Research Memorandum
ETS RM–21-01

Mapping the Redesigned *TOEIC Bridge*® Test Scores to Proficiency Levels of the Common European Framework of Reference for Languages

Jonathan Schmidgall

March 2021

ETS Research Memorandum Series

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

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

Mapping the Redesigned *TOEIC Bridge*® Test Scores to Proficiency Levels of the Common European Framework of Reference for Languages

Jonathan Schmidgall
ETS, Princeton, New Jersey, United States

March 2021

Corresponding author: J. Schmidgall, E-mail: jschmidgall@ets.org

Suggested citation: 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.

Action Editor: John Norris

Reviewers: Kathryn Hille and Spiros Papageorgiou

Abstract

The redesigned *TOEIC Bridge*® tests are designed to measure the reading, listening, speaking, and writing proficiency of beginning to low-intermediate English learners in the context of everyday adult life. This report describes the comprehensive and multifaceted process used to enhance the meaningfulness of TOEIC Bridge test score interpretations by mapping them to Levels Pre-A1, A1, A2, and B1 of the Common European Framework of Reference for Languages (CEFR). The process adhered to best practices in educational measurement for mapping test scores to standards and closely followed the procedure recommended by the Council of Europe's manual for relating examinations to the CEFR, elaborating claims and evidence as they pertain to familiarization, specification, standard setting, and validation.

Keywords: English learners, *TOEIC Bridge*® tests, Common European Framework of Reference for Languages, test scores, standard setting, validation, score interpretations, reading, speaking, writing, listening

Table of Contents

The meaning of test scores needs to be clearly established before scores can be used effectively. One of the most important responsibilities of language test developers is to help ensure that score interpretations are meaningful to stakeholders, including test takers and score users (Bachman & Palmer, 2010). The meaning of test scores can be established and communicated in a variety of ways. Fundamentally, the knowledge, skills, or abilities assessed by the test need to be clearly stated in the construct definition, which provides a basis for test design and validation. For the redesigned *TOEIC Bridge*® tests, this information was communicated in the framework paper for the test (Schmidgall et al., 2019).

Another important way to communicate the meaning of test scores to stakeholders is by relating them to external language proficiency standards or descriptors (Tannenbaum & Cho, 2014). For many stakeholders, language proficiency standards provide a brief and accessible way to understand levels of language proficiency across broad, general levels such as beginner, intermediate, and advanced (Hudson, 2013). When language proficiency standards are used to directly inform decision-making, mapping test scores to these standards can also ensure score interpretations are more relevant to score users. For score users in this context, language proficiency standards are intertwined with policy. For example, admission to a program of study may require a minimum level of language proficiency (e.g., low intermediate). Language training courses may be offered at varying levels of proficiency, and placement into training may depend on an individual's current level of language proficiency. As a result, important decisions or evaluations may be based on whether an individual or group of learners has achieved a level of language proficiency defined by a specific set of language proficiency standards or descriptors.

One set of widely used language proficiency levels and descriptors is presented in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The CEFR was introduced in 2001 and expanded with a companion volume in 2018 (Council of Europe, 2018) in order to promote the development of language learning curriculum and provide an orientation for language teaching and learning. Through its description of language proficiency and a set of common reference levels, the CEFR also aims to promote cooperation

among various stakeholders (e.g., learners, teachers, curriculum developers, administrators, policy makers) and support the refinement and reform of language education and language qualifications, particularly in Europe. Since its introduction, it has been adapted and adopted worldwide (Figueras, 2012; Runnels & Runnels, 2019), and language tests are often expected to provide scores that can be interpreted in reference to the CEFR proficiency levels (Deygers et al., 2018).

The CEFR specifies a continuum of six major levels of language ability, from basic (Levels A1 and A2) to independent (B1 and B2) to proficient (C1 and C2) user or learner (Council of Europe, 2001, 2018). These levels are applied to the CEFR's descriptive scheme of language proficiency, which includes language competencies, activities, and strategies. Language competencies include the use of linguistic (e.g., vocabulary range and control), sociolinguistic (e.g., sociolinguistic appropriateness), and pragmatic (e.g., turn taking) knowledge. Language activities may involve reception (listening and reading comprehension), production (speaking and writing), interaction (speaking and writing), or mediation (facilitation and translation). Language strategies are elaborated in reference to language activities; for example, interactive language activities may involve strategies such as cooperating and asking for clarification. Every language competency, activity, or strategy defined by the CEFR's descriptive scheme has an associated set of descriptors that illustrate what a language user should be expected to do across the continuum of ability (i.e., A1 to C2). The CEFR manual and companion volume include dozens of descriptor sets, which are parsed by this continuum of ability. Consequently, what it means to be "at" any particular CEFR level of language ability (A1 to C2) is largely defined by the illustrative descriptors associated with the ability being referenced.

Overview of the Recommended CEFR Mapping Process

The Council of Europe's manual for mapping test scores to CEFR levels states that a test developer should make a specific claim about the relationship between test scores and the CEFR and support that claim with theoretical and empirical evidence (Council of Europe, 2009). In keeping with the descriptive scheme of the CEFR, this relationship involves specifying the intended interpretation about language ability based on test scores—and thus, which CEFR descriptors are most relevant—as well as empirical research to relate test scores (or ranges of

scores) to relevant CEFR proficiency levels. Consequently, the manual's recommended mapping process involves building an argument backed by evidence across four main stages or procedures: familiarization, specification, standard setting, and validation. These stages essentially involve three overarching activities: promoting familiarization with the CEFR (familiarization), describing the test and evidence of its quality and how the test relates to the CEFR (specification), and providing an empirical basis for relating test scores to specific CEFR proficiency levels (standard setting and validation). Although the process begins with the familiarization stage, familiarization activities should be incorporated into the subsequent stages of specification and standard setting.

The redesigned TOEIC Bridge tests (hereafter, TOEIC Bridge tests) were designed to facilitate interpretations about a test taker's CEFR level for listening, reading, speaking, and writing proficiency in English in everyday life, from Pre-A1 to B1. This report describes the aspects of test design and the research activities conducted to elaborate and support claims about how TOEIC Bridge test scores map to CEFR proficiency levels. In keeping with the recommendations of the Council of Europe's (2009) manual, this report summarizes evidence pertaining to the stages of familiarization, specification, standard setting, and validation.

Familiarization

The familiarization stage involves activities designed to promote a shared understanding of relevant CEFR levels and descriptors among project team members (Council of Europe, 2009). Typically, this stage involves documenting CEFR familiarization activities for panelists in the standard setting phase (e.g., Papageorgiou, 2010), but familiarization may be more broadly conceived to describe how knowledge of the CEFR was acquired and utilized by test developers during test development, as described in the specification stage (e.g., O'Sullivan, 2010). Thus, the manual states that familiarization is distinct from other stages in that it is expected to occur repeatedly throughout the mapping process. A higher degree of familiarization with the CEFR by all project team members (test developers, researchers) is expected to enhance the quality of the overall process, as well as the quality of panelists' judgments in standard setting studies.

The TOEIC Bridge test development team included researchers and item writers who consulted CEFR descriptors throughout the test design process, in accordance with this broader

view of familiarization. The test design process included numerous activities involving the CEFR's descriptive scheme, as described in the Specification section below. These activities required the test development team to identify, categorize, revise, and reflect upon relevant CEFR descriptors at the targeted proficiency levels.

Separately, the panelists in each of four standard setting sessions engaged in familiarization activities to ensure they had an adequate understanding of relevant CEFR levels and descriptors. These activities are referenced in the Standard Setting section below and briefly summarized here. Prior to each standard setting session, panelists were asked to carefully review a familiarization manual. The manual included an overview of the CEFR, sets of CEFR descriptors at relevant proficiency levels (i.e., Pre-A1 to B1), and an activity to elaborate features of descriptors that helped distinguish different levels of CEFR proficiency. These activities were in line with Tannenbaum and Cho's (2014) recommendation for familiarization activities in standard setting studies: Panelists need to acquire a clear understanding of relevant levels and what differentiates a level from the next highest level. Panelists were encouraged to bring their notes from this activity to the standard setting session and draw upon them during group discussions aimed at consolidating a mutual understanding of the language knowledge and skills needed to be classified at each level.

In a premeeting questionnaire, panelists also indicated their familiarity with the CEFR in general and the CEFR descriptors associated with the particular standard setting session for which they were training (e.g., familiarity with CEFR descriptors related to Spoken Production for the TOEIC Bridge Speaking session). All panelists across all sessions indicated that they were somewhat or very familiar with the CEFR in general. All panelists in the TOEIC Bridge Listening, Reading, and Writing sessions indicated that they were somewhat or very familiar with the specific CEFR descriptors relevant to their session, and 12 of 15 panelists in the TOEIC Bridge Speaking session indicated the same.

Thus, the effort to map TOEIC Bridge tests' scores to CEFR levels involved a variety of familiarization activities for both the test development team and standard setting panelists. The familiarization activities helped the test development team form clear hypotheses about the CEFR level(s) that may be required to successfully respond to different test tasks and,

consequently, the range of proficiency levels each test should be expected to evaluate. This familiarity was important because, as the Specification section explains, item specifications were developed with targeted CEFR proficiency levels in mind. A separate group of panelists were required to complete familiarization activities in advance of standard setting meetings to ensure they reflected on the meaning of and distinction between relevant CEFR levels.

Specification

The specification stage involves a description of the test's content and quality and a description of the test's (intended) relationship with the CEFR (Council of Europe, 2009). The latter description is similar to what experts characterize as the "construct congruence" between a test and the framework to which the test will be mapped (Tannenbaum & Cho, 2014). More elaborate descriptions of the content and measurement quality of the TOEIC Bridge tests are available elsewhere and will only be briefly summarized here. The construct congruence between the tests and CEFR will be more fully detailed.

Content and Measurement Quality of the TOEIC Bridge Tests

The TOEIC Bridge tests were designed to measure the reading, listening, speaking, and writing proficiency of English learners at beginning to low-intermediate levels in the context of everyday adult life. Test takers will be young adults (secondary school students) and adults for whom English is a foreign language, and their nationalities and native languages will vary. Test takers' educational backgrounds and purposes for learning English (e.g., academic purposes, occupational purposes) may also vary. The tests were designed to support selection decisions in contexts where English language proficiency is desirable or needed, to make placement decisions for instructional or training purposes, and to verify a learner's current level of proficiency to determine readiness for more advanced study (Schmidgall et al., 2019). The TOEIC Bridge tests are module-based in the sense that various combinations of the four tests can be administered based on score users' needs. The listening and reading tests are paper-based, while the speaking and writing tests are computer-delivered. For all tests, scaled scores range from 15 to 50.

All of the tests adopt a construct definition—or definition of the ability to be tested—in which test takers demonstrate their ability by using their linguistic knowledge and subcompetencies to achieve communication goals (Schmidgall et al., 2019). The relevant linguistic knowledge and subcompetencies slightly vary based on the test but generally include lexical knowledge, grammatical knowledge, discourse knowledge, phonological (or orthographic) knowledge, pragmatic competence, and strategy use.

Listening Test

The TOEIC Bridge Listening test measures the ability of beginning to lower-intermediate English language learners to understand short spoken conversations and talks in personal, public, and familiar workplace contexts. Test takers demonstrate their ability by using their linguistic knowledge and subcompetencies to achieve communication goals. Linguistic knowledge and subcompetencies include the ability to (a) understand high-frequency vocabulary and formulaic phrases (lexical knowledge); (b) understand simple sentences and structures (grammatical knowledge); (c) understand sentence-length speech and some common registers (discourse knowledge); (d) recognize and distinguish English phonemes and the use of common intonation and stress patterns and pauses to convey meaning in slow and carefully articulated speech across familiar varieties (phonological knowledge); (e) infer implied meanings, speaker roles, or context in short, simple spoken texts (pragmatic competence); and (f) understand the main idea and stated details in short spoken texts (listening strategies). The communication goals targeted by the test include comprehending simple greetings, introductions, and requests; instructions and directions; descriptions of people, objects, situations; personal experiences or routines; and other basic exchanges of information (see Schmidgall et al., 2019, p. 16).

The TOEIC Bridge Listening test consists of 50 items administered across four parts or task types and takes approximately 25 min to complete. The first part, Four Pictures, includes six items. In Four Pictures, test takers hear one short phrase or sentence spoken aloud and must choose the picture that the phrase or sentence describes. The task is designed to evaluate test takers' ability to understand simple descriptions of people, places, objects, and actions.

The second part, Question-Response, includes 20 items. In Question-Response, test takers hear a question or statement spoken aloud. Each question or statement is followed by four responses that are spoken aloud and written in the test booklet. Test takers must choose the best response to each question or statement. The task is designed to evaluate test takers' ability to understand very short dialogues or conversations on topics related to everyday life.

The third part, Conversations, includes 10 items. In Conversations, test takers hear some short conversations (i.e., dialogues) and must answer two questions about each conversation. Some conversations may include a visual (e.g., short menu, list of ticket prices) that is relevant to the conversation. After listening to a short conversation, test takers hear and read the questions in the test booklet and choose the best answer to the question from four written options.

The fourth part, Talks, includes 14 items. In Talks, test takers hear some short talks (i.e., monologues) and must answer two questions about each talk. As in the previous task, some talks may include a visual that is relevant to the talk. After listening to a short talk, test takers hear and read the questions in the test booklet and choose the best answer to the question from four options. This task is designed to evaluate test takers' ability to understand short monologues as they occur in everyday life when they are spoken slowly and clearly. Test takers are expected to use all of their linguistic knowledge and subcompetencies, including pragmatic competence.

The reliability of listening test scores is reported using a measure of internal consistency, *KR-20*, which was found to be .90 in norming samples (ETS, 2019). Reliability coefficients greater than .70 are generally considered acceptable, and coefficients greater than or equal to .90 are considered very good (Chapelle, 2013). The standard error of measurement is 3 scaled score points. In an initial validity study, Schmidgall (2020) found that the correlation between test takers' self-assessments of listening ability were correlated (r = .55) with TOEIC Bridge Listening test scores. Although this is only a moderate correlation, it compares favorably with similar research that investigates the relationship between test scores and self-assessments of language ability (for a discussion, see Schmidgall, 2020).

Reading Test

The TOEIC Bridge Reading test measures the ability of beginning to lower-intermediate English language learners to understand short written English texts in personal, public, and familiar workplace contexts and across a range of formats. Test takers demonstrate their ability by using their linguistic knowledge and subcompetencies to achieve communication goals. Linguistic knowledge and subcompetencies include the ability to (a) understand common vocabulary (lexical knowledge); (b) understand simple sentences and structures (grammatical knowledge); (c) understand the organization of short written texts in a variety of formats (discourse knowledge); (d) recognize simple mechanical conventions of written English (orthographic knowledge); (e) infer implied meanings, including context or writer's purpose, in short, simple written texts (pragmatic competence); and (f) understand the main idea and stated details in short written texts and infer the meaning of unknown written words through context clues (reading strategies). The communication goals targeted by the test include understanding nonlinear written texts; written instructions and directions; short, simple correspondence; and short information, descriptive, and expository written texts about people, places, objects, and actions (see Schmidgall et al., 2019, pp. 16–17).

The TOEIC Bridge Reading test consists of 50 items, administered across three parts or task types, and takes approximately 35 min to complete. The first part, Sentence Completion, includes 15 items. In Sentence Completion, test takers are presented with a sentence that has a missing word or phrase. Test takers must then review four options and select the word or phrase that best completes the sentence.

The second part, Text Completion, includes 15 items. In Text Completion, test takers read short texts in a variety of formats. Each short text is missing three 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 four options.

The third part, Reading Comprehension, includes 20 items. In Reading Comprehension, test takers must read everyday texts (e.g., notices, letters, forms, advertisements) and answer two or three questions about each text. The questions accompanying each text may require the

test taker to identify the main idea, identify stated details, or infer implied meanings such as the context or the writer's purpose.

The reliability of reading test scores is reported in the same manner as listening test scores: *KR-20* = .90 (ETS, 2019). As with the listening test, the standard error of measurement is 3 scaled score points. In an initial validity study, Schmidgall (2020) found that the correlation between test takers' self-assessments of reading ability were correlated (r = .54) with TOEIC Bridge Reading test scores.

Speaking Test

The TOEIC Bridge Speaking test measures the ability of beginning and lower-intermediate English language learners to carry out spoken communication tasks in personal, public, and familiar workplace contexts. Test takers demonstrate their ability by using their linguistic knowledge and subcompetencies to achieve communication goals. Linguistic knowledge and subcompetencies include the ability to (a) use high-frequency vocabulary appropriate to a task (lexical knowledge); (b) use common grammar structures to contribute to overall meaning (grammatical knowledge); (c) use simple transitions to connect ideas (discourse knowledge); (d) pronounce words in a way that is intelligible to proficient speakers of English, using intonation, stress, and pauses to pace speech and contribute to comprehensibility (phonological knowledge); and (e) produce speech that is appropriate to the communication goal (pragmatic competence). The communication goals targeted by the test include asking for and providing basic information; describing people, objects, places, and activities; expressing an opinion or plan and giving a reason for it; giving simple directions; making simple requests, offers, and suggestions; and narrating and sequencing simple events (see Schmidgall et al., 2019, p. 17).

The TOEIC Speaking test consists of six speaking tasks (eight questions overall) and takes approximately 15 min to complete. All speaking tasks have their own scoring rubric that consists of either 3 score points (Tasks 1–4) or 4 score points (Tasks 5–6).

The first two tasks, Read a Short Text Aloud and Describe a Photograph, are each repeated twice for the first four questions of the test. In Read a Short Text Aloud, test takers read aloud a short presentational text that is displayed on their screen. Test takers have 20 s to

prepare and 30 s to read the text aloud. The task is designed to evaluate a linguistic subcompetency, phonological knowledge, and use (i.e., intelligibility). In Describe a Photograph, test takers view a picture on their screen and describe it in as much detail as possible. The picture contains people engaging in activities in context, so test takers are directed to describe where the people are and what they are doing. Test takers have 30 s to prepare and 30 s to speak.

The remaining four tasks are Listen and Retell, Short Interaction, Tell a Story, and Make and Support a Recommendation. In the Listen and Retell task, test takers listen to a person talking about a topic (e.g., an announcement at a train station) and then must relate or summarize what they have just heard to someone else (e.g., to a coworker who missed the announcement). After listening to the announcement, test takers have 10 s to prepare and 30 s to speak. In the Short Interaction task, test takers use visual information on the screen (e.g., a note with a few bullet points) to complete a short communicative task (e.g., leaving a voice-mail message with several questions). Test takers have 20 s to prepare and 30 s to speak. In Tell a Story, test takers look at four pictures that illustrate a story and narrate the story in their own words. They can describe places, people, actions, and feelings. Test takers have 45 s to prepare and 45 s to speak. In Make and Support a Recommendation, test takers describe information (e.g., options for a tour), make a recommendation about it (e.g., suggest a tour option), and provide support for the recommendation. Test takers have 45 s to prepare and 60 s to speak.

The reliability of speaking test scores is reported using a measure of internal consistency appropriate to the design of the test, stratified coefficient alpha. The reliability of the speaking test is approximately .86 (Lin et al., 2019). The standard error of measurement is 4 scaled score points. In an initial validity study, Schmidgall (2020) found that the correlation between Japanese and Taiwanese test takers' self-assessments of speaking ability were moderately correlated with TOEIC Bridge Speaking test scores (r = .47 and r = .48, respectively).

Writing Test

The TOEIC Bridge Writing test measures the ability of beginning and lower-intermediate English language learners to carry out written communication tasks in personal, public, and familiar workplace contexts. Test takers demonstrate their ability by using their linguistic

knowledge and subcompetencies to achieve communication goals. Linguistic knowledge and subcompetencies include the ability to (a) use high-frequency vocabulary appropriate to a task (lexical knowledge); (b) write a sentence using simple word order, such as subject-verb-object, interrogatives, and imperatives, and use common grammatical structures to contribute to meaning (grammatical knowledge); (c) arrange ideas using appropriate connectors and sequence ideas to facilitate understanding (discourse knowledge); (d) control mechanical conventions of English to facilitate comprehensibility of text (orthographic knowledge); and (e) produce text that is appropriate to the communication goal (pragmatic competence). The communication goals targeted by the test include asking for and providing basic information; making simple requests, offers, and suggestions and expressing thanks; expressing a simple opinion and giving a reason for it; describing people, objects, places, and activities; and narrating and sequencing simple events (see Schmidgall et al., 2019, pp. 17–18).

The TOEIC Bridge Writing test includes five tasks (nine questions overall) and takes approximately 37 min to complete. The first task (Build a Sentence) is machine-scored as correct or incorrect, and the remaining tasks have their own scoring rubric that consists of either 3 score points (Tasks 2–4) or 4 score points (Task 5).

The first two tasks, Build a Sentence and Write a Sentence, are each repeated three times for the first six questions of the test. In Build a Sentence, test takers must drag and drop words (or phrases) to form a grammatically correct sentence. All of the words (or phrases) must be used to form the sentence, and there is a single key (i.e., only one correct response is possible). Test takers have 60 s to complete the sentence. In Write a Sentence, test takers view a picture on their screen and use two supplied words (or phrases) to write one sentence. Test takers have 60 s to write the sentence.

The remaining three tasks include Respond to a Brief Message, Write a Narrative, and Respond to an Extended Message. In Respond to a Brief Message, test takers must read and respond to several requests by providing suggestions and answering questions. The requests are presented as an instant message, an everyday and often informal medium of communication, but test takers are instructed to respond clearly and fully to the instant message to avoid the use of texting language. Test takers have 8 min to prepare and write a

response, which typically includes two components (e.g., give two gift suggestions and answer a question about lunch). In Write a Narrative, test takers write a short narrative about an everyday topic (e.g., a time when you helped a friend). Test takers have 8 min to prepare and write a response. In Respond to an Extended Message, test takers read and respond to questions in an e-mail. The questions in this task differ from those in the Instant Message task in that they require test takers to express a simple opinion and give reasons for the opinion. The context also differs across tasks (i.e., text message vs. e-mail), and this written task is expected to involve a greater degree of organization, development, and audience awareness. Test takers have 10 min to prepare and write a response.

The reliability of writing test scores is reported in the same manner as the speaking test and is approximately .80 (Lin et al., 2019). The standard error of measurement is 4 scaled score points. In an initial validity study, Schmidgall (2020) found that the correlation between Japanese and Taiwanese test takers' self-assessments of writing ability were moderately correlated with TOEIC Bridge Writing test scores (r = .45 and r = .61, respectively).

Construct Congruence Between the TOEIC Bridge Tests and the CEFR

One of the initial mandates for test development of TOEIC Bridge tests was the need to map scores to language proficiency standards, and the specification of the content and performance standards for the tests were directly informed by test developers' familiarization with the CEFR. The tests were developed using a mandate-driven approach to evidence-centered design in which a domain analysis was used to justify a proposed construct definition (Schmidgall et al., 2019). The domain analysis began by defining the content standard of English reading, listening, speaking, and writing proficiency in the context of everyday adult life. The conceptualization of the context—the target language use domain of "everyday adult life"—was directly informed by familiarization with the CEFR. The authors of the CEFR highlight four general domains of language use: personal, public, occupational, and academic (Council of Europe, 2001, 2018). To the extent that the context of language use is referenced in CEFR descriptors, descriptors at lower levels of proficiency tend to emphasize the personal and public domains. As learners progress into intermediate and advanced levels, they are expected to have the skills needed to use language in more demanding, specific-purposes contexts such as

occupational and academic settings. Consequently, the target language use domain of the test was defined to largely include language used in personal and public settings, as well as some general workplace settings (i.e., for tasks that target learners at high-beginner to low-intermediate proficiency levels).

As described by Schmidgall et al. (2019), the first phase of the domain analysis produced an initial construct definition that effectively served as the content standard for the test. At this point, researchers conducted a review of the CEFR descriptor scales most relevant to this content standard. This review included relevant descriptor scales from the communicative language activities of Reading Comprehension, Listening Comprehension, Spoken Production, Spoken Interaction, Written Production, Written Interaction, and Online Interaction, as well as communicative language competencies (linguistic, sociolinguistic, pragmatic). The review of descriptor scales focused on the proficiency levels relevant to the target language use domain (Pre-A1 to B1) and produced summaries that helped refine the content standard and establish the proficiency standard for subsequent stages of test development. Table 1 lists the descriptor scales included in this review.

Table 1. Common European Framework of Reference (CEFR) Descriptor Scales Included in the Domain Analysis for the TOEIC Bridge Tests

CEFR communicative language activity, strategy, or competency descriptor scales	TOEIC Bridge			
	Reading	Listening	Speaking	Writing
Reading comprehension				
Overall reading comprehension	✓			
Reading correspondence				
Reading for orientation				
Reading for information and argument				
Reading instructions				
Listening comprehension				
Overall listening comprehension		✓		
Understanding conversation between other speakers				
Listening as a member of a live audience				
Listening to announcements and instructions				
Listening to audio media and recordings				
Reception strategies	✓	✓		
Identifying cues and inferring				

CEFR communicative language activity, strategy, or competency descriptor scales	TOEIC Bridge			
	Reading	Listening	Speaking	Writing
Spoken production				
Overall spoken production				
Sustained monologue: describing experience			✓	
Sustained monologue: giving information				
Sustained monologue: putting a case				
Public announcements				
Written production				
Overall written production				✓
Creative writing				
Written reports and essays				
Spoken interaction				
Informal discussion				
Obtaining goods and services			✓	
Information exchange				
Phonological control				
Written interaction				
Overall written interaction				✓
Correspondence				
Notes, messages, and forms				
Online interaction				✓
Online conversation and discussion				
Linguistic				
General linguistic range				
Vocabulary range	✓	✓	✓	✓
Grammatical accuracy				
Vocabulary control				
Sociolinguistic	✓	✓	✓	✓
Sociolinguistic appropriateness				
Pragmatic				
Thematic development				
Coherence and cohesion	✓	✓	✓	✓
Propositional precision				
Spoken fluency				

The list of descriptor scales in Table 1 illustrates the intended alignment between the TOEIC Bridge tests and the CEFR (for Levels Pre-A1 to B1, the performance standard of the tests). For example, the construct definition for the TOEIC Bridge Reading test incorporates an analysis of descriptor scales for the language activity Reading Comprehension (overall reading comprehension, reading correspondence, reading for orientation, reading for information and

argument, reading instructions), Reception Strategies (identifying cues and inferring), and the language competencies Linguistic (general linguistic range, vocabulary range, grammatical accuracy, vocabulary control), Sociolinguistic (sociolinguistic appropriateness), and Pragmatic (thematic development, coherence and cohesion, propositional precision, spoken fluency). It does not include all descriptor scales potentially relevant to reading proficiency, such as the language activity Reading as a Leisure Activity; scales were omitted when they were judged to be less relevant to the content standard as informed by the initial mandate for test design.

The domain analysis also produced documentation that summarized expected language activities, strategies, and competencies across the CEFR proficiency levels Pre-A1 to B1. This documentation directly informed subsequent test development and was integrated into task specifications as described by Everson et al. (2019).

As a result of this process, familiarization with the CEFR directly influenced the development of the TOEIC Bridge tests and established a clear relationship between the tests and the CEFR. The content standard of the tests had substantial overlap with relevant descriptor scales from the CEFR. The performance standard of the tests was directly informed by the proficiency levels specified in the CEFR: Pre-A1, A1, A2, A2+, and B1. This extended from construct definition through task and test specifications, wherein tasks were designed to target specific ranges of proficiency as defined in the CEFR (Everson et al., 2019).

Standard Setting

The purpose of standard setting is to determine the minimum level of performance needed on a test in order to achieve specified performance standards (Hambleton & Pitoniak, 2006), such as CEFR levels (Council of Europe, 2009). The minimum level of performance needed is based on the collective judgment of a panel of experts who are trained to use a standard setting methodology. Experts have provided a number of recommendations to guide the selection of panelists and standard setting approach as well as guidance on how to document the process for the purpose of validation (see Cizek & Bunch, 2007; Tannenbaum & Cho, 2014). For example, the selection of panelists and documentation of their characteristics is a critical facet of a standard setting study as its outcome rests primarily on panelists' collective judgment. Because dozens of standard setting methodologies are available and the choice of

method may impact the results (Cizek & Bunch, 2007), the selection of an appropriate method is another critical consideration. The series of standard setting sessions reported here align closely with expert recommendations, which are further elaborated in relevant sections.

TOEIC Bridge Test Data

Prior to the standard setting study, the project team obtained TOEIC Bridge test data from psychometricians at ETS. The data were collected from two test forms administered to a total of 2,368 test takers in Brazil, Colombia, Japan, Korea, Mexico, and Taiwan as described by Lin et al. (2019). Because the listening and reading sessions used an item-centered standard setting methodology, the data included one of the test forms and its associated item statistics, including item difficulty (*p*) and item discrimination (point-biserial correlation). Because the speaking and writing sessions used a person-centered method, data consisted of representative samples of test-taker responses for each point on the speaking and writing score scale. These representative samples were obtained by identifying the most frequent score profiles (i.e., patterns of scores across tasks) associated with each point on the score scale and then gathering the test responses of several test takers with each score profile.

Panelist Selection and Training

The standard setting panel for each session consisted of 15 panelists, with the exception of the panel for the reading session, which consisted of 14 panelists. The size of each panel was in line with recommendations by experts, who have alternately suggested that a panel be composed of at least 10 judges (Tannenbaum & Cho, 2014) and up to 20 judges (Hambleton et al., 2014); the Council of Europe (2009) has recommended 12 to 15 judges. Each panel was composed of experts in language teaching, learning, and assessment affiliated with ETS. A total of 27 experts (23 test developers and four research assistants) participated in at least one session. A majority of experts participated in two sessions, but some participated in only one and several participated in all four sessions (see Appendix A). None of the experts were on the redesigned TOEIC Bridge test core project team. This selection criteria was imposed to ensure that familiarization with item and test specifications—which included expectations about the CEFR levels that different item types should be expected to target—would influence the

standard setting procedure. Typically, experts recommend that panels include representation from a diverse group of major stakeholders (Hambleton & Pitoniak, 2006), so this narrow institutional affiliation is atypical. There were several reasons for utilizing this atypical approach. By drawing upon a relatively large pool of in-house experts, ETS was able to involve a relatively large number of experts (27) while maintaining an appropriately sized panel for each session (14 to 15) and constitute a unique panel for each of the four standard setting sessions. This approach may be impractical when involving a diverse group of external stakeholders, but was manageable in ETS's situation where a large group of experts—independent of the core project team—was accessible.

Panelists completed a background questionnaire prior to participating in their first panel to provide documentation of their characteristics and relevant expertise. There were more women than men on each panel; the number of men on each panel ranged from one to six. Panelists reported their age in 10-year ranges, and although each panel included panelists from the 21 to 30 range to the over 50 range, the median age was 41 to 50 for all panels.

Panelists reported having extensive experience with English teaching and assessment as well as familiarity with the population of English learners targeted by the TOEIC Bridge tests (secondary school and adult). The average number of years of English teaching experience for panelists in each panel ranged from 13 to 15. The average number of years of English assessment experience for panelists in each panel was approximately 14. All panelists in each panel reported having some familiarity or being very familiar with the adult English language learner population. Most panelists also indicated that they had some familiarity with the secondary school English language learner population.

As previously described in the Familiarization section, most panelists had knowledge of the CEFR prior to the study. Across sessions, all panelists consistently indicated that they had some familiarity or were very familiar with the CEFR in general. Panelists also indicated that they already had some familiarity or were very familiar with the specific CEFR descriptors relevant to the session in which they would be participating, with the exception of the speaking session where three panelists indicated they were not familiar with the descriptors.

A majority of panelists reported having at least some familiarity with the TOEIC Bridge tests prior to the study, but a sizable minority in each panel indicated they were not familiar with the tests. Each panel's lack of familiarity with the test prior to the study may seem surprising, given that panelists were all affiliated with the test developer. The standard setting study project team was aware that the TOEIC Bridge test and item specifications may contain hypotheses about the relationship between test tasks and the CEFR and sought to recruit panelists with little or no prior knowledge of the TOEIC Bridge test in order to ensure the independence of judges. Consequently, the vast majority of experts in each panel (ranging from 11 to 13) indicated in the background survey that they had no knowledge of the TOEIC Bridge test or item specifications.

Panelists also engaged in a training or familiarization activity prior to each session. As partially described in the Familiarization section, panelists were asked to spend 2 hr reviewing a familiarization manual that included material and activities related to the CEFR, the relevant TOEIC Bridge test, and the standard setting methodology that would be utilized for the upcoming session.

Procedure

Each of the four sessions followed the same general procedure and lasted 1 full day (8 hr), including breaks. Each day began with a presentation by the facilitator, who introduced the purpose of the session. Next, the facilitator provided a brief overview of the TOEIC Bridge test (listening, reading, speaking, or writing) and panelists completed the test individually.

Because all of the sessions used methodologies that relied on the concept of the just-qualified candidate (JQC), the panel then focused on creating definitions of JQCs at CEFR proficiency levels A1, A2, and B1. To begin, the facilitator introduced and led a discussion of the concept of the JQC. The JQC for any proficiency level is an imagined candidate who has just crossed the threshold separating that level from the level just below it (Tannenbaum & Wylie, 2008). The JQC is imagined based on CEFR proficiency level descriptors and panelists' knowledge of language learners' developmental characteristics. Since CEFR level descriptors elaborate the characteristics of language users within each level, the descriptors pertain to a relatively wide spectrum of proficiency—not the JQC. Thus, experts must utilize their

knowledge and experience to adapt or supplement the existing descriptors with a JQC in mind. The JQC descriptions produced by the panel need to be accepted and commonly understood by the group as a whole, as they define the shared performance standard that individual panelists are expected to reference as they make judgments during the standard setting procedure (Zeidler, 2016).

After introducing and discussing the concept of the JQC, the facilitator separated the panelists into two groups who worked independently to define the CEFR A2 JQC. Panelists were randomly assigned to groups, and each group was overseen by a facilitator who assigned one panelist in the group to document the group's JQC definition. Each group discussed how existing CEFR descriptors may be modified to better describe the A2 JQC, as well other language knowledge and skills the JQC may exhibit based on their expertise. The groups reunited and presented their definitions to each other, discussing similarities and differences before arriving at a consensus definition consisting of five to seven bullet points. The panelists then separated into two groups again, with one group focused on defining the CEFR A1 JQC and the other on defining the CEFR B1 JQC. After reuniting, each group presented their definition and further refined it through group discussion. The motivation for having both groups initially work on the same JQC (i.e., A2) independently—and then discuss their results to negotiate a consensus JQC—was to reinforce the idea that each group's initially drafted JQC would be subject to refinement through discussion with equally qualified colleagues. The JQC descriptors produced by this process are reproduced in Appendices B (listening), C (reading), D (speaking), and E (writing).

The panel then completed training and practice on the standard setting method, followed by three judgment rounds. A modified Angoff method (Plake & Cizek, 2012) was used for the listening and reading sessions, and the Performance Profile approach (Tannenbaum & Cho, 2014; Zieky et al., 2008) was used for the speaking and writing sessions. The modified Angoff method is well suited for the listening and reading test and the requirements of this judgment context, and it is one of the most well-studied approaches to standard setting. The original Angoff method and modified versions of it were designed for use with multiple choice questions, as are used in the listening and reading tests. As one of the most popular standard

setting methods (or more properly, family of methods) for more than 40 years, it has been thoroughly researched and often used in language testing contexts such as mapping test scores to CEFR proficiency levels (e.g., Baron & Papageorgiou, 2014; Tannenbaum & Wylie, 2008). The Performance Profile approach is appropriate for performance-based tasks with relatively few items and has been previously used to map test scores to CEFR proficiency levels (Tannenbaum & Cho, 2014).

Listening and Reading Sessions (Modified Angoff Method)

The facilitator provided an overview of the modified Angoff method (Plake & Cizek, 2012) followed by multiple examples and group discussion about how to apply it. Panelists then completed a brief survey that allowed them to provide feedback on the training sessions and to indicate whether they were ready to proceed with the first round of standard setting judgments. This formal step occurred for two reasons: to ensure that panelists had the opportunity to raise concerns without fear of losing face with colleagues and to collect process-related documentation for the purpose of validation. The facilitator quickly reviewed the survey data and addressed any panelist concerns before proceeding to the next step.

Panelists then completed a three-round judgment process to determine the recommended minimum TOEIC scores for each of the targeted CEFR levels. In the first round, panelists made independent judgments in a prepared spreadsheet in accordance with the standard setting method for the session, focused on CEFR proficiency levels A1, A2, and B1. For each item in the test, participants first considered how many A1 JQCs—from a group of 100 A1 JQCs—would be expected to get the item correct. Panelists entered the number of A1 JQCs as a multiple of 5 from 0 to 100. Next, panelists considered how many A2 JQCs (as a multiple of 5 from 0 to 100) would get the item correct. Finally, panelists considered how many B1 JQCs would get the item correct—again, as a multiple of 5, from 0 to 100. Panelists repeated this process for all 50 items in the test. An excerpted sample of a completed spreadsheet—for Panelist 1's judgments in the listening session—is shown in Figure 1.

Figure 1. Sample of a Completed Spreadsheet for Round 1 Judgments Using the Modified Angoff Method

	A	B	C	D
1	Panelist / Item	1		
2		A1	A2	B1
3	1	80	90	100
4	2	40	70	95
5	3	40	60	80
6	4	35	60	75
7	5	45	55	70
8	6	45	65	80
9	7	70	90	100

After the first round of judgments, the facilitator presented a summary of the results to the panel, focusing on points of disagreement that were then discussed within the group. For each item and JQC that was evaluated (i.e., A1, A2, and B1), the facilitator presented the average rating (0 to 100), standard deviation, minimum, and maximum. Given the large amount of information and limited time for discussion, the facilitator highlighted and encouraged discussion around three or four items for each JQC after identifying items that had relatively high standard deviations. The presentation and discussion also included a review of item statistics (difficulty, discrimination), which were then provided to participants to reference during the next judgment round.

After a break, panelists completed a second judgment round using the prepared spreadsheet where they were given the opportunity to review all 50 test items and revise the judgments they made in the first round. This round was followed by another brief presentation by the facilitator, who explained how judgments were converted into recommended minimum cut scores for each CEFR level (A1, A2, and B1). Participants were able to view the A1, A2, and B1 cut scores that had been produced by their item-level judgments in Rounds 1 and 2, as well as the panel's average, minimum, and maximum cut scores. The facilitator then led a group discussion focused on differences in cut scores between panelists, as well as the group's current consensus recommendations (i.e., the minimum cut scores based on averages for the group).

In a third and final round of judgments, panelists entered their final recommended minimum scores into their spreadsheet. In this final round, panelists added holistic cut score judgments for A2+ and B1+. Because the CEFR contains descriptors for A2+, B1+, and B2+, it may be useful for stakeholders to have a more refined mapping that includes relevant "plus" levels (i.e., A2+ and B1+). The facilitator presented the panel's average cut score recommendation to the group, and panelists completed a final evaluation survey to provide feedback on various aspects of the session.

Speaking and Writing Sessions (Performance Profile Approach)

The facilitator introduced the Performance Profile approach (Zieky et al., 2008), and the group practiced applying it to sample test-taker responses. In contrast to the item-focused modified Angoff method, the Performance Profile approach focuses on test takers' responses to tasks (i.e., speaking or writing responses). Prior to the study, exemplar sets of test-taker responses associated with each raw score point were identified by the project team, as described earlier in the TOEIC Bridge Test Data section. The goal of the judgment task is to identify the set of test-taker responses that is best aligned with each JQC description. After the training session, panelists completed a brief survey to provide feedback and indicate whether they were ready to proceed with the judgment task.

The speaking and writing sessions used a three-round judgment procedure that largely followed the organizational structure described for the listening and reading sessions, although the judgment procedure itself differed. In the first round of judgments, participants were asked to begin by reviewing the scoring rubrics and descriptors for the A2 JQC. While panelists imagined an A2 JQC, the facilitator identified a test taker's raw score and played the audio of their response set from the TOEIC Speaking test. Panelists were encouraged to take notes while listening. The facilitator then asked the panel whether they wanted to hear another test taker's response set at a higher, lower, or the same score point. This process was repeated until all panelists expressed satisfaction with being able to individually identify the cut score for JQC A2. This entire process was then repeated for JQC A1 and B1. The writing session followed a similar procedure, except panelists were able to work independently by individually reviewing test takers' response sets rather than as a group.

After the first judgment round, the facilitator summarized the results for the panel, including the average cut scores and the range (minimum and maximum) associated with each JQC (A1, A2, B1). This was followed by a discussion of these results, and panelists offered their rationales for their judgments; for example, why an A2 JQC might be associated with a higher or lower score point based on the panelist's understanding of the JQC and the performance samples associated with various score points. This discussion was followed by a second round of judgments, repeating the same process from the Round 1 where panelists either listened to test takers' audio responses as a group (speaking session) or reviewed test takers' written responses independently (writing session). After Round 2, the facilitator again summarized the results for the panel, comparing them to Round 1 and encouraging discussion within the group. In the third judgment round, panelists were asked to make their final holistic judgments for the A1, A2, and B1 cut scores and add cut scores for A2+ and B1+. After Round 3 judgments, the facilitator presented the cut scores recommended by the panel to the group, and panelists completed a final evaluation survey.

Results

The results for each judgment round for each test are summarized in Tables 2 to 5. For each round, the mean, minimum, maximum, standard deviation, and standard error of judgment for each CEFR level are included. The standard error of judgment is an estimate of uncertainty in judgment, computed by dividing the standard deviation of judgments by the square root of the number of panelists and interpreted as an indicator of the extent to which each recommended cut score is likely to be the recommended cut score of a similarly composed panel (in terms of its expertise and training; Papageorgiou et al., 2019). The Round 3 mean scores are considered the final recommendations of the panel. All data in the tables are expressed in terms of raw scores.

Listening

The results of the listening session are summarized in Table 2. The maximum raw score for the TOEIC Bridge Listening test is 50 points. The panel's average cut score recommendations for A1, A2, and B1 were mostly consistent across rounds, although all slightly decreased from

Rounds 1 to 3. The standard deviations were initially quite large in Round 1, but decreased substantially across rounds. The standard errors of judgment also decreased across rounds.

Table 2. Standard Setting Results for the TOEIC Bridge Listening Test

Levels	Round 1			Round 2			Round 3				
	A1	A2	B1	A1	A2	B1	A1	A2	A2+	B1	B1+
Mean	20.5	32.9	42.7	21.0	33.3	42.3	19.1	30.9	35.7	41.0	45.1
Minimum	7.9	18.8	32.4	10.2	19.5	33.2	15.0	25.0	32.0	39.0	43.0
Maximum	35.2	42.5	47.9	29.9	39.6	48.2	22.0	37.0	41.0	44.0	46.0
SD	8.5	6.8	4.6	5.0	5.1	4.0	2.3	3.0	2.4	1.7	0.8
SEJ	2.2	1.7	1.2	1.3	1.3	1.0	0.6	0.8	0.6	0.4	0.2

Note. SEJ = standard error of judgment.

Reading

The results of the reading session are summarized in Table 3. The maximum raw score for the TOEIC Bridge Reading test is 50 points. The panel's cut score recommendations for A1, A2, and B1 were similar across rounds, consistently decreasing by 1 or 2 points. The standard deviations and standard errors of judgment decreased across the rounds.

Table 3. Standard Setting Results for the TOEIC Bridge Reading Test

Levels	Round 1			Round 2			Round 3				
	A1	A2	B1	A1	A2	B1	A1	A2	A2+	B1	B1+
Mean	18.2	32.6	43.0	16.6	31.6	42.3	15.6	30.6	36.1	41.6	46.2
Minimum	9.4	29.1	39.7	9.4	27.9	38.9	12.0	28.0	34.0	40.0	45.0
Maximum	26.8	36.0	45.6	24.8	36.8	46.2	18.0	35.0	40.0	45.0	48.0
SD	4.7	2.3	1.8	4.3	2.6	2.5	1.5	1.6	1.7	1.5	1.2
SEJ	1.3	0.6	0.5	1.1	0.7	0.7	0.4	0.4	0.4	0.4	0.3

Note. SEJ = standard error of judgment.

Speaking

The results of the speaking session are summarized in Table 4. The maximum raw score for the TOEIC Bridge Speaking test is 34 points. The panel's cut score recommendations for A1, A2, and B1 were extremely consistent across rounds. The standard deviations and standard errors of judgment decreased across the rounds.

Table 4. Standard Setting Results for the TOEIC Bridge Speaking Test

Levels	Round 1			Round 2			Round 3				
	A1	A2	B1	A1	A2	B1	A1	A2	A2+	B1	B1+
Mean	17.7	24.7	29.1	17.9	24.3	28.9	17.8	24.3	27.0	28.8	32.2
Minimum	15.0	22.0	27.0	17.0	23.0	28.0	17.0	23.0	26.0	28.0	30.0
Maximum	22.0	28.0	30.0	20.0	27.0	30.0	20.0	27.0	28.0	29.0	33.0
SD	2.1	1.9	0.8	1.0	1.1	0.6	1.0	1.0	0.8	0.4	1.1
SEJ	0.5	0.5	0.2	0.3	0.3	0.2	0.3	0.3	0.2	0.1	0.3

Note. SEJ = standard error of judgment.

Writing

The results of the writing session are summarized in Table 5. The maximum raw score for the TOEIC Bridge Writing test is 32 points. As with the speaking session, the panel's cut score recommendations for A1, A2, and B1 were consistent across rounds. The standard deviations and standard errors of judgment decreased from Round 1 to Round 2 and retained comparably low levels of variance in Round 3.

Table 5. Standard Setting Results for the TOEIC Bridge Writing Test

Levels	Round 1			Round 2			Round 3				
	A1	A2	B1	A1	A2	B1	A1	A2	A2+	B1	B1+
Mean	13.5	18.2	24.0	13.3	18.1	23.6	13.3	18.0	21.3	23.7	28.4
Minimum	12.0	15.0	21.0	12.0	15.0	22.0	12.0	15.0	19.0	22.0	25.0
Maximum	16.0	20.0	27.0	15.0	19.0	25.0	15.0	19.0	23.0	25.0	31.0
SD	1.6	1.5	1.7	0.8	1.1	1.0	0.8	1.1	1.2	0.9	1.8
SEJ	0.4	0.4	0.4	0.2	0.3	0.3	0.2	0.3	0.3	0.2	0.5

Note. SEJ = standard error of judgment.

Poststudy Adjustments

A complete standard setting process should incorporate additional sources of information beyond the recommendations obtained from an expert panel (Geisinger & McCormick, 2010). These additional sources may include a consideration of organizational or societal needs, the error of measurement, or results from different standard setting sessions or techniques. In mapping the TOEIC Reading and Listening test to the Vietnam National Standard, Tannenbaum and Baron (2015) recommended that decision makers consider raising or lowering

the recommended cut scores by 1 SEM based on their needs. Based on feedback from decision makers and additional data analyses (including an investigation of the impact of revised cut scores on admissions decisions), Papageorgiou and his colleagues advised lowering the CEFR cut scores for the *TOEFL iBT*® using the standard error of measurement (Papageorgiou, Tannenbaum, et al., 2015).

Several considerations led to adjustments in the recommended cut scores using a multistep procedure. An overriding consideration for the project team was whether the reliability of the test could justify the use of five cut scores, including the "plus" levels (A2+, B1+). With this consideration in mind, psychometricians evaluated the classification consistency and accuracy (Livingston & Lewis, 1995) of various combinations of the recommended cut scores after they had been converted to weighted raw scores, including the following:

- Five cut scores/six levels (Pre-A1, A1, A2, A2+, B1, B1+)
- Four cut scores/five levels (Pre-A1, A1, A2, A2+, B1; Pre-A1, A1, A2, B1, B1+)
- Three cut scores/four levels (Pre-A1, A1, A2, B1)
- Two cut scores/three levels (A1, A2, B1)

Classification accuracy indicates the proportion of test takers who would be correctly classified into the same score level as their true score level. Classification consistency estimates the proportion of test takers who would be classified into the same score level if they took two parallel test forms. Although there is no strict "rule of thumb" for acceptable classification accuracy and consistency (Young & Yoon, 1998), applied research has expressed support for values higher than .60 (e.g., Papageorgiou, Morgan, & Becker, 2015; Papageorgiou, Xi, et al., 2015; Powers et al., 2016). For each test, and each combination of cut scores, estimates of classification accuracy and consistency were obtained for each proposed cut score. Overall estimates of classification accuracy and consistency were obtained for each combination of recommended cut scores as well. All of these estimates were examined to determine which combinations of cut scores yielded acceptable estimates of classification consistency and accuracy. The results of these analyses indicated that classification consistency and accuracy were too low for operational use when four or five cut scores were used. For the 5 cut-score

combination, overall estimates of classification accuracy ranged from .55 to .70, and estimates of classification consistency ranged from .44 to .61. For the 4 cut-score combination, classification accuracy ranged from .62 to .78, and classification consistency ranged from .51 to .71. For the 3 cut-score combination, classification accuracy ranged from .69 to .85, and classification consistency ranged from .60 to .79. For the 2 cut-score combination, classification accuracy ranged from .76 to .86, and classification consistency ranged from .68 to .80.

Following these initial analyses, the project team revisited the panelists' recommendations to see if slight adjustments to cut scores may improve the classification consistency and accuracy for the four cut-score or five cut-score models. The team looked to secondary data sources to justify any proposed modifications. For the listening and reading tests, the team examined the concordance between the redesigned and classic TOEIC Bridge test scores, and between the classic TOEIC Bridge and TOEIC test scores. Because these tests had been independently mapped to the CEFR and were part of the TOEIC family of assessments, the project team examined the coherence of the proposed cut scores with the score concordance tables in mind. The project team also considered the possibility of slight adjustments to cut scores within 1 *SD* of the current recommendations. For all of the tests, the team considered the practical implications of how recommended cut scores mapped to scaled scores. As a result, slight modifications to recommended cut scores were proposed for the listening and reading tests.

Psychometricians on the project team conducted another round of analyses of classification consistency and accuracy using several variations of the four cut scores/five levels combination. The results showed that the slight adjustments made to the cut scores resulted in similar estimates of classification consistency and accuracy, which were insufficient to justify the use of four cut scores and five levels. Therefore, a decision was made to report CEFR mapping for three cut scores and four levels (Pre-A1, A1, A2, and B1). For this combination of cut scores, the estimates of classification accuracy and consistency were .81 and .74, respectively, for the listening test; .85 and .79 for the reading test; .69 and .60 for the speaking test; and .73 and .65 for the writing test. The final recommended cut scores, converted to scale scores, are shown in Table 6.

Table 6. Final Recommended Cut Scores

Redesigned TOEIC Bridge test	Score scale range	Minimum score		
		A1	A2	B1
Listening	15–50	16	26	39
Reading	15–50	19	34	45
Speaking	15–50	23	37	43
Writing	15–50	20	32	43

The final recommended cut scores reflected the consideration of multiple sources of information while placing primary emphasis on panelists' judgment. The cut scores in Table 6 are scaled score conversions of panelists' raw score recommendations with just one minor exception—the Listening A1 cut score, which was adjusted from 15 to 16. Although panelists recommended cut scores for CEFR proficiency levels A2+ and B1+, analyses of classification consistency and accuracy determined that the misclassification rate using these cut scores would be too high for operational testing. Consequently, the team concluded that the three cut scores and four levels model was empirically defensible, conceptually sound, and was likely to meet the practical needs of stakeholders. Thus, Table 6 summarizes the claim about the relationship between TOEIC Bridge test scaled scores and CEFR proficiency levels Pre-A1, A1, A2, and B1. Since the A1 cut score is higher than the minimum scaled score for each test, scaled scores below the A1 cut score are interpreted as Pre-A1.

Validation

Validation is a critical step in the process of linking or mapping cut scores to performance descriptors, as it helps to establish the meaningfulness and credibility of the cut scores (Tannenbaum & Cho, 2014). Three primary sources of evidence are relevant to standard setting: procedural, internal, and external evidence (Council of Europe, 2009; Tannenbaum & Cho, 2014).

Procedural Evidence

Procedural evidence adds credibility to outcome of standard setting when it establishes that the panel was appropriately selected and qualified, that training procedures were

effective, and that the judgment process was conducted appropriately. The procedural evidence for this study draws upon feedback provided by the panelists over the course of each study.

One source of procedural evidence derived from panelist feedback comes from the survey each panelist completed after training and prior to beginning judgment rounds. Table 7 summarizes the results of this survey. The table includes the average panelist response to each question for each session, which used a 4-point Likert-type scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*). The high averages suggest that after the training sessions, panelists believed that they understood the purpose of the session, understood the definition of JQC, and understood the judgment task ahead of them. None of the panelists strongly disagreed with any statement in any session. All panelists across all sessions indicated that they were ready to proceed to the judgment rounds.

Table 7. Panelist Feedback After Training

Feedback	Listening	Reading	Speaking	Writing
I understand the purpose of the study.	3.8	3.9	3.7	3.9
The facilitator explained things clearly.	3.9	3.9	3.6	3.9
I understand the definition of the just qualified candidate.	3.9	3.9	3.9	3.9
The training in the standard setting method adequately prepared me to make my standard setting judgments.	3.6	3.8	3.5	3.9
The opportunity to practice helped clarify the standard setting task for me.	3.7	3.9	3.7	3.8
I understand how to make the standard setting judgments.	3.6	3.8	3.7	3.9
Are you ready to proceed and to make your first standard setting judgments? (% Yes)	100%	100%	100%	100%

Note. 1 = *strongly disagree*, 4 = *strongly agree.*

Another source of procedural evidence based on panelists' feedback comes from the evaluation survey, completed after the final recommended cut scores had been presented to the panel. The first section of the evaluation survey included five or six Likert-type questions similar to those included in the previous survey. These questions and panelists' average responses are shown in Table 8. These results are consistent with the findings of the previous

survey, and the high averages suggest that panelists believe they understood the purpose of the session and were satisfied with training activities and opportunities for feedback and discussion. None of the panelists in any session strongly disagreed with any of the statements.

Table 8. Panelists' Final Evaluations

Evaluation	Listening	Reading	Speaking	Writing
I understood the purpose of the study.	3.8	4.0	3.9	3.9
The instructions and explanations provided by the facilitator were clear.	3.8	3.9	3.5	3.9
The training in the standard setting method was adequate to give me the information I needed to complete my assignment.	3.6	3.9	3.5	3.9
The explanation of how the recommended cut score is computed was clear.	3.6	3.9	3.7	4.0
The opportunity for feedback and discussion between rounds was helpful.	3.6	4.0	3.7	3.9
The inclusion of the item and task data was helpful.	3.9	4.0		

Note. 1 = *strongly disagree*, 4 = *strongly agree.*

The second section of the evaluation survey asked panelists to quantify the extent to which different factors influenced their standard setting judgments. These factors and panelists' average responses are shown in Table 9. Panelists quantified the influence of each factor using a 3-point scale (1 = *not influential*, 2 = *influential*, 3 = *very influential*). On average, panelists rated most factors somewhere between influential and very influential, but tended to emphasize the importance of the definition of the JQC, the knowledge and skills required to answer each test question, and item-level data (for listening and reading). The factor with the lowest average score across sessions—and thus, considered comparatively less influential by candidates—was "the cut scores of other panel members." These results suggest that panelists were attending to the factors that should be influencing their judgments, with particular emphasis on the definition of the JQC and the knowledge and skills required to answer each test question.

Table 9. The Influence of Factors That Guided Panelists' Standard Setting Judgments

Factors	Listening	Reading	Speaking	Writing
Definition of the just qualified candidate	2.7	2.6	2.9	2.9
Between-round discussions	2.3	2.6	2.7	2.3
Knowledge and skills required to answer each test question	2.8	2.6	2.8	2.7
Cut scores of other panel members	1.9	2.2	2.1	2.0
(Panelists') own professional experience	2.1	2.3	2.7	2.6
Item-level data	2.7	2.6		

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

The final section of the evaluation survey asked panelists to quantify their comfort level with the panel's recommended cut scores. Panelists indicated their comfort using a 4-point scale (1 = *very uncomfortable*, 2 = *uncomfortable*, 3 = *comfortable*, 4 = *very comfortable*). Table 10 summarizes panelists' average responses for each recommended cut score for each test. The average responses are high, between 4 (*very comfortable*) and 3 (*comfortable*) on the scale for all tests and cut scores. None of the panelists in any of the sessions indicated that they were very uncomfortable with any of the recommended cut scores, with the exception of one panelist in the reading session who indicated they were very uncomfortable with the A1, A2, and B1 cut scores. However, this panelist provided strongly positive feedback for all other survey questions, so it is possible that they misread the scale when completing this survey question.

Table 10. Panelists' Comfort Level With the Recommended Cut Scores

Redesigned TOEIC Bridge test	A1	A2	A2+	B1	B1+
Listening	3.4	3.5	3.6	3.4	3.4
Reading	3.5	3.6	3.6	3.4	3.0
Speaking	3.6	3.4	3.7	3.3	3.1
Writing	3.9	3.8	3.7	3.6	3.3

Note. 1 = *very uncomfortable*, 4 = *very comfortable.*

Internal Evidence

Internal evidence addresses issues of consistency: for example, the consistency of judgments between and within panelists and the consistency of judgments between panels (Tannenbaum & Cho, 2014). One common way to evaluate the consistency of panelists'

judgments is to examine their variability between and within rounds. The standard deviation of cut scores for each round, as shown in Tables 2 to 5, is an indicator of the variation in cut scores for each round and is expected to reduce in size across rounds as panelists incorporate feedback from group discussions into their ratings. This general pattern occured in all four sessions. The sessions differed in terms of the variation that was observed in judgments during the initial rounds; for example, the first round of the listening session had relatively large standard deviations for cut scores (4.6 to 8.5) while the first round of the speaking session had much lower standard deviations (1.5 to 1.7). These differences could be due to differences in standard setting approaches, but regardless, the small standard deviations reported in Round 3 across sessions provides evidence that panelists' judgments were consistent or in agreement with one another.

The standard error of judgment provides an estimate of the extent to which the panel's recommended cut scores would be replicated by a different panel (Tannenbaum & Cho, 2014). Again, this estimate should be relatively small and is expected to decrease across rounds of judgment. The results reported in Tables 2 to 5 conform to these expectations, and all standard errors of judgment reported for Round 3 judgments were less than 1 raw score point. These results suggest that the recommended cut scores would be similar if a new panel with similar characteristics were to replicate the study.

External Evidence

External evidence is used to evaluate whether independent sources of information align with the conclusions of standard setting (Council of Europe, 2009). In a preliminary validity study, Schmidgall (2020) collected test takers' self-assessments with respect to various "can-do" statements for each of the four TOEIC Bridge tests. Many of the can-do statements corresponded directly to CEFR descriptors, and the results were summarized in tables that showed the percentage of test takers at each CEFR level (or TOEIC Bridge proficiency level) that believed they could perform each task. For example, TOEIC Bridge Speaking test takers were asked a can-do statement associated with CEFR proficiency level A2 (Council of Europe, 2018, p. 85) if they could "handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself." Based on their TOEIC Bridge Speaking test

scaled score, 32% of Pre-A1 test takers reported they could perform the task, as did 39% of A1 test takers, 56% of A2 test takers, and 76% of B1 test takers (see Schmidgall, 2020, p. 6). Thus, a majority of test takers categorized by the TOEIC Bridge test as CEFR A2 (or above) believed they could perform the task associated with CEFR proficiency level A2, while a majority of test takers categorized at lower levels (A1, Pre-A1) did not. Although this perfect alignment between can-do statement and TOEIC Bridge–based CEFR level classification did not occur for every can-do statement, the results generally followed this pattern and provide initial external validation evidence.

Conclusion

This report described the process used to establish a claim about the relationship between the redesigned TOEIC Bridge tests and CEFR proficiency levels Pre-A1, A1, A2, and B1. The process was guided by expert recommendations for mapping test scores to proficiency levels (Tannenbaum & Cho, 2014), as well as the specific process recommended for mapping tests to CEFR levels (Council of Europe, 2009). The process included four stages: familiarization, specification, standard setting, and validation. The documentation of the familiarization stage established how the various stakeholders involved in the process developed and applied their knowledge of the CEFR. The documentation of the specification stage described the content and measurement quality of the TOEIC Bridge tests, as well as the construct congruence between the tests and the CEFR. The description of the standard setting study detailed how an expert panel was convened and trained to produce recommended cut scores, as well as the poststudy adjustments made to finalize the claim about the relationship between TOEIC Bridge test scaled scores and CEFR proficiency levels Pre-A1, A1, A2, and B1. Finally, the documentation for the validation stage summarized the procedural, internal, and external evidence that support this claim.

References

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford.

Baron, P., & Papageorgiou, S. (2014). *Mapping the TOEFL Primary test onto the Common European Framework of Reference* (Research Memorandum No. RM-14-05). ETS.

Chapelle, C. A. (2013). Reliability in language assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell. https://doi.org/10.1002/9781405198431.wbeal1003

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing 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: A manual*.

Council of Europe. (2018). *Companion volume with new descriptors*. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Deygers, B., Zeidler, B., Vilcu, D., & Hamnes Carlsen, C. (2018). One framework to unite them all? Use of the CEFR in European university entrance policies. *Language Assessment Quarterly, 15*(1), 3–15. https://doi.org/10.1080/15434303.2016.1261350

ETS. (2019). *TOEIC Bridge Listening and Reading tests: Score user guide*.

Everson, P., Duke, T., Garcia Gomez, P., Carter Grissom, E., Park, E., & Schmidgall, J. (2019). *Development of the redesigned TOEIC Bridge tests* (Research Memorandum No. RM-19-10). ETS.

Figueras, N. (2012). The impact of the CEFR. *ELT Journal, 66*(4), 477–485. https://doi.org/10.1093/elt/ccs037

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). Praeger.

Hambleton, R. K., Pitoniak, M. J., & Copella, J. M. (2014). Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 47–76). Routledge.

Hudson, T. (2013). Standards-based testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 479–494). Routledge.

Lin, P., Cid, J., & Zhang, J. (2019). *Field study statistical analysis for the redesigned TOEIC Bridge tests* (Research Memorandum No. RM-19-09). ETS.

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications on test scores. *Journal on Educational Measurement*, *32*(2), 179–197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

O'Sullivan, B. (2010). The City & Guilds Communicator examination linking project: A brief overview with reflections on the process. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manua*l (pp. 33–49). Cambridge University Press.

Papageorgiou, S. (2010). Linking international examinations to the CEFR: The Trinity College London experience. In W. Martyniuk (Ed.), *Aligning tests with the CEFR: Reflections on using the Council of Europe's draft manual* (pp. 145–158). Cambridge University Press.

Papageorgiou, S., Morgan, R., & Becker, V. (2015). Enhancing the interpretability of the overall results of an international test of English-language proficiency. *International Journal of Testing*, *15*(4), 310–336. https://doi.org/10.1080/15305058.2015.1078335

Papageorgiou, S., Tannenbaum, R. J., Bridgeman, B., & Cho, Y. (2015). *The association between TOEFL iBT test scores and the Common European Framework of Reference (CEFR) levels* (Research Memorandum No. RM-15-06). Princeton, NJ: ETS.

Papageorgiou, S., Wu, S., Hsieh, C.-N., Tannenbaum, R. J., & Cheng, M. (2019). *Mapping the TOEFL iBT test scores to China's Standards of English Language Ability: Implications for score interpretation and use* (Research Report No. TOEFL-RR-89). ETS. https://doi.org/10.1002/ets2.12281

Papageorgiou, S., Xi, X., Morgan, R., & So, Y. (2015). Developing and validating band levels and descriptors for reporting overall examinee performance. *Language Assessment Quarterly, 12*(2), 153–177. https://doi.org/10.1080/15434303.2015.1008480

Plake, B. S., & Cizek, G. J. (2012). Variations on a theme: The modified Angoff, extended Angoff, and yes/no standard setting methods. In G. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (pp. 181–199). Routledge.

Powers, D., Schedl, M., & Papageorgiou, S. (2016). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing, 34*(2), 175–195. https://doi.org/10.1177/0265532215623582

Runnels, J., & Runnels, V. (2019). Impact of the Common European Framework of Reference–A bibliometric analysis of research from 1990–2017. *CEFR Journal–Research and Practice, 1*, 18–32.

Schmidgall, J. (2020). *The redesigned TOEIC Bridge tests: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-20-07). ETS. https://doi.org/10.1002/ets2.12288

Schmidgall, J., Oliveri, M. E., Duke, T., & Carter Grissom, E. (2019). *Justifying the construct definition for a new language proficiency assessment: The redesigned TOEIC Bridge tests – Framework paper* (Research Report No. RR-19-30). ETS. https://doi.org/10.1002/ets2.12267

Tannenbaum, R. J., & Baron, P. A. (2015). *Mapping TOEIC scores to the Vietnamese National Standard: A study to recommend English language requirements for admissions and graduation from Vietnamese universities* (Research Memorandum No. RM-15-08). ETS.

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, 11*(3), 233–249. https://doi.org/10.1080/15434303.2013.869815

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 iBT Research Report No. 6). ETS. https://doi.org/10.1002/j.2333-8504.2008.tb02120.x

Young, M. J., & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment* (CSE Technical Report No. 475). Center for Research on Evaluation, Standards and Student Testing, University of California, Los Angeles.

Zeidler, B. (2016). Getting to know the minimally competent person. In C. Docherty & F. Barker (Eds.), *Language assessment for multilingualism: Proceedings of the ALTE Paris conference, April 2014* (pp. 251–269). Cambridge University Press.

Zieky, M. J., Perie, M., & Livingston, S. A. (2008). *Cutscores: A manual for setting standards of performance on educational and occupational tests*. ETS.

Appendix A. Panelists' Assignments to Sessions

Panelist	Session			
	Listening	Reading	Speaking	Writing
1	✓			
2	✓		✓	✓
3			✓	
4	✓		✓	
5	✓	✓		
6	✓	✓		
7	✓			
8			✓	✓
9	✓	✓		✓
10	✓	✓	✓	✓
11		✓	✓	✓
12	✓	✓		
13				✓
14			✓	✓
15		✓	✓	✓
16			✓	✓
17	✓	✓		
18		✓	✓	
19	✓	✓		✓
20	✓	✓		✓
21			✓	✓
22			✓	
23	✓	✓	✓	✓
24		✓		✓
25	✓		✓	
26			✓	✓
27	✓	✓		
Total	15	14	15	15

Appendix B. Panelists' Just Qualified Candidate (JQC) Descriptors for Listening Comprehension

CEFR Level B1

- Can understand multiple main points/topics beyond the sentence-level in extended speech around the listener
- Can understand some important details when explicitly stated (e.g., instructions, technical information, agreement/disagreement)
- Can understand clear and relatively slow, standard speech
- Can understand public announcements with minimum interference from background noise
- Can understand familiar/straightforward topics, but with new (not personal, has schema but not that particular information) information

CEFR Level A2

- Can understand sentence-level discourse
- Can understand slow, articulated clear speech
- Can understand outline, essential information, main point in short, simple exchanges/messages/ monologues
- Can follow simple, routine instructions beyond the listener's immediate environment
- Can understand high frequency words and phrases (vocabulary)

CEFR Level A1

- Can recognize high frequency words, short phrases, and formulaic expressions when delivered slowly and clearly with pauses and repetition (speed)
- Can recognize high frequency words, short phrases, and formulaic expressions with minimal reliance on visual and nonverbal cues (channel)

- Can recognize high frequency words, short phrases, and formulaic expressions about familiar routines and everyday contexts (immediacy of context/topic)

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

CEFR Level B1

- Can understand longer (multi-sentence), straightforward, familiar texts.
- Can understand topics that are unfamiliar as long as the information is direct and explicit.
- Can follow the plot of simple stories/comics with a linear, clear storyline
- Can move beyond high-frequency vocabulary to sometimes infer meaning of unfamiliar words from context or use a dictionary
- Can identify salient details within a text
- Can understand descriptions of feelings, events, and places within straightforward, simply written articles and guides
- Can recognize simple discourse markers (all of a sudden, therefore, however, conjunctions) to connect ideas
- Can understand straightforward, simply written text related to his/her profession

CEFR Level A2

- Can understand short, simple texts (sentences/simple discourse as opposed to just words and phrases) on familiar topics
- Can understand short, simple personal letters, e-mails, and narratives
- Can understand concrete texts, can find predictable information
- Can locate specific information in straightforward phrases in signs, instructions, (bulleted) lists, menus, etc.
- Can understand short, straightforward texts with high frequency words, with or without visual support

- Can understand some main points in short, descriptive texts with simple, predictable language
- Can apply basic grammatical knowledge (tenses, agreement, plurals, etc.)

CEFR Level A1

- Can understand short, connected texts if supported with graphics (illustrated stories/narratives)
- Can understand very high frequency words and short phrases, especially if there is visual (and/or telegraphic) support provided
- Can understand nonlinear texts reflecting everyday life/situations (e.g., basic instructions) when supported by illustrations in a predictable format (e.g., floor maps, timetables (simple schedules), menus, labels, pamphlet, how-to guide)
- Can recognize familiar/famous very basic phrases, names, dates, etc. in everyday situations

Appendix D. Panelists' Just Qualified Candidate (JQC) Descriptors for Speaking

CEFR Level B1

- Can minimally manage (initiate, participate, close) a conversation in both routine and nonroutine situations
- Can tell/retell a story by connecting and sequencing events, with some errors
- Can begin to use a range of language functions (e.g., make a complaint, offer advice, compare and contrast alternatives)
- Can begin to provide minimal reasons and simple justifications for opinions and advice
- Can pronounce words and phrases in a generally clear manner, requiring minimal listener effort
- Can use sufficient vocabulary to support some limited discussion of topics like work, travel, activities, events

CEFR Level A2

- Can provide simple information about people, places, things, and events beyond the self, present time, and immediate environment
- Can ask and answer simple questions (e.g., where, when) to engage in short, simple transactions
- Can use simple, high-frequency vocabulary to identify and describe for familiar, everyday events
- Can use short, basic sentence patterns (e.g., SVO) with the most basic connectors (e.g. first, then; and, but) using simple tenses and aspects with frequent errors
- Can pronounce familiar words and formulaic phrases clearly (with some proper stress and intonation), though overall production requires some listener effort
- Can state a preference (e.g., likes, dislikes) without elaboration

CEFR Level A1

- Can produce simple information about familiar people and places in concrete situations
- Can describe simple aspects about everyday things with some advance preparation
- Can make and respond in a limited way to simple requests in familiar contexts
- Can produce a limited repertoire of high-frequency words and phrases relevant to familiar and routine events (e.g., time, numbers, dates, prices, days of the week)
- Can state a preference when addressed clearly and slowly
- Can produce only short, mainly formulaic utterances with frequent pausing and some routine errors
- Can pronounce simple words and phrases; overall, requires significant listener effort to understand

Appendix E. Panelists' Just Qualified Candidate (JQC) Descriptors for Writing

CEFR Level B1

- Can tell a simple story
- Can write simple descriptions of real or imagined events and things outside of the present
- Can write straightforward, connected texts on a range of topics of interest
- Can ask for or give simple clarification
- Can use a range of vocabulary (i.e., not just high-frequency) to respond to various tasks; may use strategies to compensate for limited vocabulary and structures
- Can express preference/opinion and support it using basic vocabulary with limited elaboration
- Can use a limited range of grammatical structures in a nonformulaic manner with occasional errors

CEFR Level A2

- Can present information in a limited logical sequence using simple connectors with simple phrases and sentences
- Can use a limited range of grammatical structures with some errors that might obscure meaning
- Can use high-frequency vocabulary to appropriately respond to a task, although a response to a task may be incomplete with meaning partially obscured
- Can begin to adjust writing style/register appropriately according to the purpose of the task
- Can convey personal or familiar information (e.g., short notes expressing thanks or apology)
- Can express preference/opinion using basic vocabulary without elaboration

CEFR Level A1

- Can begin to construct isolated phrases and short formulaic sentences using simple words and basic expressions, but with systematic errors
- Can write short, simple messages using isolated phrases conveying information of a personal nature (e.g., family, likes/dislikes)
- Can use common, high-frequency formulaic expressions with minor errors that don't obscure meaning
- Can write basic phrases describing familiar, everyday objects
- Can begin to express basic ideas in more than one sentence, with frequent errors that often obscure meaning
- Can use connector "and"