Round 3 judgments in terms of scale score units (0–25).

Table 10. Standard Setting Results From Round 3 for Listening (Scale Score Units: 0–25)

Statistic	A2	B1	B2	C1
Mean	8	14	18	21
Median	10	14	18	21
Mode	10	14	18	21
Maximum	10	16	22	23
Minimum	5	11	18	20
Range	5	5	4	3
SD	1.44	1.10	1.08	0.83
n	14	14	14	14
SEJ	0.38	0.29	0.29	0.22

Agreement among panelists was higher for the higher-level cut scores than for the lower-level cut scores; the *SD* was largest for the A2 cut score (1.44) and smallest for the C1 cut score (0.83). However, the *SD* and *SEJ* for all four cut scores were reasonably small. Even the largest *SEJ*, 0.38 for the A2 cut score, represents a likely variation of just 0.38 points for this cut score judgment among possible panels with similar background and training. This is less than half of a scale score point, making it likely that another panel would make the same judgment in terms of the A2 cut score (expressed in whole-number scale score units). The *SEJ*s for the B1, B2, and C1 cut scores are even smaller, indicating that the same judgment in scale-score units from another panel would be even more likely.

Table 11 presents the results of the three rounds of judgment for reading, in the ability units used by the panelists.

Table 11. Standard Setting Results From Three Rounds of Judgments for Reading (Ability Units: 5.6–41.4)

Round	Statistic	A2	B1	B2	C1
1	Mean	8.8	13.4	19.9	27.2
	Median	8.3	12.2	17.8	26.4
	Mode	8.3	12.2	17.3	30.6
	Maximum	12.2	18.2	28.4	41.4
	Minimum	5.6	11.4	14.9	18.8
	Range	6.6	6.8	13.5	22.6
	SD	1.78	2.44	4.02	6.29
	SEJ	0.45	0.61	1.01	1.57
2	Mean	8.5	12.5	20.1	29.3
	Median	8.3	12.1	20.7	30.6
	Mode	8.3	11.4	20.7	30.6
	Maximum	10.6	15.7	24.3	34.9
	Minimum	5.6	11.4	17.2	22.7
	Range	5.0	4.3	7.1	12.2
	SD	1.22	1.39	2.38	2.78
	SEJ	0.31	0.35	0.60	0.69
	Scale score cut	6	11	18	22
3	Mean	8.4	12.2	20.0	30.0
	Median	8.3	12.0	20.7	30.6
	Mode	8.3	11.4	20.7	30.6
	Maximum	10.6	15.7	24.3	34.9
	Minimum	5.6	11.4	17.2	27.3
	Range	5.0	4.3	7.1	7.6
	SD	1.13	1.18	2.16	1.94
	SEJ	0.28	0.30	0.54	0.49
	Scale score cut	6	11	18	22

For this assessment, the *SD* and *SEJ* decreased—sometimes markedly—for all four judgments, across all three rounds, indicating increasing panelist agreement as the process unfolded. In this case, the scale-score cuts resulting from the Round 2 panelist judgments were all spaced at least four scale score points apart, and panelist judgments for Round 3 resulted in the same four scale score cuts as for Round 2.

Table 12 presents information about the cut score judgments for reading from Round 3 in terms of scale-score units.

Table 12. Standard Setting Results From Round 3 for Reading (Scale Score Units: 0–25)

Statistic	A2	B1	B2	C1
Mean	6	11	18	22
Median	6	11	19	23
Mode	6	10	19	23
Maximum	9	15	21	23
Minimum	4	10	16	22
Range	5	5	5	1
SD	1.22	1.44	1.61	0.50
n	14	14	14	14
SEJ	0.32	0.38	0.43	0.13

In this case, the *SD* and *SEJ* showed the highest degree of panelist agreement for the C1 cut score and the lowest degree of panelist agreement for the B2 cut score. However, even the B2 cut score had an *SEJ* (0.43) of less than half of a scale-score point, indicating likelihood that another panel would result in the same scale-score cuts for reading as well.

Table 13 presents the standard setting results for the three rounds of panelist judgments for speaking, specific to each speaking task type and using rubric score point units. Each panelist's judgment regarding the expected scores of 100 JQCs was first converted into an average JQC score for that task type, prior to the calculation of the summary statistics across all the panelists as presented in Table 13.

For each task type, the *SD* and *SEJ* decreased or remained steady for each judgment over the three rounds, indicating increasing panelist agreement overall. The mean values for judgments remained fairly stable across rounds. There is also a general pattern of more agreement (smaller *SD* and *SEJ*) for the higher-level judgments (e.g., C1) than for the lower-level ones (e.g., A2), although this is not exactly present in every case. It is also notable that the panelist judgments generally spanned most of the score range for each task type, indicating that the panelists found an overall match between the range of ability delineated by the rubric-specified score points and the range of ability represented by CEFR levels A1–C1.

Table 13. Standard Setting Results From Three Rounds of Judgments for Speaking (Task-Specific Units)

Statistic	Round 1				Round 2				Round 3			
	A2	B1	B2	C1	A2	B1	B2	C1	A2	B1	B2	C1
Read Aloud (Rating: 0–4)												
Mean	1.6	2.4	3.1	3.8	1.7	2.5	3.1	3.8	1.7	2.6	3.2	3.8
Median	1.5	2.5	3.3	3.8	1.7	2.6	3.2	3.8	1.7	2.6	3.2	3.8
Maximum	2.6	3.3	3.8	4.0	2.0	2.8	3.7	4.0	2.0	2.8	3.6	4.0
Minimum	1.1	1.8	2.5	3.4	1.4	2.3	2.7	3.6	1.4	2.4	2.8	3.7
Range	1.5	1.5	1.3	0.6	0.6	0.5	1.0	0.4	0.6	0.5	0.8	0.3
SD	0.4	0.4	0.4	0.2	0.2	0.2	0.3	0.1	0.2	0.1	0.2	0.1
SEJ	0.1	0.1	0.1	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.0
Listen and Repeat (Rating: 0–5)												
Mean	1.7	2.6	3.6	4.4	1.8	2.7	3.7	4.5	1.8	2.7	3.7	4.5
Median	1.6	2.6	3.6	4.6	1.8	2.7	3.6	4.5	1.8	2.7	3.7	4.6
Maximum	2.4	3.4	4.5	4.9	2.6	3.3	4.4	4.9	2.6	3.3	4.4	4.9
Minimum	1.1	1.9	3.0	3.8	1.3	2.3	3.4	4.1	1.3	2.5	3.4	4.1
Range	1.4	1.6	1.5	1.1	1.3	1.0	1.0	0.8	1.3	0.9	1.1	0.8
SD	0.4	0.4	0.4	0.3	0.3	0.3	0.3	0.2	0.3	0.2	0.3	0.2
SEJ	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Opinion (Rating: 0–5)												
Mean	1.6	2.6	3.6	4.5	1.7	2.7	3.7	4.6	1.7	2.7	3.7	4.6
Median	1.7	2.7	3.7	4.6	1.8	2.7	3.7	4.6	1.8	2.7	3.7	4.6
Maximum	2.4	3.3	4.3	4.8	2.6	3.4	4.0	4.8	2.6	3.4	4.0	4.8
Minimum	0.5	1.6	2.5	3.4	1.1	2.0	3.2	4.2	1.1	2.0	3.2	4.2
Range	1.9	1.7	1.8	1.4	1.5	1.4	0.8	0.6	1.5	1.4	0.8	0.6
SD	0.6	0.4	0.4	0.4	0.4	0.4	0.2	0.2	0.4	0.3	0.2	0.2
SEJ	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.1	0.0

Table 14 presents the Round 3 results for speaking in scale score units (0–25). The *SD* and *SEJ* values indicate highest panelist agreement for the C1 cut score and lowest panelist agreement for the A2 cut score, although—as with the other assessments—the *SEJ* value for the A2 cut score (0.44) is still less than half of a scale score point, indicating that another panel would likely recommend the same cut scores in scale score units. The speaking cut scores are each spaced at least four scale-score points apart.

Table 14. Standard Setting Results From Round 3 for Speaking (Scale Score Units: 0–25)

Statistic	A2	B1	B2	C1
Mean	6	12	18	22
Median	6	12	18	23
Mode	6	12	18	23
Maximum	11	15	20	24
Minimum	4	11	16	21
Range	7	4	4	3
SD	1.65	1.14	1.19	1.02
n	14	14	14	14
SEJ	0.44	0.30	0.32	0.27

Table 15 presents the three rounds of writing results in rubric score point units specific to each task type. Each panelist's judgment regarding the expected scores of 100 JQCs was first converted into an average JQC score for that task type, prior to the calculation of the summary statistics across all the panelists as presented in Table 15.

As with speaking, the mean panelist judgments for each task type were similar across rounds, while panelist agreement for each judgment increased or remained steady across rounds. The *SD*s and *SEJ*s were higher overall in Round 1 for the lower level judgments (e.g., A2) than for the higher level ones (e.g., C1), but by Round 3 the *SD*s and *SEJ*s were similarly low for all judgments, indicating increasing panelist agreement for the most initially differing judgments.

Panelists utilized almost the full range of task-specific scores for each task type, indicating that for writing (as with speaking) there seems to be a match between the range of ability specified within the task-specific rubrics and the range of ability described by CEFR levels A1–C1.

Table 15. Standard Setting Results From Three Rounds of Judgments for Writing (Task-Specific Units)

Statistic	Round 1				Round 2				Round 3			
	A2	B1	B2	C1	A2	B1	B2	C1	A2	B1	B2	C1
Build a Sentence (Rating: 0–2)												
Mean	1.2	1.4	1.7	1.9	1.2	1.5	1.7	1.9	1.2	1.5	1.7	1.9
Median	1.1	1.4	1.7	1.9	1.2	1.4	1.7	1.9	1.2	1.4	1.7	1.9
Maximum	1.4	1.8	2.0	2.0	1.4	1.8	1.9	2.0	1.3	1.8	1.9	2.0
Minimum	1.0	1.3	1.5	1.7	1.1	1.3	1.5	1.7	1.1	1.3	1.5	1.7
Range	0.4	0.5	0.5	0.3	0.3	0.5	0.4	0.3	0.2	0.5	0.4	0.3
SD	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
SEJ	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Write a Brief Email (Rating: 0–5)												
Mean	1.8	2.8	3.8	4.6	1.8	2.8	3.9	4.6	1.7	2.7	3.9	4.6
Median	1.8	2.8	3.8	4.7	1.8	2.8	3.9	4.7	1.8	2.8	3.9	4.7
Maximum	2.9	3.6	4.2	4.9	2.6	3.4	4.2	4.8	2.1	3.1	4.2	4.8
Minimum	1.2	2.3	3.2	4.0	1.2	2.4	3.5	4.1	1.2	2.4	3.5	4.1
Range	1.7	1.4	1.0	1.0	1.4	1.0	0.7	0.7	0.9	0.7	0.7	0.7
SD	0.4	0.4	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2
SEJ	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Write an Extended Email (Rating: 0–5)												
Mean	1.6	2.7	3.7	4.5	1.7	2.7	3.7	4.5	1.6	2.7	3.7	4.5
Median	1.6	2.7	3.7	4.5	1.6	2.7	3.7	4.6	1.7	2.7	3.7	4.5
Maximum	2.9	3.6	4.1	4.8	2.7	3.4	4.0	4.7	2.0	3.1	4.0	4.7
Minimum	1.2	1.9	3.0	3.9	1.2	2.2	3.4	4.2	1.2	2.2	3.4	4.2
Range	1.7	1.8	1.1	0.9	1.5	1.2	0.6	0.6	0.9	1.0	0.6	0.5
SD	0.4	0.4	0.3	0.3	0.3	0.3	0.2	0.2	0.2	0.3	0.2	0.2
SEJ	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.0
Write for an Online Discussion (Rating: 0–5)												
Mean	1.5	2.4	3.5	4.4	1.5	2.5	3.6	4.4	1.5	2.5	3.6	4.4
Median	1.4	2.3	3.6	4.4	1.5	2.5	3.6	4.4	1.5	2.5	3.6	4.4
Maximum	2.6	3.5	4.0	4.7	2.5	3.1	3.8	4.6	2.1	2.8	3.8	4.6
Minimum	0.7	1.9	2.8	4.0	1.1	2.1	3.4	4.1	1.1	2.1	3.4	4.2
Range	1.9	1.6	1.3	0.7	1.4	1.0	0.4	0.5	1.0	0.7	0.4	0.4
SD	0.5	0.5	0.4	0.2	0.3	0.3	0.1	0.1	0.3	0.2	0.1	0.1
SEJ	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.0

Table 16 presents the writing cut score results in scale score units.

Table 16. Standard Setting Results From Round 3 for Writing (Scale Score Units: 0–25)

Statistic	A2	B1	B2	C1
Mean	8	13	18	22
Median	8.5	13	18	22
Mode	9	13	18	22
Maximum	10	15	21	24
Minimum	6	11	17	21
Range	4	4	4	3
SD	1.14	0.95	1.08	0.85
n	14	14	14	14
SEJ	0.30	0.25	0.29	0.23

The mean results place all four cut scores at least four scale-score points apart from one another, and the *SEJ*s are all small, ranging from 0.23 to 0.30. Thus, across all four assessments (listening, reading, speaking, and writing), the *SEJ*s indicate that it is likely another panel with similar background and training would have made judgments resulting in the same whole-point scale score cuts.

Poststudy Adjustments

Standard setting studies generally rely primarily on the judgments of a panel of trained experts in order to establish cut scores. However, there may also be reasons to make small adjustments to the panelist-recommended cut scores for operational use.

Some considerations that may be relevant for any proposed poststudy adjustments to cut scores include the following (Geisinger & McCormick, 2010; Mills, 1995):

- the impact (relative seriousness) of false positive and false negative classification errors
- results from different standard setting studies or methods
- observed score distributions
- organizational or societal needs
- opportunities to retest

Tannenbaum and Baron (2015) also recommend raising or lowering cut scores by up to one standard error of measurement (SEM) according to the application of relevant considerations to the standard setting situation and decision makers' needs.

For the TOEIC Link, many of these considerations are difficult to analyze due to the fact that the assessments are new and without operational data at the time of writing this report. Therefore, poststudy adjustments are limited to the category of "results from different standard setting studies or methods," insofar as all four assessments comprising the TOEIC Link draw from items and task types also found on other English proficiency tests, which are themselves already mapped to the CEFR. For listening and reading, the IRT theta value for each cut score can be compared between the TOEIC Link and the TOEIC; see Tannenbaum and Wylie (2008) for the study that maps TOEIC scores to the CEFR levels. For speaking, mean task-level scores for test takers at each cut score can be compared between the TOEIC Link and other English language proficiency tests for each task type (Read Aloud, Listen and Repeat, and Express an Opinion); see Davis, Garcia Gomez, et al. (2023) for the study that maps TOEFL Essentials scores to the CEFR levels, and Papageorgiou et al. (2021) for the TOEFL Essentials design framework. For writing, mean task-level scores for test takers at each cut score can also be compared between the TOEIC Link and other English language proficiency tests for each task type (Build a Sentence, Brief Email, Extended Email, and Online Discussion); see Schmidgall (2021) for the study that maps the TOEIC Bridge scores to the CEFR levels, and Davis, Garcia Gomez, et al. (2023) and Papageorgiou et al. (2021) for TOEFL Essentials.

Listening

Because the TOEIC Link Listening assessment draws from TOEIC Listening items that are already calibrated using IRT, it is possible to use theta values to compare scores between the two tests. Table 17 presents the panelist-recommended TOEIC Link Listening cut scores from this study, in comparison to the score that matches the theta value at RP = 0.67 that corresponds to the CEFR cut scores for TOEIC Listening, as previously mapped by another study (Tannenbaum & Wylie, 2008). The final row of Table 17 presents the research team revision for the CEFR A2, B1, B2, and C1 cut scores for the TOEIC Link Listening assessment.

Table 17. Recommended TOEIC Link Listening Cut Scores in Context

Source of recommendation	A2 cut score	B1 cut score	B2 cut score	C1 cut score
TOEIC Link panelist judgments	8	14	18	21
Comparison to TOEIC	5	13	19	24
Research team revision	8	14	18	22

As shown in Table 17, the cut scores for B1 and B2 are very close between the panelist judgments for this study and the corresponding previous TOEIC mapping. The difference of one point is small, showing solid congruence between the two mappings. The TOEIC Listening mapping was performed in 2008, and some small updates to the TOEIC Listening assessment have taken place since that time, so for these two cut scores it seems preferable to prioritize the judgments of the panel for this current study while considering the implications of the existing mapping between TOEIC and CEFR.

For the A2 cut score, there is a three-point difference between the current panel's recommendation (8) and the comparable TOEIC Listening cut score (5). After some discussion, it was noted that a scale score of 5 is likely to be at or below the chance score (a score expected to be achieved just by random guessing), given the composition of the TOEIC Link Listening assessment with multiple-choice items with three to four response options each. Considering the high levels of agreement among our panelists, we decided to prioritize the panelist-recommended A2 cut score of 8 for the TOEIC Link Listening.

The C1 cut score also has a three-point difference between the current panel's recommendation (21) and the corresponding TOEIC Listening cut score (24). In the standard setting study, panelists were shown how their judgments translated into cut scores after Round 2, which prompted discussion among the panelists as to whether the C1 cut score may be too low. After this discussion, some panelists revised their C1 judgments to require a higher ability value, resulting in an increase in the C1 scale score cut from 20 in Round 2 to 21 in Round 3. The Round 3 value was still only three scale score points away from the B2 scale score cut (18). Therefore, the research team recommends a small adjustment to the panelist-recommended C1 cut score, shifting it from 21 to 22, both in accordance with the panel's sense that the C1 cut score needed to be higher, and to promote congruence among CEFR mappings for the TOEIC family of tests.

Reading

As with listening, the TOEIC Link Reading assessment draws from TOEIC Reading items that are already calibrated using IRT, so here too it is possible to use theta values to compare scores between the two tests. Table 18 presents the panelist-recommended TOEIC Link Reading cut scores from this study, in comparison to the score that matches the theta value at RP = 0.67 that corresponds to the CEFR cut scores for TOEIC Reading, as previously mapped by another study (Tannenbaum & Wylie, 2008). The final row of Table 18 presents the research team revision for the CEFR A2, B1, B2, and C1 cut scores for the TOEIC Link Reading assessment.

Table 18. Recommended TOEIC Link Reading Cut Scores in Context

Source of recommendation	A2 cut score	B1 cut score	B2 cut score	C1 cut score
TOEIC Link panelist judgments	6	11	18	22
Comparison to TOEIC	6	14	19	23
Research team revision	6	12	18	22

As shown in Table 18, the panelist-recommended TOEIC Link Reading cut scores for A2, B2, and C1 are all no more than one point away from the corresponding TOEIC Reading cut scores. Therefore, the research team decided to fully follow the recommendations of the panel for the current study for these three cut scores. For the B1 cut score, however, there is a three-point difference between the panelist-recommended TOEIC Link Reading cut score (11) and the corresponding TOEIC Reading cut score (14). In this case, there is also no issue with the spacing of cut scores for adjacent levels, as the panelist-recommended TOEIC Link Reading cut score for B2 (18) is seven points higher than for B1 (11). After discussion, the research team recommends the B1 cut score to be 12, just one point higher than the panelist-recommended value, while also supporting coherence among CEFR mappings within the TOEIC family of tests.

Speaking

For speaking, the comparison to other mappings is based on overlapping task types between the TOEIC Link Speaking and other English language proficiency tests. Each of the three task types in the TOEIC Link Speaking (Read Aloud, Listen and Repeat, and Express an Opinion) is shared with one of two other tests, designated as reference tests for this study. These shared task types all have very similar specifications, the same range of score points, and

similar rubrics. However, it should be noted as a caveat that the specifications and the rubrics are not identical in all cases. Therefore, the comparisons that follow should be treated as exploratory, not definitive; and emphasis has been placed on the judgments of the panel for the current study, as this is the only panel that has made judgments on the basis of the exact specifications and rubrics of the TOEIC Link Speaking assessment.

Operational data were obtained for 107 test takers who took a reference test containing items very similar or identical to the Read Aloud items and the Listen and Repeat items used by the TOEIC Link panelists to make their judgments. Matching items were used in this way in order to limit any source of variation in terms of the relative difficulty of individual items within a certain task type. For the Express an Opinion task type, operational data were obtained for 20 test takers who took a reference test containing an item very similar or identical to the one used in this study. For both data sets, test takers were sorted by total score, which was then categorized by CEFR level as per the pre-existing mappings of the reference tests to the CEFR. Then, for test takers with total scores at the very bottom of CEFR levels A2, B1, B2, and C1, mean scores on the overlapping items were calculated. This approach provides a rough way to find the expected task-level scores on the targeted items for test takers whose overall abilities are right at the cut scores for each CEFR level. Again, the items and rubrics were not absolutely identical in all cases, and the targeted construct may vary somewhat among these English proficiency tests, making these direct comparisons more exploratory than definitive. Limited available data is another caveat. However, in the absence of operational data for the TOEIC Link, this approach allows for some cross-checking of panelist judgments against some form of real-world test-taker performance data.

Table 19 presents, for each TOEIC Link Speaking task type, the Round 3 panelist judgments for that task type (rows labeled TOEIC Link in the Test column), in comparison with average scores on each task type from operational data of reference test takers with overall scores at the cut score for that CEFR level (rows labeled Reference test in the Test column).

Table 19. Panelist-Recommended TOEIC Link Speaking Cut Scores Compared to Reference Tests

Task type	Test	Average item score at A2 cut	Average item score at B1 cut	Average item score at B2 cut	Average item score at C1 cut
Read Aloud	TOEIC Link	1.7	2.6	3.2	3.8
	Reference test	2.4	2.7	3.2	3.5
Listen and Repeat	TOEIC Link	1.8	2.7	3.7	4.5
	Reference test	2.1	2.9	3.7	4.2
Express an Opinion	TOEIC Link	1.7	2.7	3.7	4.6
	Reference test	2.1	2.9	3.4	4.1

It is evident from Table 19 that the operational data across task types show that real-world test takers performed similarly to the panelist judgments at the B1 and B2 cut scores. In fact, this close correspondence for these two levels in the middle can be seen as support for the overall congruence between panelist judgments and operational test-taker performance. However, there is some divergence at both the A2 and the C1 cut scores. For A2, panelist judgments were lower than operational test-taker performance across all three task types. For C1, panelist judgments were higher than operational test-taker performance across all three task types. These two phenomena are seen as connected, pointing perhaps to more of an ideal differentiation among test-taker abilities in the minds of panelists, while real-world test-taker scores tend to demonstrate more regression to the mean. For example, the C1 panelist judgments for the Listen and Repeat task type and the Express an Opinion task type were both within half a point of the maximum possible score for these item types (4.5 and 4.6, respectively, out of a possible 5 points). This close proximity to the maximum possible score may be a bit unrealistic, even for C1 JQC test takers. It is important for cut scores to not be overly extreme at either end of the score scale. This avoids the risk of systematically producing false-positive classifications at the lower end (i.e., classifying A1 test takers as A2 due to their scores being closer to the middle), while also systematically producing false-negative classifications at the upper end (i.e., classifying C1 test takers as B2 due to their scores also being closer to the middle).

In light of these considerations, the research team made a poststudy adjustment of the panelist-recommended A2 and C1 cut scores to both be one point closer to the middle of the

score scale. Table 20 presents the panel's recommendation and the research team's revisions for the TOEIC Link Speaking cut scores for CEFR levels A2, B1, B2, and C1.

Table 20. Recommended TOEIC Link Speaking Cut Scores After Adjustments

Source of recommendation	A2 cut score	B1 cut score	B2 cut score	C1 cut score
TOEIC Link panelist judgments	6	12	18	22
Research team revision	7	12	18	21

Writing

The poststudy adjustments for the TOEIC Link Writing assessment followed a similar process as for the TOEIC Link Speaking assessment. Operational data were accessed for identical or very similar items found in reference tests for the Build a Sentence (*n* = 27), Brief Email (*n* = 61), Extended Email (*n* = 60), and Online Discussion (*n* = 59) task types. Unfortunately, limited available data was an issue here, as with speaking, but clear patterns were still evident.

Table 21 presents, for each TOEIC Link Writing task type, the Round 3 panelist judgments for that task type (rows labeled TOEIC Link in the Test column), in comparison with average scores on each task type from operational data of reference test takers with overall scores at the cut score for that CEFR level (rows labeled Reference test in the Test column).

Table 21. Panelist-Recommended TOEIC Link Writing Cut Scores Compared to Reference Tests

Task type	Test	Average item score at A2 cut	Average item score at B1 cut	Average item score at B2 cut	Average item score at C1 cut
Build a Sentence	TOEIC Link	1.2	1.5	1.7	1.9
	Reference test	1.0	2.0		
Brief Email	TOEIC Link	1.7	2.7	3.9	4.6
	Reference test	1.6	2.2	3.3	4.0
Extended Email	TOEIC Link	1.6	2.7	3.7	4.5
	Reference test	1.7	2.6	3.4	4.0
Online Discussion	TOEIC Link	1.5	2.5	3.6	4.4
	Reference test	1.6	2.4	3.3	3.9

Note. The reference test for the Build a Sentence task type is not mapped to Common European Framework of Reference for Languages (CEFR) levels higher than B1, and so the cells corresponding to the B2 and C1 cuts for this task type are blank.

As shown in Table 21, the cut scores with the most similarity between the panelist judgments and the operational data from reference tests are A2 and B1; the operational data

show higher mean scores than the panelist judgments for some task types, and lower mean scores for others, but the task-level scores at these two cuts seem roughly equivalent overall. Meanwhile, the comparisons at B2 and C1 show a pattern with panelist judgments consistently higher than test-taker performance from the other tests with overlapping task types. This pattern has some similarities to speaking, in that panelist judgments tended to be more extreme than operational data, but in this case the affected cut scores are the two highest ones (B2 and C1), rather than the lowest and the highest ones as in speaking (A2 and C1). The disparity is greatest for the C1 cut score, but is clearly present for the B2 cut score as well.

In light of this pattern, and given the same consideration raised regarding speaking (regression to the mean in real-world data), the research team decided to lower both the B2 and the C1 panelist-recommended cut scores, by one point each, for operational use. Table 22 presents the panel's recommendation and the research team's revisions for the TOEIC Link Writing cut scores for CEFR levels A2, B1, B2, and C1.

Table 22. Recommended TOEIC Link Writing Cut Scores After Adjustments

Source of recommendation	A2 cut score	B1 cut score	B2 cut score	C1 cut score
TOEIC Link panelist judgments	8	13	18	22
Research team revision	8	13	17	21

With all poststudy adjustments complete, Table 23 presents the research team's recommendation for the TOEIC Link CEFR cut scores for CEFR levels A2, B1, B2, and C1.

Table 23. Summary of Recommended CEFR Cut Scores for TOEIC Link Assessments Following Poststudy Adjustments

Assessment	A2 cut score	B1 cut score	B2 cut score	C1 cut score
Listening	8	14	18	22
Reading	6	12	18	22
Speaking	7	12	18	21
Writing	8	13	17	21

In a final step, cut score recommendations were aligned across skills. This step was taken to simplify the interpretation and use of test scores while preserving the original cut score recommendations as shown in Table 23. To establish recommended cut scores that would be common across assessments (e.g., the same numeric score corresponding to the A2 cut

score for all four TOEIC Link assessments), we used a central tendency or balanced approach (i.e., estimated the mean) while rounding up to the nearest scale score. This helped minimize the number of adjustments needed for each of the base scales while incorporating a slight preference for reducing false-positive classification errors.[1] Psychometricians then adjusted the base scale for the four measures while preserving the original cut score recommendations, which resulted in the final recommended cut scores shown in Table 24.

Table 24. Summary of Final Recommended Common European Framework of Reference for Languages (CEFR) Cut Scores for TOEIC Link Assessments

Assessment	A2 cut score	B1 cut score	B2 cut score	C1 cut score
Listening	8	13	18	22
Reading	8	13	18	22
Speaking	8	13	18	22
Writing	8	13	18	22

Validity Evidence

When examining the results of a standard setting study, it is important to consider validity evidence. Standard setting results are typically supported by three categories of validity evidence: procedural, internal, and external (Hambleton & Pitoniak, 2006; Kane, 2001; Tannenbaum & Cho, 2014). In this section, we present procedural and internal validity evidence for this study and end with recommendations for the collection of external validity evidence once the TOEIC Link assessment is launched operationally.

Procedural Validity

Evidence to support procedural validity corroborates the clarity, appropriateness, and fairness of the standard setting processes used to generate the recommended cut scores. For this study, the process used to generate recommended cut scores included a standard setting study and poststudy adjustment procedure, as documented earlier in this report. Documentation of study materials and protocols, as well as the description of standard setting procedures the panel experienced, can be one source of procedural validity; another source of evidence can come from panelists themselves, such as ratings on facilitator explanations, satisfaction level with the suggested cut scores, and feedback on the entire meeting procedures (Hambleton & Pitoniak, 2006). The earlier section on the standard setting procedure offers

evidence for the procedural validity given that it includes detailed descriptions of the pre-meeting familiarization tasks and the judgment procedures using the Bookmark method as well as the Expected Task Score method. In addition, a detailed description of the selection of panelists provides evidence for the procedural validity, emphasizing that the panel comprised experts who had relevant experience and expertise in teaching English to the target test takers for the TOEIC Link assessments and were fairly diverse in terms of biographic and geographic backgrounds.

Another important source of procedural validity comes from the feedback given by the panelists themselves. We collected feedback from panelists via online surveys at different times during the standard setting meetings. The Ready to Proceed survey was filled out at an early stage of the standard setting meetings: It was given to panelists after they received explanations and training on the specific standard setting procedure (e.g., Bookmark method), but prior to the first round of cut score judgments. The Ready to Proceed survey asked a series of questions regarding panelists' understanding of the standard setting procedure and their level of satisfaction with the training they received. They were also asked to indicate their readiness for the upcoming judgment task and to enter any comments regarding what kind of additional explanation or training they might need to be fully prepared for the judgment tasks. Another survey, the Meeting Evaluation survey, was administered toward the end of the standard setting meeting to collect panelist feedback after they had completed the judgment tasks. The Meeting Evaluation survey included a set of questions asking the degree to which standard setting procedures were clear and easy to follow, how comfortable panelists were with the panel's recommended cut scores generated after three rounds of group discussions and judgments, and the extent to which TOEIC Link assessment tasks elicited evidence for the English language skills across the four CEFR levels.

Table 25 summarizes the results of the Ready to Proceed survey that was administered to ensure that panelists received sufficient explanations and training prior to the first round of cut score judgments. Panelists indicated their level of agreement with six statements that focused on different aspects of the explanations and training they received, such as their overall understanding of the study purpose and the definition of JQC, the clarity of facilitator

explanations, the effectiveness of the training, and their readiness for the judgment tasks. Table 25 shows the average ratings for the statements in the TOEIC Link Listening and Reading, Speaking, and Writing assessments. Given that the training for the listening and reading assessments was given in one sitting, the panelists were asked to indicate their rating for the two assessments combined. The average ratings for the statements were relatively high across different assessments, with the lowest rating being for the statement, "The training in the standard setting method adequately prepared me to make my standard setting judgments," for the listening and reading assessments (3.4). A potential reason for the lower average ratings for the listening and reading assessments may be that this was when panelists had received explanations for standard setting methods and training for the first time in this standard setting study. In fact, average ratings for the same statement in the later assessments (speaking and writing) were quite high (3.9 for both); panelist familiarity with the standard setting study procedures might have contributed a higher level of agreement for the same statement. Panelists generally had a good understanding of the study purpose (ranging from 3.8 to 4.0) and the definition of JQC (3.6–4.0), which laid the foundation for them to fully engage in the cut score judgment procedures. In addition, panelists indicated that they had an adequate level of practice (3.8–4.0) and understanding of how the standard setting judgments are done (3.6–3.9).

Table 25. Average Ratings on Panelists' Perceptions About Training

Statement	Listening and Reading	Speaking	Writing
I understand the purpose of the study.	3.8	4.0	4.0
The facilitator explained things clearly.	3.5	4.0	4.0
I understand the definition of the Just Qualified Candidate (JQC).	3.6	4.0	3.9
The training in the standard setting method adequately prepared me to make my standard setting judgments.	3.4	3.9	3.9
The opportunity to practice helped clarify the standard setting task for me.	3.8	3.9	4.0
I understand how to make the standard setting judgments.	3.6	3.9	3.9

Note. 1 = *strongly disagree*; 2 = *disagree*; 3 = *agree*; 4 = *strongly agree*

The ready-to-proceed form concluded with a question, “Are you ready to proceed and to make your first standard setting judgment for Listening/Reading/Speaking/Writing?” (see Table 26). If the answer was no, we asked the follow-up question, “If No, what other information or explanation do you need before making your first standard setting judgment?” For the listening and reading assessments, all panelists but one selected Yes for this question. One panelist who answered no indicated that they needed more examples of judgment tasks. We addressed their concerns and ensured that they were ready to proceed. For the speaking and writing assessments, all panelists selected Yes for this question.

Table 26. Number of Panelists With Readiness for the Judgment Task

Question	Listening		Reading		Speaking		Writing	
	Yes	No	Yes	No	No	No	Yes	No
Are you ready to proceed and to make your first standard setting judgment?	13	1	13	1	14	0	14	0

Another form of evidence for procedural validity is based on the feedback gained through the Meeting Evaluation survey, which was administered at the end of each meeting day (a total of three Meeting Evaluation surveys for the study). The survey asked a series of questions regarding their entire experience with the standard setting study, starting from the preparation activities they completed independently prior to the whole-group meeting and covering various aspects of the meeting procedures. Table 27 presents the results of the survey. All panelists indicated that they strongly agreed with the statements, “I understood the purpose of the study,” and “feedback and discussion between judgement task rounds was helpful” (an average of 4.0). Other statements also received a very high total average rating with almost all panelists expressing that they strongly agreed or agreed. Particularly, panelists felt that the instructions and explanations for the process of the judgment task and how the recommended cut score was determined were clear to them (an average of 3.9), except for the statement for listening and reading on the first day: “The instructions and explanations provided by the facilitators were clear” (an average of 3.7). The only statement that received a rating of Disagree was from one panelist was for the statement, “The CEFR familiarization activity (sorting CEFR descriptors in the online survey) was helpful” (an average of 3.7 to 3.8).

This familiarization activity was a pre-meeting assignment that they had to complete independently; while the exact reason for the low rating by this particular panelist is unknown, it may be because they were already familiar with the CEFR levels, which might have caused the assignment to be perceived as less effective in preparing for the whole-group meeting.

Table 27. Panel's Feedback on Overall Meeting Procedure

Statement	Listening and Reading	Speaking	Writing
I understood the purpose of the study.	4.0	4.0	4.0
The CEFR familiarization activity (sorting CEFR descriptors in the online survey) was helpful.	3.7	3.8	3.8
The instructions and explanations provided by the facilitators were clear.	3.7	3.9	3.9
The explanation of the process for the judgment task helped me complete my assignment.	3.9	3.9	3.9
The explanation of how the recommended cut scores are computed was clear.	3.9	3.9	3.9
Feedback and discussion between judgement task rounds was helpful.	4.0	4.0	4.0

Note. Common European Framework of Reference for Languages (CEFR); 1 = *strongly disagree*; 2 = *disagree*; 3 = *agree*; 4 = *strongly agree*

Table 28 shows the survey results for the panelists' understanding of the distinguishing features of the five CEFR levels and the degree to which they were satisfied with the JQC descriptors they created collectively as a group. The average rating ranged from 3.7 to 3.9 across skills for both statements, with a majority of the panelists selecting Strongly Agree. However, the average rating for the listening and reading assessments was relatively lower than for speaking and writing. While none of the panelists selected Strongly Disagree or Disagree for these statements, the proportion of Agree was larger for the listening and reading assessments. Because listening and reading were the first assessments panelists worked on, they might have struggled to have a clear distinction of language features between different CEFR levels, or the discussion and debate might not have been fully optimized compared to the later meetings for speaking and writing. Panelists may have also been more familiar, as teachers, with the task of reviewing test-taker performances for speaking and writing, as

compared to the task of reviewing multiple-choice test items for listening and reading. Even so, panelists indicated a high level of agreement to the two statements across the four assessments.

Table 28. Panel's Feedback on Features of Common European Framework of Reference for Languages (CEFR) Levels and Just Qualified Candidate (JQC) Descriptors

Statement	Listening	Reading	Speaking	Writing
I understood the distinguishing features of the 5 CEFR levels, A1 to C1 for [skill].	3.7	3.7	3.9	3.9
I was satisfied with the descriptors of the Just Qualified Candidates (JQCs) produced and used during the meeting for [skill].	3.8	3.7	3.9	3.9

Note. 1 = *strongly disagree*; 2 = *disagree*; 3 = *agree*; 4 = *strongly agree*

Table 29 demonstrates how panelists characterized the standard setting meetings regarding four different aspects: efficiency, coordination, clarity, and satisfaction. The total average rating for all four aspects ranged from 4.7 to 5.0, indicating that the standard setting meeting was perceived extremely positively from a process perspective. While the ratings for listening and reading were lower compared to those for speaking and writing, which is the same pattern as we saw in earlier survey results, panelists were very satisfied with the meeting process (an average ranging from 4.9 to 5.0), providing strong evidence supporting procedural validity for this study.

Table 29. Panel's Rating on Aspects of Meeting Process

Characteristics of the meeting	Listening and Reading	Speaking	Writing
Efficiency (1 = *inefficient*; 5 = *efficient*)	4.7	5.0	4.9
Coordination (1 = *uncoordinated*; 5 = *coordinated*)	4.8	5.0	4.9
Clarity (1 = *confusing*; 5 = *understandable*)	4.8	5.0	4.9
Satisfaction (1 = *dissatisfying*; 5 = *satisfying*)	4.9	5.0	4.9

Note. Panelists used a five-point scale with 1 being the lowest and 5 being the highest rating.

The survey also asked which of the factors influenced their cut score judgments: JQC definitions, group discussions, summary statistics, and/or their own professional experience. As Table 30 shows, all of the factors were generally influential (none of the panelists indicated Not

Influential to any of the factors). Particularly, JQC definitions and their own professional experience in teaching English as well as developing assessment materials largely influenced their judgements (an average ranging from 2.8 to 2.9). While the JQC definition was a relatively new concept for some panelists, the fact that panelists relied on the JQC definition when making cut score judgments is a good indicator that the recommended cut scores were based on what JQC can do at a given CEFR level. In addition, it is not surprising that panelists' professional experience was most influential given that they have been in the field of English language teaching for an extended period of time (with an average of 14 years of teaching experience). Panelists also indicated that the between-rounds discussions and summary statistics were generally influential (an average ranging from 2.7 to 2.9).

Table 30. Factors Influencing Panelist Score Judgments

Factors	Listening	Reading	Speaking	Writing
The definition of the Just Qualified Candidate (JQC)	2.8	2.9	2.9	2.9
The between-round discussions	2.7	2.8	2.9	2.9
The summary statistics presented after each round of judgments	2.7	2.8	2.7	2.9
My own professional experience	2.9	2.9	2.9	2.9

Note. 1 = *not influential*; 2 = *influential*; 3 = *very influential*

Lastly, the survey asked the degree to which panelists felt comfortable with panel's recommended cut scores after three rounds of judgements and group discussions. This question was particularly important to investigate whether panelists felt their independent cut score judgments aligned well with the mean judgements of the entire panel and whether their individual opinions were reflected in the recommended cut scores. This is an important piece of validity evidence given that their recommendations were largely kept intact, with a few small poststudy adjustments made by the research team. Table 31 presents the raw number of panelists who indicated each comfort level with the panel's recommended cut scores using a 4-point Likert scale, and the average rating for each CEFR level, A2–C1.

Table 31. Panelists' Comfort Levels With the Recommended Cut Scores

	Listening						Reading				
Level	1	2	3	4	Average	Level	1	2	3	4	Average
A2	0	0	1	13	3.9	A2	0	0	1	13	3.9
B1	0	0	1	13	3.9	B1	0	0	1	13	3.9
B2	0	0	2	12	3.9	B2	0	0	4	10	3.7
C1	0	0	4	10	3.7	C1	0	0	2	12	3.9
	Speaking						Writing				
Level	1	2	3	4	Average	Level	1	2	3	4	Average
A2	0	0	1	13	3.9	A2	0	0	1	13	3.9
B1	0	0	0	14	4.0	B1	0	0	1	13	3.9
B2	0	0	1	13	3.9	B2	0	0	0	14	4.0
C1	0	0	0	14	4.0	C1	0	0	0	14	4.0

Note. 1 = *very uncomfortable*; 2 = *somewhat uncomfortable*; 3 = *somewhat comfortable*; 4 = *very comfortable*

As Table 31 shows, overall, panelists were very comfortable with the panel's recommended cut scores. None of the panelists indicated discomfort with any of the group's recommended cut scores across the assessments and the CEFR levels. The majority of panelists selected Very Comfortable with each cut score recommendation. Looking at the patterns for each assessment, the number of panelists who indicated they were only Somewhat Comfortable was generally higher at higher CEFR levels for listening and reading. For example, four panelists indicated that they were only Somewhat Comfortable with the C1 listening and B2 reading cut scores. For speaking and writing, almost all panelists were very comfortable with the recommended cut scores for all CEFR levels, which may reflect more panelist familiarity with making judgments regarding spoken or written test-taker performances, as distinct from multiple-choice test items. Other potential reasons might be that the TOEIC Link assessments better differentiate test-taker ability at lower levels for listening and reading, and/or that the boundary between B2 and C1 cut scores might have been less clear to the panelists compared to the lower levels of A2 and B1. In fact, the standard deviation for the Reading B2 cut score

(1.61) was larger than that for the other CEFR levels (A2: 1.22, B1: 1.44, and C1: 0.50), although the listening C1 cut score did not show a particularly high standard deviation compared to other CEFR levels. In summary, the Ready to Proceed and Meeting Evaluation surveys demonstrated that the panel had a positive experience overall with the meeting processes, the recommended cut scores, and the efficacy of the TOEIC Link assessments for the target population of English language learners, supporting the procedural validity evidence for this study.

The final aspect of the procedure for this study comprised the poststudy adjustments, as documented previously. While these adjustments required judgments on the part of the research team, all of these decisions were made very carefully, in consideration of all of the evidence, and with much discussion among the members of the research team. For each cut score, the default was to uphold the recommendation of the panel, due to the fact that the level of agreement and satisfaction among the panelists had been so high for this study. Adjustments were only made sparingly, in order to support congruence among the CEFR mappings of English language proficiency tests with the same or very similar task types, and in no case was any cut score adjusted by more than one point during the poststudy adjustment procedure.

Internal Validity

Internal validity in a standard setting study indicates the extent to which panelists are consistent in their judgments (Hambleton & Pitoniak, 2006). For example, it is desirable for both the *SD* and the *SEJ* of panelists' judgments at every stage to be small in proportion to the score scale of the test. Increasing panelist agreement across rounds of judgments, as shown in decreasing *SD* and *SEJ* values from earlier to later rounds, is another source of internal validity evidence. Finally, while it is important for panelists to feel free to adjust their judgments from one round to the next, it is a good sign if average panelist judgments remain relatively consistent, especially between the final two rounds.

The standard setting results shown in this report display all of these desirable qualities that provide internal validity evidence for this study and its results. Across all four skills and at all four CEFR cut scores, the *SEJ*s were all less than 0.5 in scale score units, indicating that another panel with similar background and training would likely make judgments that would

result in a cut score at the same scale-score point. The *SD*, *SEJ*, and range of panelist judgments also decreased consistently across rounds for all of these judgments, with just a few exceptions in which it remained similar (and still small). Finally, the mean panelist judgments resulted in the exact same scale-score cuts between Round 2 and Round 3 in most cases, with small exceptions in a few cases where careful analysis and discussion among panelists after Round 2 resulted in a small but intentional shift in one of the cut scores, reflecting an updated understanding among panelists that they gained during the standard setting process.

External Validity

External validity refers to the comparison of study results with relevant data external to the study itself, to ensure reasonableness of results within a wider framework (Hambleton & Pitoniak, 2006). The poststudy adjustment process, as documented earlier in this report, has already incorporated some external evidence into the cut scores recommended by the research team. We might expect these adjustments to enhance the external validity of the cut scores, but further research is still needed. Because the TOEIC Link assessments are new and without available operational data at the time of the writing of this report, sources of external validity evidence may be limited. Therefore, it is recommended that some of the following steps be taken to gather and evaluate external validity evidence once operational data for the TOEIC Link are available.

Evaluate Score Congruence Among Task Types for Speaking and Writing

In this study, panelists made judgments about how JQCs at each of the four CEFR levels would likely score on each task type for speaking and writing. Once operational data are available, it will be possible to evaluate whether test takers at a JQC score for one task type are also at the corresponding JQC score for the other task types. If the panelist judgments for one of the task types is out of sync with the others—for example, if test takers at the panelist-recommended JQC B1 task-level score for the speaking Read Aloud task type and Opinion task type have consistently higher (or consistently lower) mean scores than the panelist-recommended JQC B1 task-level score for the speaking task type of Listen and Repeat—then the task-level score judgments should be reexamined in light of the operational data.

Evaluate Scores of Test Takers Who Take Both the TOEIC Link and Another Related English Proficiency Test, Such as the TOEIC or the TOEIC Bridge

Because both the TOEIC and the TOEIC Bridge are already mapped to the CEFR, it would be possible to compare the reported CEFR levels between tests for test takers who took two or more related tests within a relatively short timeframe. One caveat here is that the CEFR mappings for all these tests reflect panelist judgment, which has an inherent subjective element. Therefore, it cannot be assumed that the results of earlier mapping studies are inherently more legitimate than those of later mapping studies. In addition, it is standard practice to evaluate and even update mappings over time, especially when test content or other specifications have been revised in some way. However, keeping these limitations in mind, it would still be useful to compare mappings in this way, and to consider possible implications for updating the CEFR mapping of the TOEIC Link and/or other related assessments in the future, to ensure coherence in score interpretation among the TOEIC family of tests.

Evaluate Relative Frequencies of Operational Scores Mapped to Each CEFR Level

This approach may not yield any firm recommendations regarding the score mapping, but it may be worthwhile in terms of evaluating the impact of the score mapping on test takers. Once sufficient operational data are available, the number of test takers at each score point and, therefore, within each CEFR level, on each assessment, can be examined. If some score points and CEFR levels have very few test takers while others have many, this pattern and its impact can be considered in light of anything else we may know about the test-taker population and/or typical ability distributions among test takers for English proficiency tests. If this process indicates that many TOEIC Link test takers have scores that are mapped to CEFR levels higher—or lower—than what we might expect for a typical English proficiency test taking population, then this might open an investigation into whether the cut scores from this study might be overly lenient or overly strict. Again, this approach should be used with caution, as new test taking populations may differ in their characteristics from other test taking populations in ways that are unexpected but real. Any differences among similar tests in relative frequencies of test takers mapped to various CEFR levels may be legitimate, but still worth investigating.

Limitations and Future Recommendations

As with the setting of cut scores on any new test, this study is limited by a lack of operational data specific to the TOEIC Link. While shared items and task types from other English language proficiency tests provide helpful reference information, including direct comparisons of test-taker performance on items shared in common between tests, it is recommended that the findings of this study be revisited in the future once operational data are available for test-taker performance on the TOEIC Link assessments.

For the listening and reading assessments, the collection of sufficient operational test-taker performance data will allow for the results of this study to be further examined in several ways. First, the stability and comparability of IRT parameters can be compared between the items as administered in the TOEIC Link assessments and the same items as originally administered in the TOEIC Listening and Reading test. Some statistical variation at the item level is expected due to population differences, so it will be important to evaluate the overall impact of the item-level statistical differences on reported scores. The impact of scoring stability on the classification of scores into CEFR levels, based on the results of this study, can then be evaluated in turn. Finally, as noted in the External Validity section, relative frequencies of operational scores mapped to each CEFR level can be evaluated in light of any available information about the expected distribution of the test-taker population among the CEFR levels.

For the speaking and writing assessments, operational data can provide both authentic test-taker responses and scoring patterns. These constructed-response assessments were mapped to CEFR levels in this study using the Expected Task Score method, in part because operational test-taker responses were not yet available and therefore could not be incorporated into the standard setting process. Once available, test-taker responses can be reviewed by content and scoring experts, and classified into CEFR levels based on their expert judgment. These qualitative classifications of test-taker responses can then be compared to the CEFR classifications indicated by the total score of each test taker, as determined by the cut scores resulting from this study. Any patterns of disagreement between the score-based versus the qualitative CEFR classifications can then be examined. In addition, scoring patterns can

themselves be examined for congruence among task types, and for consistency with other relevant information about the test-taking population, as described further in the External Validity section.

Conclusion

This report presents the standard setting procedures and outcomes to map scores on the newly developed TOEIC Link assessments (listening, reading, speaking, and writing) to CEFR levels. The development of the assessments is first described, including construct definitions, task types, and scoring models; construct congruence between the assessments and CEFR scales is then analyzed. The standard setting process itself is then explained, beginning with the methodology and the preparation of panelists and continuing through the defining of the JQC and the procedures for panelist judgments during the three days of panel meetings. The standard setting results are presented, and poststudy adjustments in light of available data from overlapping task types in other English proficiency tests are described and explained. Finally, the results are supported by procedural, internal, and limited external validity evidence, with recommendations for the future gathering of additional external validity evidence.

References

Cizek, G. J. (2012). An introduction to contemporary standard setting. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 3–14). Routledge. https://doi.org/10.4324/9780203848203

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Sage. https://doi.org/10.4135/9781412985918

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR): A manual.* https://rm.coe.int/1680667a2d

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

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

Davis, L., Norris, J., Papageorgiou, S., & Sasayama, S. (2023). Balancing construct coverage and efficiency: Test design, security, and validation considerations for a remotely-proctored online language test. In K. Sadeghi & D. Douglas (Eds.), *Fundamental considerations in technology mediated language assessment* (pp. 49–63). Routledge. https://doi.org/10.4324/9781003292395-5

ETS. (2022). *TOEIC® Listening and Reading test: Examinee handbook*. Author. https://www.ets.org/pdfs/toeic/toeic-listening-reading-test-examinee-handbook.pdf

Figueras, N., Kaftandjieva, F., & Takala, S. (2013). Relating a reading comprehension test to CEFR levels: A case of standard setting in practice with focus on judges and items. *The Canadian Modern Language Review, 69*(4), 359–385. http://dx.doi.org/10.3138/cmlr.1723.359

Geisinger, K. F., & McCormick, C. A. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29*(1), 38−44. http://dx.doi.org/10.1111/j.1745-3992.2009.00168.x

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). ACE and Praeger.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Lawrence Erlbaum Associates.

Mills, C. N. (1995). Establishing passing standards. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 219–252). Buros.

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. https://www.ets.org/Media/Research/pdf/RM-21-03.pdf

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The Modified Angoff, Extended Angoff, and Yes/No standard setting methods. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 181–199). Routledge.

Reckase, M. D. (2023). *The psychometrics of standard setting: Connecting policy and test scores.* CRC. https://doi.org/10.1201/9780429156410

Schmidgall, J. (2021). *Mapping the redesigned* TOEIC Bridge® *test scores to proficiency levels of the Common European Framework of Reference for Languages* (Research Memorandum No. RM-21-01). ETS. https://www.ets.org/Media/Research/pdf/RM-21-01.pdf

Schmidgall, J., Huo, Y., Cid, J., & Wei, Y. (2024). *Investigating fairness claims for a general-purposes assessment of English proficiency for the international workplace: Do full-time employees have an unfair advantage over full-time students?* (Research Report No. RR-24-06). ETS. https://doi.org/10.1002/ets2.12380

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping scores from the* TOEFL Junior® *Comprehensive test onto the Common European Framework of Reference (CEFR)* (Research Memorandum No. RM-15-13). ETS. https://www.ets.org/Media/Research/pdf/RM-15-13.pdf

Tannenbaum, R. J., & Cho, Y. (2014). Critical factors to consider in evaluating standard-setting studies to map language test scores to frameworks of language proficiency. *Language Assessment Quarterly, 7*(3), 233–249. https://doi.org/10.1080/15434303.2013.869815

Tannenbaum, R. J., & Katz, I. R. (2013). Standard setting. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 3. Testing and assessment in school psychology and education* (pp. 455–477). American Psychological Association. https://doi.org/10.1037/14049-022

Tannenbaum, R. J., & Wylie, E. C. (2008). *Linking English language test scores onto the Common European Framework of Reference: An application of standard setting methodology* (TOEFL Research Report No. TOEFL ibt-06). ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Notes

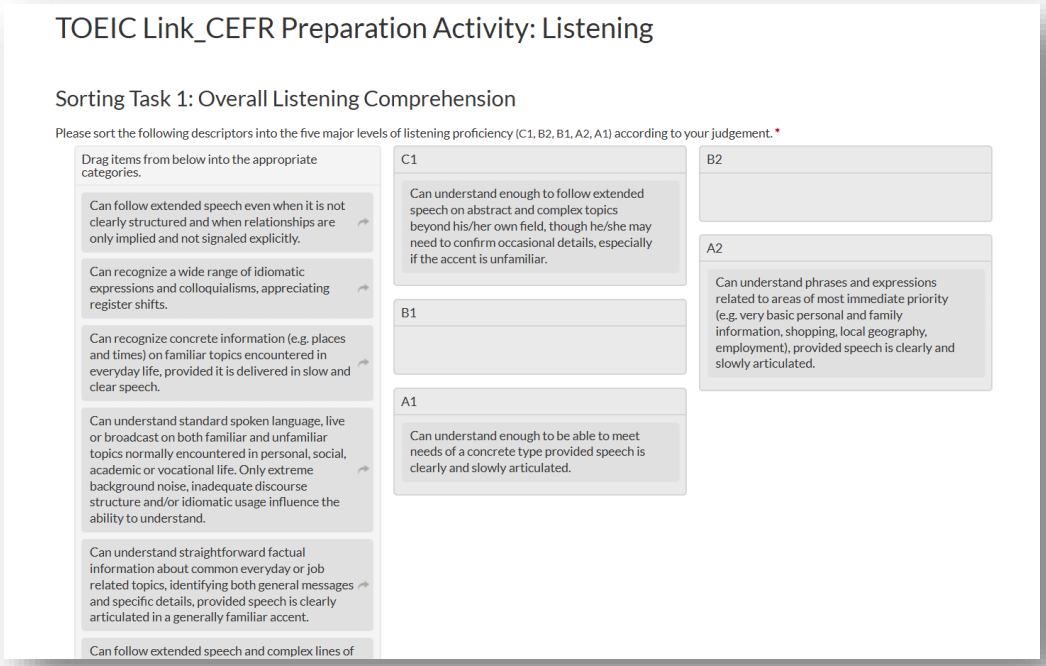
[1] Rounding up to the nearest scale score may slightly reduce false-positive classifications errors at the expense of slightly increasing false-negative classification errors. In other words, this approach slightly decreases the possibility that some test takers classified at a given CEFR level do not actually have the proficiency needed to be classified at that level, but slightly increases the possibility that some test takers may have their CEFR-level classification underestimated. Ultimately, score users should consider their decision-making needs and priorities (e.g., the seriousness of false positive versus false negative classification errors) when utilizing a cut score for classification and decision-making purposes.

List of Appendices

Appendix A. The Standard Setting Panelists, Affiliation, Institution Type, and Country

Panelist name	Affiliation	Institution type	Country
John Arlon	Junia Hautes Etudes d'Ingénieur	Business school	France
Nicholas Josephs	University Le Havre FST	University	France
Teresa Hancock	Wall Street English	English language training institution	Italy
Steven Menditto	Wall Street English, Inlingua	English language training institution	Italy
Taner Yapar	TOBB University of Economics and Technology	University	Türkiye
Seyma Dogru	TOBB University of Economics and Technology	University	Türkiye
Deniz Akkuş	Şişli Terakki High School	High school	Türkiye
Elif Kantarcıoğlu	Bilkent University	University	Türkiye
María Esther Pavón Solana	Review Quality Benemerita Universidad Autonoma de Puebla	University	Mexico
Guadalupe Valades	Universidad Marista Valladolid	University	Mexico
Nimbe Gallegos	Campus Universitario Siglo 21	University	Mexico
Sally Hatfield	The American School for Women at Water for Ishmael	Language school for immigrant/refugee women	USA
Ahmet Dursun	University of Chicago	University	USA
Paula Price	ETS	Organization	USA

Appendix B. Sample Screenshots of First Preparation Activity for Panelists



TOEIC Link_CEFR Preparation Activity: Listening

Sorting Task 1: Overall Listening Comprehension

Please sort the following descriptors into the five major levels of listening proficiency (C1, B2, B1, A2, A1) according to your judgement. *

Drag items from below into the appropriate categories.

- Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signaled explicitly.
- Can recognize a wide range of idiomatic expressions and colloquialisms, appreciating register shifts.
- Can recognize concrete information (e.g. places and times) on familiar topics encountered in everyday life, provided it is delivered in slow and clear speech.
- Can understand standard spoken language, live or broadcast on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influence the ability to understand.
- Can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent.
- Can follow extended speech and complex lines of

C1

Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar.

B1

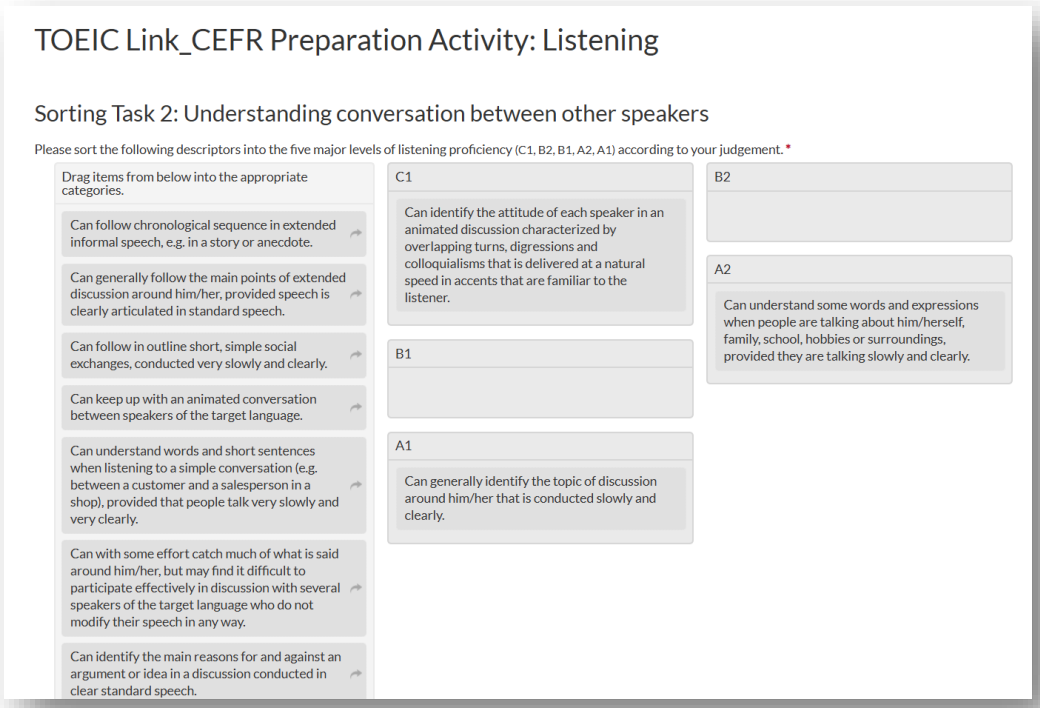
A1

Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated.

B2

A2

Can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment), provided speech is clearly and slowly articulated.



TOEIC Link_CEFR Preparation Activity: Listening

Sorting Task 2: Understanding conversation between other speakers

Please sort the following descriptors into the five major levels of listening proficiency (C1, B2, B1, A2, A1) according to your judgement. *

Drag items from below into the appropriate categories.

- Can follow chronological sequence in extended informal speech, e.g. in a story or anecdote.
- Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard speech.
- Can follow in outline short, simple social exchanges, conducted very slowly and clearly.
- Can keep up with an animated conversation between speakers of the target language.
- Can understand words and short sentences when listening to a simple conversation (e.g. between a customer and a salesperson in a shop), provided that people talk very slowly and very clearly.
- Can with some effort catch much of what is said around him/her, but may find it difficult to participate effectively in discussion with several speakers of the target language who do not modify their speech in any way.
- Can identify the main reasons for and against an argument or idea in a discussion conducted in clear standard speech.

C1

Can identify the attitude of each speaker in an animated discussion characterized by overlapping turns, digressions and colloquialisms that is delivered at a natural speed in accents that are familiar to the listener.

B1

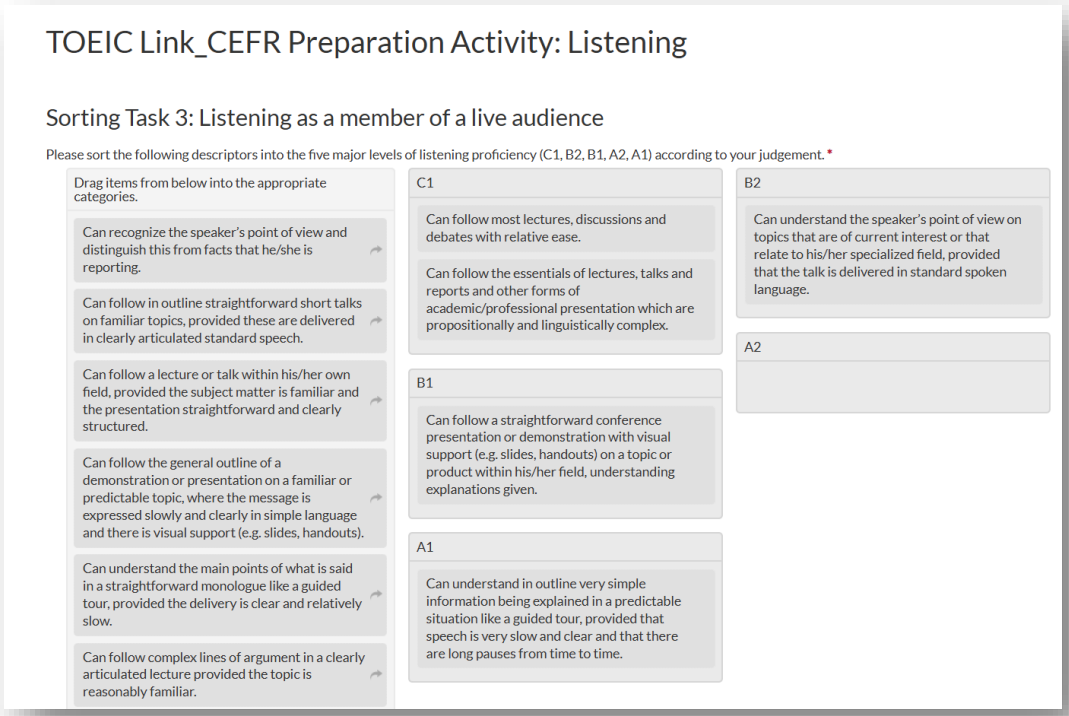
A1

Can generally identify the topic of discussion around him/her that is conducted slowly and clearly.

B2

A2

Can understand some words and expressions when people are talking about him/herself, family, school, hobbies or surroundings, provided they are talking slowly and clearly.



TOEIC Link_CEFR Preparation Activity: Listening

Sorting Task 3: Listening as a member of a live audience

Please sort the following descriptors into the five major levels of listening proficiency (C1, B2, B1, A2, A1) according to your judgement. *

Drag items from below into the appropriate categories.

- Can recognize the speaker's point of view and distinguish this from facts that he/she is reporting.
- Can follow in outline straightforward short talks on familiar topics, provided these are delivered in clearly articulated standard speech.
- Can follow a lecture or talk within his/her own field, provided the subject matter is familiar and the presentation straightforward and clearly structured.
- Can follow the general outline of a demonstration or presentation on a familiar or predictable topic, where the message is expressed slowly and clearly in simple language and there is visual support (e.g. slides, handouts).
- Can understand the main points of what is said in a straightforward monologue like a guided tour, provided the delivery is clear and relatively slow.
- Can follow complex lines of argument in a clearly articulated lecture provided the topic is reasonably familiar.

C1

- Can follow most lectures, discussions and debates with relative ease.
- Can follow the essentials of lectures, talks and reports and other forms of academic/professional presentation which are propositionally and linguistically complex.

B2

- Can understand the speaker's point of view on topics that are of current interest or that relate to his/her specialized field, provided that the talk is delivered in standard spoken language.

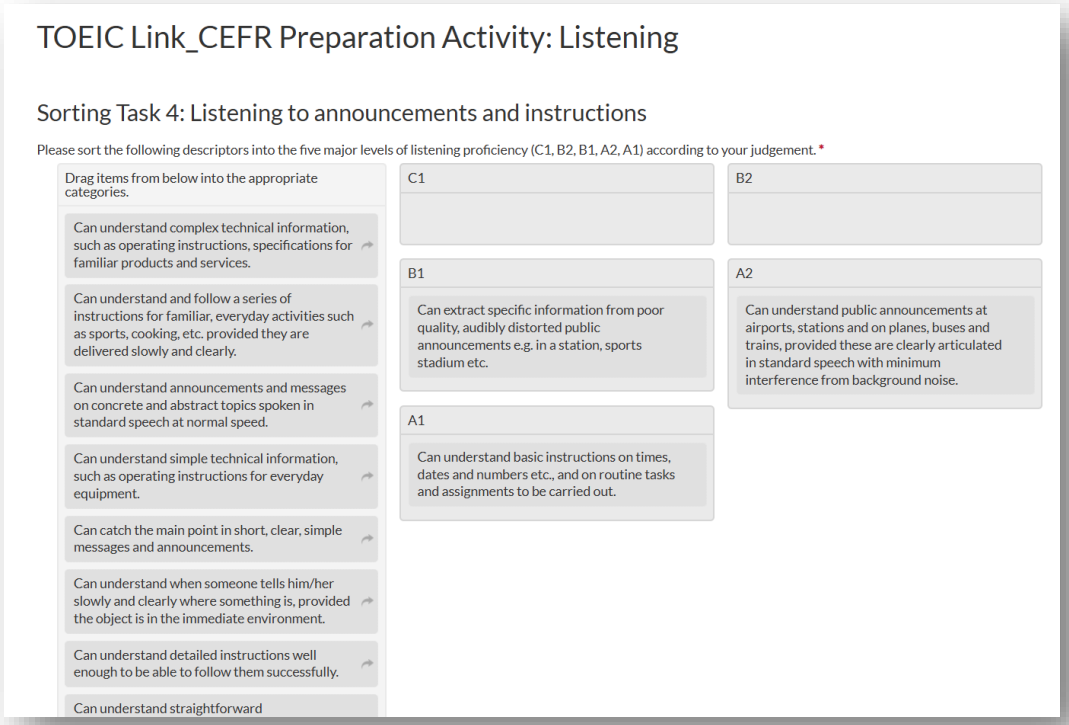
B1

- Can follow a straightforward conference presentation or demonstration with visual support (e.g. slides, handouts) on a topic or product within his/her field, understanding explanations given.

A2

A1

- Can understand in outline very simple information being explained in a predictable situation like a guided tour, provided that speech is very slow and clear and that there are long pauses from time to time.



TOEIC Link_CEFR Preparation Activity: Listening

Sorting Task 4: Listening to announcements and instructions

Please sort the following descriptors into the five major levels of listening proficiency (C1, B2, B1, A2, A1) according to your judgement. *

Drag items from below into the appropriate categories.

- Can understand complex technical information, such as operating instructions, specifications for familiar products and services.
- Can understand and follow a series of instructions for familiar, everyday activities such as sports, cooking, etc. provided they are delivered slowly and clearly.
- Can understand announcements and messages on concrete and abstract topics spoken in standard speech at normal speed.
- Can understand simple technical information, such as operating instructions for everyday equipment.
- Can catch the main point in short, clear, simple messages and announcements.
- Can understand when someone tells him/her slowly and clearly where something is, provided the object is in the immediate environment.
- Can understand detailed instructions well enough to be able to follow them successfully.
- Can understand straightforward

C1

B2

B1

- Can extract specific information from poor quality, audibly distorted public announcements e.g. in a station, sports stadium etc.

A2

- Can understand public announcements at airports, stations and on planes, buses and trains, provided these are clearly articulated in standard speech with minimum interference from background noise.

A1

- Can understand basic instructions on times, dates and numbers etc., and on routine tasks and assignments to be carried out.

Appendix C. Panelists' Just Qualified Candidate (JQC) Descriptors

Listening

Listening skills of just-qualified A2

- Can generally understand slow, clear speech on concrete, everyday topics of personal relevance.
- Can easily understand high frequency words and phrases (vocabulary) on concrete, everyday topics.
- Can understand the main topic of a conversation or talk, although accommodations (such as examples, repetition, or clarification) are often required in everyday topics.
- Can pick out some important information from short, straightforward commercials, recordings, presentations, conversations between other speakers, etc., especially with the support of visuals.
- Can sometimes understand simple, multi-step directions that are outside of the immediate environment.

Listening skills of just-qualified B1

- Can usually understand clear standard speech with relatively slow delivery.
- Can generally understand familiar topics in everyday and/or work-related contexts.
- Can understand main ideas in extended speech with significant support (e.g. repetition, paraphrasing, visual/graphic).
- Can understand some important details when explicitly stated (e.g., instructions, technical information, agreement/disagreement).
- Can sometimes deduce the meaning of unknown words within familiar contexts.
- Can sometimes make basic inferences using visual or contextual clues (situation, vocabulary).

Listening skills of just-qualified B2

- Can understand explicit factual information about familiar topics in extended speech, identifying both main ideas and specific details.
- Can sometimes follow complex lines of argument (including point of view) on familiar topics when explicit markers are used.
- Can sometimes draw inferences and make predictions from extended speech on familiar topics.

- Can sometimes understand main points on unfamiliar or abstract topics.
- Can understand oral texts of some semantic and syntactic complexity on familiar topics.

Listening skills of just-qualified C1

- Can generally follow extended, complex or abstract speech on a wide variety of familiar and unfamiliar topics even when that speech is not clearly structured.
- Can often follow extended speech even when relationships between speakers are not explicit.
- Can usually understand inference, shifts in register, and implied meaning.
- Can identify the attitudes of speakers in a wide variety of contexts.
- Can usually understand natural speech (unmodified, unaccommodated) even when it contains idiomatic expressions and colloquialism.

Reading

Reading skills of just-qualified A2

- Can comprehend short texts (beyond a single sentence) with a decreasing dependence on visual cues.
- Can find information in predictable texts, such as calendars, signs, schedules, instructions (patterns, formats).
- Can understand simple texts based on everyday topics using familiar or high frequency vocabulary.
- Can understand simple personal communications (e-mails, text messages).
- Can recognize basic discourse markers (also, but, so, because).
- Can use basic grammatical knowledge (tenses, agreement, plurals, etc.) to aid comprehension.

Reading skills of just-qualified B1

- Can understand familiar everyday and/or workplace-related texts of more than one paragraph as long as it contains straightforward sentences.
- Can understand contexts that are unfamiliar as long as the information is direct and explicit (if visual support is provided).
- Can generally make basic inferences from stated information.
- Understanding vocabulary from context, including derivational morphemes.
- Can identify salient, explicitly stated details within a text.

- Can understand descriptions of feelings, events, and places within straightforward, simply written articles and guides.
- Can generally follow the plot of everyday and workplace-related narrative texts with a linear, clear storyline and identify a sequence of events.
- Can increasingly understand discourse markers (suddenly, therefore, however, conjunctions) to connect ideas.

Reading skills of just-qualified B2

- Can usually understand an author's general point of view.
- Can understand a broad range of reading vocabulary, including some colloquial language (have difficulty with some idioms).
- Can use a wide variety of discourse markers and text structures (e.g., cause and effect, compare/contrast) to achieve understanding.
- Can scan well-organized, coherent, extended texts or multiple sources to identify the relevance of information given enough time.
- Can adapt reading strategies to different text styles and formats.

Reading skills of just-qualified C1

- Can fully understand most extended, abstract/complex texts, including those outside their area of specialty or interest.
- Can often infer attitudes, emotions, intentions, and opinions implied in a text and understand shifts in register.
- Can understand most texts of various genres in multiple sources/texts, re-reading as needed.
- Can usually understand idiomatic expressions, colloquial language, and technical vocabulary across a range of contexts.
- Can understand a broad range of grammatical structures to derive meaning from most complex texts.
- Can analyze and synthesize from multiple texts to some extent.

Speaking

Speaking skills of just-qualified A2

- Some grammatical, lexical, pronunciation errors will impede parts of the message.

- Can speak with limited fluency on familiar topics with frequent pauses/hesitations that may require a sympathetic interlocutor.
- Can provide limited description in familiar contexts and talk about immediate needs.
- Can present an opinion with very limited supporting detail in familiar contexts.
- May use a range of simple cohesive devices (e.g., because, but, so).
- Can use simple, short, rehearsed phrases on personal/familiar tasks. Speech characterized by simple, memorized, fixed/formulaic phrases.
- Attempts to use correct intonation and stress but may not sound natural. Accent is strongly influenced by the other language(s) they speak.
- Can sometimes understand and accurately repeat parts of the shorter utterances.
- Attempts to repeat some longer utterances but will generally be unsuccessful.

Speaking skills of just-qualified B1

- Grammatical, lexical, pronunciation errors may impede parts of the message, but the overall message is clear.
- Can speak with some degree of fluency with frequent pauses/hesitations that may require the listener's effort to understand.
- Can provide straightforward descriptions in familiar contexts.
- Can briefly provide arguments, reasons, and support of opinion in familiar contexts.
- Can use a limited range of more complex cohesive devices with some accuracy (e.g., however, in addition).
- Can use a wide range of simple language with some attempts at using more complex forms.
- Can use some limited features of appropriate intonation, sometimes places stress correctly, and recognizes and articulates most familiar words clearly; accent is usually influenced by the other language(s) they speak.
- Can generally understand and accurately repeat some shorter utterances. Can sometimes understand and repeat longer utterances, although responses may be incomplete.

Speaking skills of just-qualified B2

- Can speak with some degree of fluency, spontaneity with highly proficient English speakers with minimal strain on either party.
- Grammatical and lexical errors usually do not impede message; however, a few words may be ambiguous because of pronunciation.

- Can give clear and detailed descriptions and presentations on a wide range of subjects related to field of interest.
- Uses some idiomatic language and appropriate collocations and some complex forms in some contexts.
- Speech may contain occasional pauses/hesitations when searching for patterns and expressions.
- Employs a limited number of cohesive devices. There is some jumpiness in longer contributions.
- Can sustain point of views by providing limited relevant explanations and arguments, discussions.
- Can use some features of appropriate intonation, generally places stress correctly and articulates individual sounds clearly; accent only has some influence by the other language(s) they speak, but has little or no effect on intelligibility.
- Can usually understand and accurately repeat shorter utterances. Can generally understand and attempt to successfully repeat longer utterances with varying degree of success.
- Has sufficient vocabulary to speak about familiar contexts, with a wider range in field of interest.

Speaking skills of just-qualified C1

- Can produce clear and detailed descriptions and extended speech on complex topics regardless of contexts with some success.
- Can develop an appropriate argument systematically in well-structured language, with supporting details and/or examples and concluding appropriately.
- Usually uses effective, precise, flexible language at length without difficulty and can self-correct and adjust register when necessary.
- Can use most features of appropriate intonation, generally places stress correctly and articulate individual sounds clearly; accent does not affect intelligibility.
- Has a substantial lexical repertoire, including idiomatic language and appropriate collocations, although some errors may persist.
- Uses complex grammatical structures, although limited errors may persist.

Writing

Writing skills of just-qualified A2

- Can generally produce a series of straightforward sentences on familiar and everyday subjects with some degree of intelligibility.
- Can briefly give straightforward opinions about topics of personal interest.
- Can communicate using basic vocabulary on familiar and everyday subjects, although writing will show interference from other language(s).
- Can usually compose sentences with a range of simple cohesive devices (and, but, so).
- Can use basic grammatical forms, although writing will show interference from other language(s).

Writing skills of just-qualified B1

- Can sometimes produce straightforward, generally intelligible texts on variety of familiar subjects with some degree of success.
- Can briefly give reasons and opinions with very limited support.
- Can communicate using a limited range of vocabulary on familiar subjects and restricted use of appropriate register and conventions, although writing may show interference from other language(s).
- Can usually compose linear sequential text with some use of basic cohesive devices.
- Can use a wide range of simple language with some attempts at using more complex forms, although writing may show interference from other language(s).
- Can sometimes write compound sentences using appropriate grammar and correct word order. Simple sentences are often correct, although errors may impede comprehension at times.
- Can sometimes write simple sentences using appropriate grammar and correct word order.

Writing skills of just-qualified B2

- Can usually produce clear texts with some details on a variety of familiar subjects.
- Can explain their view with limited support, including describing advantages and disadvantages, causes and effects, and/or similarities and differences.
- Can usually write compound and complex sentences using appropriate grammar and correct word order.

- Can usually communicate clearly and effectively, including using a good range of vocabulary on familiar subjects and limited use of appropriate register and conventions.
- Can use common colloquialisms and idiomatic expressions to some extent.
- Can usually synthesize information and views from a number of sources with varying degree of success.
- Can usually compose texts that follow standard structural patterns with limited use of cohesive devices.

Writing skills of just-qualified C1

- Can usually produce complex, well-structured and well-developed texts on a variety of both abstract and concrete topics.
- Can elaborate their view to some degree, including describing advantages and disadvantages, causes and effects, and/or similarities and differences.
- Can write mostly accurately in a natural style using appropriate grammar and correct word order.
- Can usually communicate clearly and effectively using language flexibly in an appropriate tone and style adjusting to shifts in register.
- Can usually use a range of common colloquialisms and idiomatic expressions.
- Can produce texts by synthesizing information and views from a number of sources to draw an informed conclusion with some degree of success.
- Can usually compose texts efficiently with a variety of cohesive devices and organizational patterns.
- Minor errors do not impede general comprehension.

Appendix D. Scoring Rubrics for Speaking and Writing

Speaking Rubrics

Scoring Rubric: Read Aloud

Score **Response Description**

4 **The passage is read with ease; speech is fluid and intelligible; units of meaning are clearly marked.**

A typical response exhibits all of the following.

- Reading is fluid, with little hesitation or self-correction, and the rate of speech is appropriate.
- Intonation and pausing are used to group words in meaningful phrases.
- Speech is highly intelligible; minor mispronunciations or other-language influences may be present that do not impair intelligibility.
- The text is fully and accurately reproduced.

3 **The passage is read with little difficulty; it may require minor listener effort.**

A typical response exhibits most of the following.

- Pacing is mostly even but with minor hesitation/choppiness.
- Intonation may be monotone at times, but words are mostly grouped in meaningful phrases.
- Intelligibility is generally sustained, although there may be occasional minor lapses.
- Occasional mispronunciations may require minor listener effort.
- The text is fully reproduced, with no more than minor changes.

2 **The passage is read with some difficulty; the content is partially clear.**

A typical response exhibits one or more of the following.

- Pacing is uneven or slow; frequent or inappropriate hesitations cause noticeable choppiness; sentence level meanings may be hard to follow.
- Intelligibility is repeatedly affected by inaccuracies in pronunciation or intonation.
- Intonation is monotone or inappropriate for the meaning; there may be partially effective grouping of words into meaningful phrases.
- The text is mostly reproduced; variations (substitutions, omissions) may be present.

1 **The passage is read with noticeable difficulty; the content may be mostly unintelligible.**

A typical response exhibits one or more of the following.

- The speaking rate may be very slow, or the text is mostly read in short chunks, or multiple or extended pauses result in speech that is largely staccato.
- The response is mostly unintelligible; only isolated words or phrases from the text are understandable.
- The text is substantially incomplete, or the response consists of isolated phrases; variations (substitutions, omissions) may be common.

0 **No response OR the response is entirely unintelligible OR there is no English in the response OR the test taker does not read the text provided.**

Scoring Rubric: Listen and Repeat

Score **Response Description**

5 **The response exactly repeats the prompt.**

A typical response exhibits the following.

- The response is fully intelligible and is an exact repetition of the prompt.

4 **The response captures the meaning expressed in the prompt, but it is not an exact repetition.**

A typical response exhibits the following.

- Minor changes in words or grammar are present that do not substantially change the meaning of the prompt. For example:
  - one or two function words may be missing or changed,
  - a content word may be missing (in longer stimuli) or replaced with a related word,
  - markers of tense/aspect/number may be missing or incorrect, or
  - two words may be transposed.
- One or two content words may be ambiguous because of imprecise pronunciation. The speaker may self-correct but successfully completes the response.

3 **The response is essentially full, but it does not accurately capture the original meaning.**

A typical response exhibits the following.

- The response contains a majority of the content words or ideas in the prompt.
  - Multiple function words may be changed or missing; one or more content words may be missing or substantively changed.
- The response is a full sentence.
- In some cases, intelligibility issues cause occasional difficulty in understanding meaning. The speaker may struggle over a word or phrase or run words together, reducing intelligibility.

2 **The response is missing a significant part of the prompt and/or is highly inaccurate.**

A typical response exhibits one or more of the following.

- A large portion of the prompt is missing, and important original meaning is left out.
  - The speaker may repeat the first part of the sentence. Then the speaker may stop or fill with inaccurate content and/or include the last few words.
- The response is not a self-standing sentence; meaning is fragmentary.
- Intelligibility is low; the response would be difficult to understand for a listener unfamiliar with the prompt.

1 **The response captures very little of the prompt or is largely unintelligible.**

A typical response exhibits one or more of the following.

- A minimal response of a few words is made; most of the prompt is missing.
- The response is recognizable as an attempt to repeat the prompt, but it is mostly unintelligible.

0 **No response OR the response is entirely unintelligible OR there is no English in the response OR the content is entirely unconnected to the prompt (or consists only of phrases such as "I don't know").**

Scoring Rubric: Express an Opinion

A successful response (Score 4 or 5) indicates that the test taker can create connected, sustained discourse appropriate to the typical workplace.

Score	Score Description
5	**The response fulfills the demands of the task and is readily intelligible, sustained, and coherent. It is characterized by ALL of the following:** • The response is generally well developed; relationships between ideas are clear. • The speech is clear with a generally well-paced flow. It may include minor lapses or minor difficulties with pronunciation or intonation patterns that do not affect overall intelligibility. • The response demonstrates automatic and effective use of grammar, with good control of basic and complex structures. Some minor errors may be noticeable, but they do not obscure meaning. • The use of vocabulary is effective, with allowance for occasional minor inaccuracy.
4	**The response addresses the task and is generally intelligible, sustained, and coherent.** • The response is somewhat developed; relationships between ideas are mostly clear, with occasional lapses. • Minor difficulties with pronunciation, intonation, or pacing are noticeable and may require listener effort at times, although overall intelligibility is not significantly affected. • The response demonstrates fairly automatic and effective use of grammar but may be somewhat limited in the range of structures used. • The use of vocabulary is fairly effective. Some vocabulary may be inaccurate or imprecise.
3	**The response addresses the task and is basically intelligible, but development is limited.** • The response provides some development. However, it provides little elaboration, repeats itself with no new information, is vague, or is unclear. Relationships between ideas may be unclear at times. • The speech is basically intelligible, though listener effort may be needed because of unclear articulation, awkward intonation, or choppy rhythm/pace; meaning may be obscured in places. • The response demonstrates limited control of grammar; for the most part, only basic sentence structures are used successfully. • The use of vocabulary is limited.
2	**The response attempts to address the task but has significant limitations. Parts of the response may not be sustained or may be unintelligible or incoherent.** • The response expresses only basic ideas and is vague or repetitious. Relationships between ideas may be unclear. • Consistent difficulties with pronunciation, stress, and intonation cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; there may be long pauses and frequent hesitations. • Control of grammar severely limits the expression of ideas and clarity of connections among ideas. • The use of vocabulary is severely limited or highly repetitious.
1	**The response is limited to reading the prompt or the directions aloud OR the response consists of isolated words or phrases, or mixtures of the first language and English.**
0	**No response OR no English in the response.**

Writing Rubrics

Scoring Rubric: Write a Brief Email

Score **Response Description**

5 **A fully successful response**

The response is well elaborated and shows consistent facility in language use.
A typical response shows all of the following.

- The e-mail accomplishes the task and is fully elaborated with relevant supporting details.
- A range of grammar and vocabulary is used effectively (e.g., to create a degree of fluidity, precision, idiomaticity, or expressiveness).
- Language use is accurate, although minor errors or nonidiomatic uses may occur. However, the intended meaning is fully clear and cohesive throughout.
- Appropriate social conventions (e.g., politeness, organization of information, and formulation of actions such as requests, etc.) are consistently used.

4 **A generally successful response**

The response is elaborated, although minor inconsistencies in language use are present.
A typical response shows most of the following.

- The e-mail accomplishes the task and is supported with suitable elaboration.
- A range of grammar and vocabulary is used appropriately, but the range or effectiveness may not be sustained throughout the response.
- Inaccuracies in language use may be noticeable and may cause minor ambiguities in sentence-level meanings, but the overall response is clear.
- Mostly appropriate social conventions are used, but usage may not be consistent.

3 **A partially successful response**

The response shows some elaboration and/or facility in language use, but it is limited in one or more aspects.
A typical response partially accomplishes the task, but it may show either of the following patterns.

- The response shows a reasonable degree of elaboration, with limitations elsewhere.
  - A somewhat limited range of grammar and vocabulary is used; attempts at more complex structures or vocabulary are not entirely successful.
  - Limitations in range, accuracy, or mechanics may affect the readability or clarity of part or all of the response.
- The response shows limited elaboration, which reduces the effectiveness of the message.
  - An adequate range of grammar and vocabulary is used; limitations in language use do not greatly affect the overall clarity of the message.

2 **A mostly unsuccessful response**

The response is limited in content and in facility of language use.
A typical response shows some or all of the following.

- Very limited elaboration is present, which significantly affects the effectiveness of the e-mail.
- A limited range of grammar and vocabulary is used, although some attempt is made to produce sentence-level language beyond simple clauses or basic vocabulary.
- Errors in language use or mechanics may cause the intended meaning to be unclear.

1 **An unsuccessful response**

The content of the response is very limited, with little evidence of ability to produce extended text.
A typical response shows some or all of the following.

- Little or no elaboration is provided.
- The e-mail consists of telegraphic language (i.e., short and/or disconnected phrases and simple sentences).
- Serious errors in language use may be present.

0 **The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt, or consists of arbitrary keystrokes.**

Scoring Rubric: Write an Extended E-mail

Score **Response Description**

5 **A fully successful response**

The response is effective, is clearly expressed, and shows consistent facility in the use of language.
A typical response displays the following.

- Elaboration that effectively supports the communicative purpose
- Effective syntactic variety and precise, idiomatic word choice
- Consistent use of appropriate social conventions (e.g., politeness, register, organization of information, and formulation of actions such as requests, refusals, criticisms, etc.)
- Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there*/*their*)

4 **A generally successful response**

The response is mostly effective and easily understood. Language facility is adequate to the task.
A typical response displays the following.

- Adequate elaboration to support the communicative purpose
- Syntactic variety and appropriate word choice
- Mostly appropriate social conventions
- Few lexical or grammatical errors

3 **A partially successful response**

The response generally accomplishes the task. Limitations in language facility may prevent parts of the message from being fully clear and effective.
A typical response displays the following.

- Elaboration that partially supports the communicative purpose
- A moderate range of syntax and vocabulary
- Some noticeable errors in structure, word forms, use of idiomatic language, and/or social conventions

2 **A mostly unsuccessful response**

The response reflects an attempt to address the task, but it is mostly ineffective. The message may be limited or difficult to interpret.
A typical response may display the following.

- Limited or irrelevant elaboration
- Some connected sentence-level language, with a limited range of syntax and vocabulary
- An accumulation of errors in sentence structure and/or language use

1 **An unsuccessful response**

The response reflects an ineffective attempt to address the task. The message may be limited to the point of being unintelligible.
A typical response may display the following.

- Very little elaboration, if any
- Telegraphic language (i.e., short and/or disconnected phrases and sentences) with a very limited range of vocabulary
- Serious and frequent errors in the use of language
- Minimal original language; any coherent language is mostly borrowed from the stimulus.

0 **The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt, or consists of arbitrary keystrokes.**

Scoring Rubric: Write for an Online Discussion

Score **Response Description**

5 **A fully successful response**

The response is a relevant and very clearly expressed contribution to the online discussion, and it demonstrates consistent facility in the use of language.

A typical response displays the following.

- Relevant and well-elaborated explanations, exemplifications, and/or details
- Effective use of a variety of syntactic structures and precise, idiomatic word choice
- Almost no lexical or grammatical errors other than those expected from a competent writer writing under timed conditions (e.g., common typos or common misspellings or substitutions like *there*/*their*)

4 **A generally successful response**

The response is a relevant contribution to the online discussion, and facility in the use of language allows the writer's ideas to be easily understood.

A typical response displays the following.

- Relevant and adequately elaborated explanations, exemplifications, and/or details
- A variety of syntactic structures and appropriate word choice
- Few lexical or grammatical errors

3 **A partially successful response**

The response is a mostly relevant and mostly understandable contribution to the online discussion, and there is some facility in the use of language.

A typical response displays the following.

- Elaboration in which part of an explanation, example, or detail may be missing, unclear, or irrelevant
- Some variety in syntactic structures and a range of vocabulary
- Some noticeable lexical and grammatical errors in sentence structure, word form, or use of idiomatic language

2 **A mostly unsuccessful response**

The response reflects an attempt to contribute to the online discussion, but limitations in the use of language may make ideas hard to follow.

A typical response displays the following.

- Ideas that may be poorly elaborated or only partially relevant
- A limited range of syntactic structures and vocabulary
- An accumulation of errors in sentence structure, word forms, or use

1 **An unsuccessful response**

The response reflects an ineffective attempt to contribute to the online discussion, and limitations in the use of language may prevent the expression of ideas.

A typical response may display the following.

- Words and phrases that indicate an attempt to address the task but with few or no coherent ideas
- Severely limited range of syntactic structures and vocabulary
- Serious and frequent errors in the use of language
- Minimal original language; any coherent language is mostly borrowed from the stimulus

0 **The response is blank, rejects the topic, is not in English, is entirely copied from the prompt, is entirely unconnected to the prompt, or consists of arbitrary keystrokes.**

www.ets.org