ETS TOEIC®

A Compendium of Studies

VOLUME IV

The Research Foundation for the Redesigned TOEIC Bridge® TESTS

Jonathan Schmidgall, Editor

TOEIC® COMPENDIUM OF STUDIES: VOLUME IV

Foreword 0.2

Ida Lawrence

Preface 0.3

Jonathan Schmidgall

Section I: Developing the Redesigned *TOEIC Bridge*® Tests

Justifying the Construct Definition for a New Language Proficiency Assessment: The Redesigned *TOEIC Bridge*® Tests—Framework Paper 1.1

Jonathan Schmidgall, Maria Elena Oliveri, Trina Duke, and Elizabeth Carter Grissom

Development of the Redesigned *TOEIC Bridge*® Tests 2.1

Philip Everson, Trina Duke, Pablo Garcia Gomez, Elizabeth Carter Grissom, Elizabeth Park, and Jonathan Schmidgall

Field Study Statistical Analysis for the Redesigned *TOEIC Bridge*® Tests 3.1

Peng Lin, Jaime Cid, and Jiayue Zhang

Section II: Accumulating Evidence to Support Claims

Mapping the Redesigned *TOEIC Bridge*® Test Scores to Proficiency Levels of the Common European Framework of Reference for Languages 4.1

Jonathan Schmidgall

The Redesigned *TOEIC Bridge*® Tests: Relations to Test-Taker Perceptions of Proficiency in English 5.1

Jonathan Schmidgall

Making the Case for the Quality and Use of a New Language Proficiency Assessment: Validity Argument for the Redesigned *TOEIC Bridge*® Tests 6.1

Jonathan Schmidgall, Jaime Cid, Elizabeth Carter Grissom, and Lucy Li

FOREWORD

Over the years, English has become the global language of communication. Organizations around the world have come to recognize that English-language proficiency is a key to competitiveness. For more than 40 years, the *TOEIC*® testing program has provided assessments that enable corporations, government agencies, and educational institutions throughout the world to evaluate a person's ability to communicate in English in the workplace. Millions of TOEIC tests are administered annually for more than 14,000 organizations across more than 160 countries.

ETS is proud of the substantial research base that supports all of the assessments we offer. Research guides us not only as we develop new products, services, tools, and learning solutions, but also as we continually improve existing ones, including those in the TOEIC program (e.g., the *TOEIC Bridge*® tests, the TOEIC Listening and Reading test, and the TOEIC Speaking and Writing tests). Offerings like these are essential to meeting our overall mission—to advance quality and equity in education for people worldwide.

This fourth TOEIC program compendium is a compilation of selected work conducted by ETS Research & Development staff since the third compendium was published in 2018. The focus of this research is making certain that TOEIC tests and test scores remain not only reliable, fair, and valid, but also meaningful, useful, and responsive to the needs of organizations.

We hope you find this compendium to be valuable. As with the previous compendia, we welcome your comments and suggestions.

Ida Lawrence

Senior Vice President
Research & Development Division
ETS

PREFACE

This is the fourth volume in the *TOEIC*® Program Compendium series, which focuses on the research foundation for TOEIC assessments. The first volume was published in 2010 and focused on the redesigned TOEIC Listening and Reading test and the newly developed TOEIC Speaking and Writing tests. The second and third volumes were published in 2013 and 2018, respectively, and covered a variety of topics related to the TOEIC and *TOEIC Bridge*® tests, including the refinement of the TOEIC Listening and Reading, Speaking, and Writing tests. The themes explored across these volumes, and also framing the current volume, include refinement, revision, renewal; monitoring and controlling quality; and accumulating evidence to support claims about test use. The first theme—refinement, revision, renewal—is explored in chapters describing how the design of TOEIC tests is periodically revisited to continue to meet the needs of stakeholders. The second theme reflects the importance of monitoring and empirically investigating the measurement quality of the test, or issues related to reliability, validity, and fairness. The third theme builds upon the second to support the use of test scores to make decisions and to evaluate claims about the intended consequences of *TOEIC* test use and of decisions based on test scores.

This volume in the series differs from previous volumes in that it is entirely focused on the redesigned TOEIC Bridge tests, intended to measure basic to intermediate English proficiency in everyday life and common workplace scenarios. In early 2017, a team of ETS researchers, psychometricians, and test developers began meeting with TOEIC program staff to revisit the design of the TOEIC Bridge test. Based on input from key stakeholders, the TOEIC program established a mandate for a redesigned four-skills (listening, reading, speaking, and writing) TOEIC Bridge assessment. Over the course of the next several years, the research team conceptualized the redesigned assessment, developed new items and tests, and conducted preliminary research to support the operational launch of the tests.

This volume is organized into two main sections, echoing the major themes of the TOEIC Program Compendium series. The first section, "Developing the Redesigned TOEIC Bridge Tests," includes a collection of three chapters that describe the full scope of the test development process. This process utilized an evidence-centered design methodology, a rigorous and systematic approach to test design that is further described in relevant chapters.

The first chapter, the test framework paper, describes the first step of the test development process: establishing a definition of the language knowledge, skills, and abilities that would be evaluated by the redesigned (listening, reading) or new (speaking, writing) tests. This process began by translating the mandate for test design into a theory of action, or visual depiction of how components of an assessment should be used to make decisions to facilitate specific outcomes. This theory of action informed a domain analysis, which explored relevant theoretical and empirical research to document the rationale for how English listening, reading, speaking, and writing ability for everyday adult life would be defined for the purpose of assessment.

The second chapter continues the narrative of test development by describing how definitions of ability drove the development of prototype test tasks and test forms. As this chapter shows, there was an explicit link between the targeted definitions of ability and test tasks throughout the development process. The chapter also describes how performance data, input from test takers, and input from raters contributed to the design process throughout, from the pilot study to the field test.

The third chapter in this volume concludes the test development narrative by summarizing the results of a field study that was used to evaluate the statistical properties of the tests. The chapter describes how the field study was conducted and summarizes the results of analyses that have implications for claims about the measurement quality of the tests.

The second main section, "Accumulating Evidence to Support Claims," includes two chapters that describe research conducted to investigate and elaborate the meaning of test scores and a final chapter that synthesizes the evidence presented throughout this volume into a coherent narrative about the quality of the assessment and its intended use.

In the fourth chapter, the process used to map redesigned *TOEIC Bridge* test scores to Common European Framework of Reference for Languages (CEFR) levels is described. As detailed in the chapter, the process was comprehensive and multifaceted, adhering to best practices in educational measurement for mapping test scores to standards while closely following the Council of Europe's manual for relating examinations to the CEFR.

The fifth chapter details a study in which redesigned TOEIC Bridge test scores were compared to an external criterion of test takers' language abilities: their self-assessments of the extent to which they can perform various language tasks. The results of this study provide validity evidence and help expand the meaning of test scores by further elaborating the types of language activities test takers probably can (or cannot) do at different proficiency levels.

Finally, the sixth chapter describes how the main claims in a "validity argument" communicate a narrative about the qualities that make a test useful, and it elaborates an initial validity argument for the redesigned TOEIC Bridge tests. This validity argument includes claims about the measurement quality of test scores (i.e., their consistency or reliability) and score interpretations (i.e., their meaningfulness, impartiality, and generalizability), as well as the intended uses of the tests.

This volume was produced for two audiences. First and foremost, it is for those interested in or impacted by the design, quality, and intended uses of the redesigned TOEIC Bridge tests: key stakeholders such as test takers, score users, and teachers. This volume also illustrates a test development and research program that is rigorous yet practical, which may interest students, researchers, and practitioners in language assessment.

Jonathan Schmidgall

SECTION I: DEVELOPING THE REDESIGNED *TOEIC BRIDGE*® TESTS

JUSTIFYING THE CONSTRUCT DEFINITION FOR A NEW LANGUAGE PROFICIENCY ASSESSMENT: THE REDESIGNED *TOEIC BRIDGE*® TESTS—FRAMEWORK PAPER

Jonathan Schmidgall, Maria Elena Oliveri, Trina Duke, and Elizabeth Carter Grissom

BACKGROUND

In this framework paper, we describe the purpose of the redesigned *TOEIC Bridge*® tests and justification of their construct definitions. In doing so, we elaborate the rationale for the interpretation and use of test scores. This is a foundational step in the test design process that provides the basis for initial assumptions about the meaning of test scores and serves as a reference for subsequent validity research (American Educational Research Association et al., 2014; Bachman & Palmer, 2010).

We begin with a discussion of the purpose and intended uses of the assessment and key stakeholder groups and propose a logic model that outlines the relationships among assessment components, intended uses, and intended outcomes. This forms the basis of a mandate for test design. It also establishes connections among test purpose, test design, and validation (Fulcher, 2013).

We contextualize the rest of the framework paper within an evidence-centered design (ECD) approach to test design and development (Mislevy et al., 2003). Although the ECD approach consists of five layers of analysis, the framework paper focuses primarily on the first layer, domain analysis.

Our approach to domain analysis reflects an interactionalist approach to construct definition, in which context and abilities interact to form the construct (Bachman, 2007). Thus, we begin by elaborating a clearer definition of our language use domain, "everyday adult life." Next, we survey research literature and relevant developmental proficiency standards to highlight the knowledge, skills, and abilities relevant to beginner to low-intermediate general English proficiency. This information is synthesized in our definitions of the constructs of reading, listening, speaking, and writing ability for beginner to low-intermediate levels of general English proficiency in the context of everyday adult life.

Test Purpose and Intended Uses

The redesigned TOEIC Bridge tests measure beginning to low-intermediate English language proficiency in the context of everyday adult life. In order to accommodate the particular needs of score users, the redesigned TOEIC Bridge tests include modules for listening and reading, speaking, and writing. If score users are interested in an evaluation of overall language proficiency or communicative competence, all four skills should be tested.

The tests are primarily intended to be used for selection, placement, and readiness purposes. Some score users may wish to use the test to determine whether applicants to vocational or training institutions have

a threshold level of English proficiency that is needed or desirable (i.e., selection) to benefit from further English language training. Other score users may use information about English proficiency for the purpose of placing students or employees into English language training courses or programs of study at beginner to low-intermediate proficiency levels. Additionally, some score users (i.e., test takers) may wish to use the information obtained about their English proficiency to determine their readiness to take *TOEIC*® tests or for more advanced study.

Several secondary uses of the test were also considered in the design of the test. Some score users may want to use test section scores to track or benchmark development or improvement over time in order to monitor growth in language skills or overall proficiency. Others may wish to use subscores or other performance feedback in order to identify their relative strengths and weaknesses with respect to different language skills.

Stakeholders

The stakeholders of a test are those who are either directly affected (primary stakeholders) or indirectly affected (secondary stakeholders) by the use of the test (Bachman & Palmer, 2010). Those directly affected—primary stakeholders—are the individuals whose proficiency is being evaluated (test takers) and those who use the scores to make important decisions (score users, including teachers). Those indirectly affected—secondary stakeholders—are the individuals who may have a stake in the use of the test due to its impact on their work or experience (e.g., teachers who are not necessarily score users).

Test takers are young adults (high school/secondary school and older) and adults for whom English is a second or foreign language, and their nationalities and native languages (L1) will vary. Test takers' educational backgrounds and purpose for learning English (e.g., general purposes, academic purposes, occupational purposes) may also vary. Score users will typically be administrators (e.g., at vocational training institutions) and managers (e.g., at organizations and institutions). Teachers may be primary or secondary stakeholders and will be affected if the redesigned TOEIC Bridge tests are used for placement into language training courses. Teachers may also benefit from the use of the test to track proficiency and potentially monitor progress and the use of any information provided by the test to inform remedial instruction.

A Logic Model for Redesigned TOEIC Bridge Tests

Ultimately, tests are used to promote particular outcomes, effects, or consequences. With this in mind, intended outcomes should be elaborated from the beginning of a test design project and inform the design of the test itself (Bachman & Palmer, 2010; Norris, 2013). Bachman and Palmer (2010) advanced this view through the use of an argument-based approach to test use, which begins with test developers consulting with score users to establish claims about desirable outcomes (e.g., hiring employees with appropriate English language skills). Test developers then work backward to determine the types of decisions that facilitate the intended outcomes (e.g., a selection decision), the interpretations about abilities needed to facilitate equitable decisions, the scores that are needed to facilitate meaningful and impartial interpretations, and finally, characteristics of test performances needed to produce scores that are reliable or consistent.

Another approach that establishes a link between test components, intended uses, and outcomes is the theory of action (Bennett, 2010; Patton, 2002, pp. 162–164). The theory of action uses a logic model to illustrate how components of the test (such as scores) are expected to facilitate particular actions (i.e., decisions), which in turn are intended to produce particular effects (i.e., outcomes or consequences). In the logic model, arrows indicate hypothesized causal links: For example, an arrow between test components and a particular action mechanism implies a claim about the relevance of the test for a particular use. When fully developed, the logic model is expanded to a theory of action by providing documentation that explicitly states each claim and provides a summary of the evidence backing the claim.

As a preliminary step, we specified a logic model for the redesigned TOEIC Bridge tests that reflects their purpose and intended uses (see Figure 1). These uses are formalized in the diagram as hypothesized actions. Each hypothesized action is expected to produce intermediate and ultimate effects. Based on the actions and effects we intend to support and promote, we specified components of the tests that we believe are necessary.

In the logic model, there are three primary hypothesized actions that the test will be designed to support: selection, placement, and determining readiness for TOEIC tests or more advanced study. There are two additional hypothesized actions that the test developer would like to support, identified in dashed boxes in the logic model: monitoring growth or progress and using test information to identify learners' strengths and weaknesses. Several components of the redesigned TOEIC Bridge tests will be necessary to support these actions: test section scores, and mapping or concordance with external standards (e.g., Common European Framework of Reference [CEFR] A1 to B1) and TOEIC tests. We intend these actions to have specific intermediate and ultimate effects.

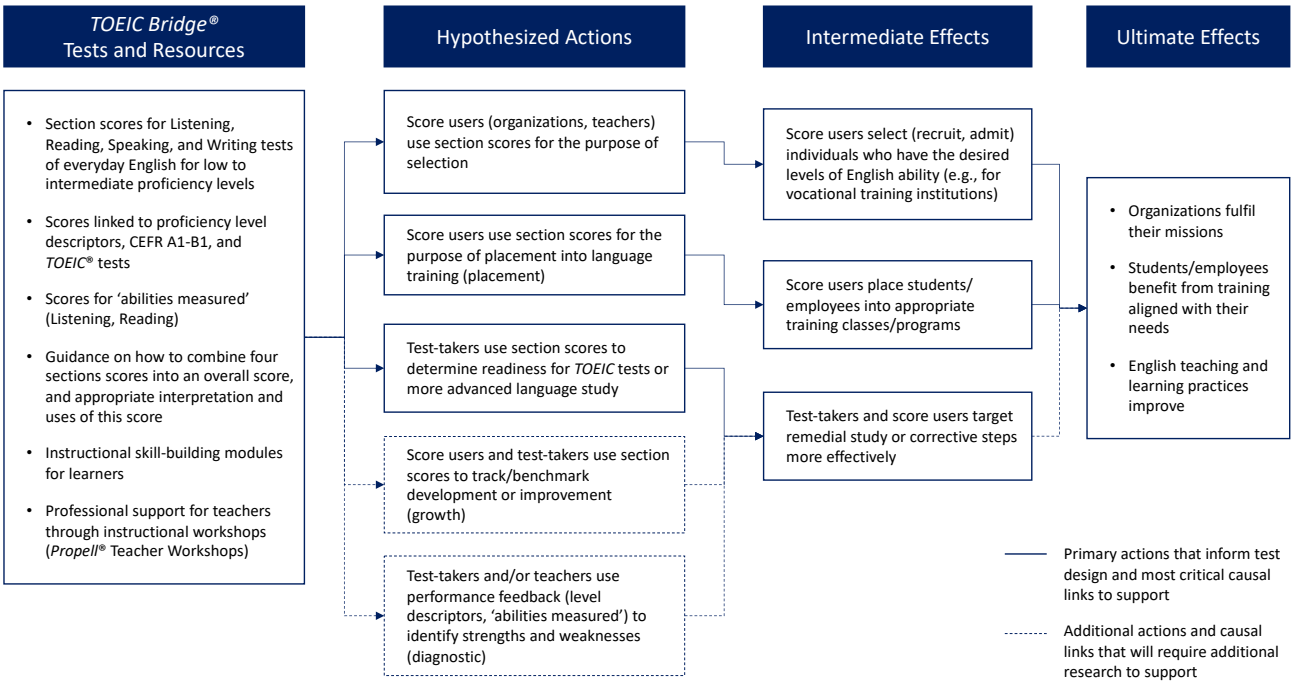

Figure 1. A logic model for the redesigned TOEIC Bridge tests.

EVIDENCE-CENTERED DESIGN AND TEST DEVELOPMENT

With the intended uses, effects, and test components specified in the logic model, we began to conceptualize the design of the test within an ECD framework (Mislevy et al., 2003). ECD is a systematic approach to test design that helps identify, map, and categorize activity patterns associated with a particular context or practice to render test takers' implicit behaviors and attitudes observable and assessable in an operational assessment. Although conceived as a general approach to test design and development, ECD has been utilized by several language assessment programs (Chapelle et al., 2008; Hines, 2010; Kenyon, 2014).

The ECD model has five layers: (a) domain analysis, (b) domain modeling, (c) the conceptual assessment framework (CAF), (d) assessment implementation, and (e) assessment delivery (Mislevy & Yin, 2012). Each layer includes different concepts and entities, representations, purposes, and questions. There is an implied iteration between these layers as developers move back and forth between the layers. Figure 2 illustrates the roles, associated activities, and resulting activity for the first three layers of ECD (Riconscente et al., 2015). The red boxes identify the aspects of the ECD process that are addressed by this framework paper.

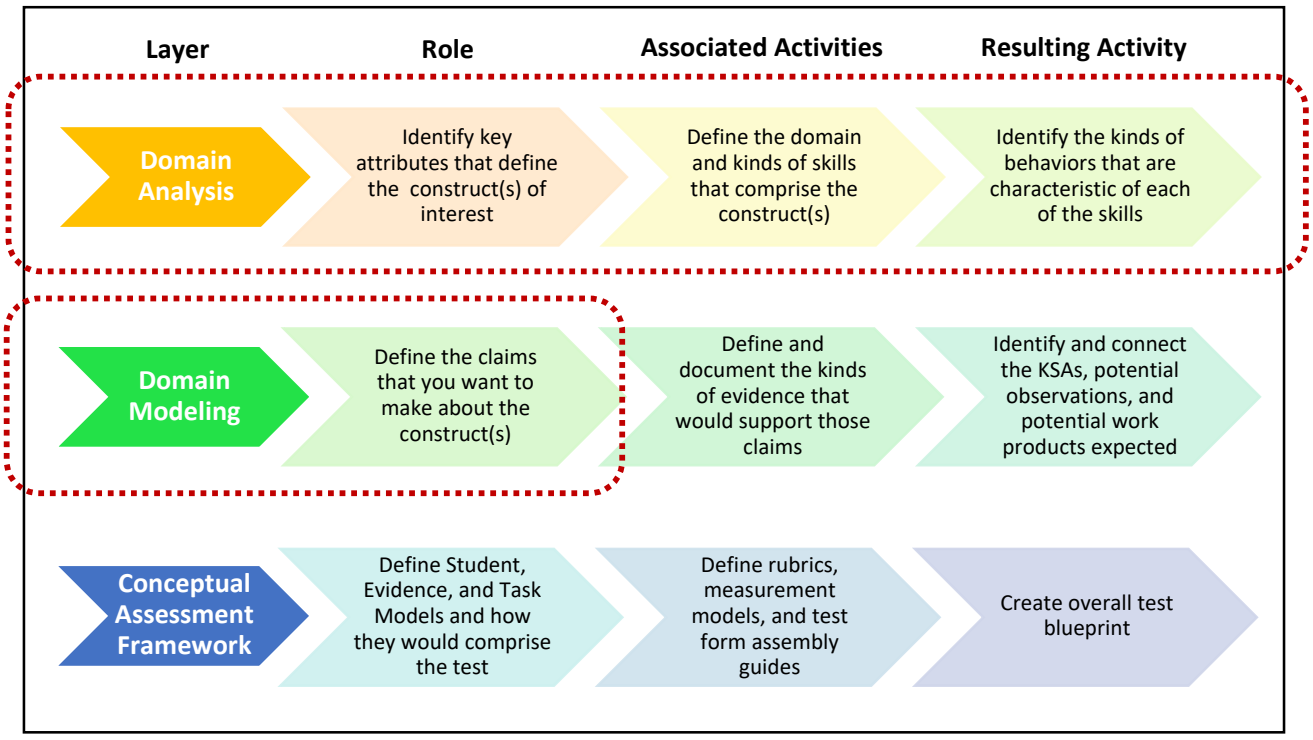

Figure 2. Activities within the first three layers of the evidence-centered design assessment development process, and the focus of the framework paper.

The purpose of the first layer, domain analysis, is to identify the key attributes that define the constructs of interest. In language assessment, construct definition typically entails elaborating ability-in-context (Bachman, 2007): knowledge, skills, and abilities (KSAs) and the target language use (TLU) domain. Activities at this stage of ECD typically include conducting systematic literature reviews of frameworks, taxonomies, and assessments and may include consulting with subject-matter experts and industry-related stakeholders to identify the key features of the construct(s) of interest, the kinds of skills that comprise it, and the kinds of behaviors that characterize each skill.

In the second layer, domain modeling, the information gleaned in the domain analysis is parsed into assessment design patterns (Wei et al., 2008). Design patterns elaborate key attributes of the test, including its rationale, focal KSAs, potential observations, characteristic features, and variable features. They form the initial narrative for the design of the test and the basis for the development of test specifications in subsequent ECD layers.

The third layer of ECD is the CAF, which is used for the assembly of the entire assessment by generating a test blueprint (which should include the desired performances to elicit and work products to capture, the features of tasks or items, and constraints for the development of the assessment). The CAF includes the student, evidence, and task models that specify the elements of an operational assessment design (Mislevy et al., 2003). The student model is conceptualized in terms of the construct, assessment purpose, and the target population(s). The evidence model structures thinking about the kinds of performances (their salient features captured as observable variables) that provide evidence of a test taker's standing on the KSAs as deemed important for the construct. Considerations for how to elicit the desired evidence about the defined construct occur in the task model. These considerations include identifying the types of situations necessary to best elicit behaviors that demonstrate proficiency in the desired KSAs. All of the information from the design patterns is brought together to populate the student, evidence, and task models. The assessment is specified in terms of its content, how it will be delivered, features of the test-taking environment, and test administration instructions. The CAF documents how items/tasks can be varied to create additional test forms. It also documents how test developers update their beliefs about test takers' proficiency based on their work products. In other words, the CAF specifies the operational elements, models, and data structures that instantiate the assessment argument. It structures the data that will be produced and makes sense of them in a way that permits interpretable and meaningful score-based inferences, in accordance with the assessment argument. The CAF also serves another purpose: examining the impact the assessment may have on test takers and different populations. Reviewing the elements of the operational assessment at this stage helps the developer ensure that inferences from the overall performances are appropriate and the construct coverage is adequate.

After the assessment is deployed operationally (see Mislevy & Yin, 2012, for a discussion of the assessment delivery and assessment implementation layers), the ECD-based assessment argument can be extended into an assessment use argument using a formal argument-based approach to validation (e.g., Bachman & Palmer, 2010; Kane, 2011). Evidence collected throughout the ECD process can provide initial backing to support claims about test scores, score interpretations, and test use.

DOMAIN ANALYSIS: CONCEPTUALIZING BEGINNER TO LOW-INTERMEDIATE GENERAL ENGLISH PROFICIENCY FOR EVERYDAY ADULT LIFE

Language proficiency may be conceptualized as ability-in-context or from an interactionalist perspective (e.g., Bachman, 2007; Chalhoub-Deville, 2003; Chapelle, 1998; Xi, 2015). This involves three essential components: the language knowledge required to facilitate performance, communicative strategies to support performance, and a description of the performance context itself. The performance context is often referred to as the TLU domain (Bachman & Palmer, 2010). Once the TLU domain is broadly defined (e.g., everyday adult life), communicative tasks that are typical of the domain are identified and their features are elaborated using a task characteristic framework (e.g., Bachman & Palmer, 2010) or another principled approach to specifying contextual features of tasks (e.g., Xi, 2015). The underlying language knowledge (e.g., lexical knowledge) and processes (e.g., lexical retrieval) needed to successfully perform tasks in the domain form another component. Communicative strategies are often linked to particular tasks and reflect the use of language to achieve a communicative purpose or functional goal and may be articulated more broadly (e.g., reading to find information) or narrowly (e.g., ability to identify essential information in complex sentences in text). Documentation of these components is the product of the domain analysis stage and provides the basis for domain modeling, the next layer in the ECD process.

Figure 3 illustrates how the stages of the domain analysis described in this section were structured to provide the basis for construct definition for redesigned TOEIC Bridge tests.

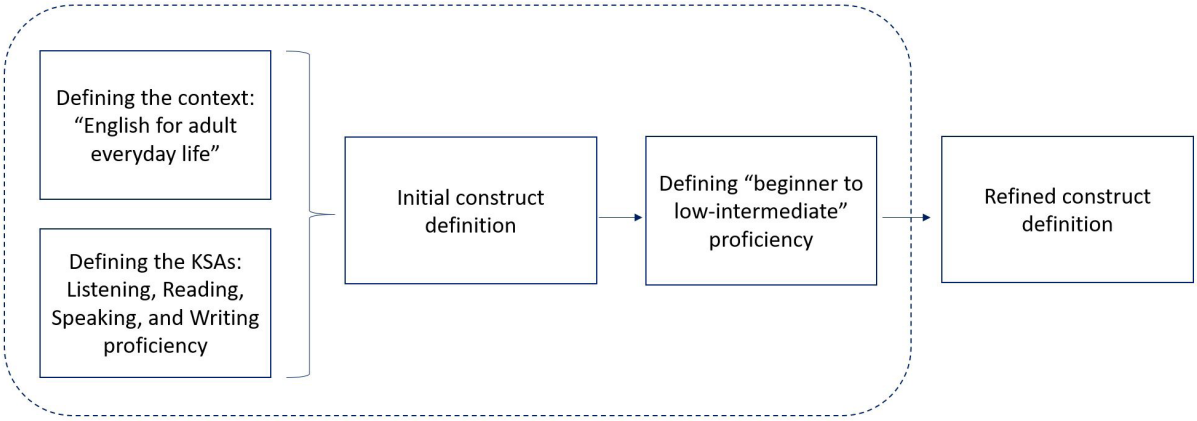

Figure 3. Domain analysis as the basis of construct definition for redesigned TOEIC Bridge tests.

The domain analysis began with a review of literature that may inform the definition of the TLU domain, English for everyday adult life (see the subsection entitled Defining the TLU Domain of Everyday Adult Life). The purpose of this activity was to elaborate the contextual features of the TLU domain (i.e., general features of the setting) relevant to construct definition and test design. We then summarized literature related to the constructs of listening comprehension, reading comprehension, speaking proficiency,

and writing proficiency for second or foreign language (L2) learners in the subsection entitled Defining English Reading, Listening, Speaking, and Writing Proficiency. These initial reviews provided the theoretical basis for construct definition within an interactionalist approach, highlighting relevant abilities and contexts that should be incorporated into construct definitions for redesigned TOEIC Bridge tests. Given the mandate to evaluate proficiency at beginner to low-intermediate levels—and map test-based interpretations about proficiency with levels of the CEFR, Canadian Language Benchmarks (CLB), and American Council on the Teaching of Foreign Languages (ACTFL) language proficiency standards—we then conducted a thorough review of relevant levels of these standards with our definitions of proficiency in mind in the subsection entitled Defining Beginner to Low-Intermediate English Proficiency. This evaluation informed the refined version of the construct definition, presented in the section Construct Definition for an Assessment of Beginning to Low-Intermediate English Language Proficiency for Everyday Life.

Defining the Target Language Use Domain of Everyday Adult Life

Broadly, researchers make a distinction between general and specific-purpose TLU domains (Douglas, 2000). This distinction is made based on the degree to which the TLU domain is concretely and narrowly specified; in other words, the communicative context of a general purposes domain is more varied and resistant to precise description (Douglas, 2001). Although definitions of general and specific-purpose domains are typically based on a theoretical model of language ability or acquisition, the nature of specific-purpose domains facilitates a more detailed analysis of relevant communicative tasks and language abilities.

Although a broad distinction between general and specific-purpose domains can be maintained, it might be helpful to view the specificity of TLU domains as a continuum with general purposes on one end and specific purposes on the other (Knoch & Macqueen, 2016). TLU domains that are more narrowly and concretely defined (e.g., English for aviation) will have a higher degree of specificity than those that are more broadly or abstractly defined (e.g., English for the workplace). When the degree of specificity is high, the language abilities and contextual features relevant to the domain can be more clearly articulated. For more general domains where the degree of specificity is low, researchers or test developers may need to rely on taxonomies to describe features of the TLU domain that should be represented in the assessment procedure to facilitate generalizations about language abilities.

The TLU domain of everyday adult life as conceptualized for the redesigned TOEIC Bridge tests is expected to fall toward the general-purposes end of a specificity continuum. Given this conceptualization of the TLU domain, we considered a number of relevant taxonomies to further elaborate what the general, everyday adult life TLU domain may or may not include. Our review of relevant literature and test documentation identified four approaches that could contribute to the conceptualization of everyday adult life: the social-ecological model of concentric circles, the CEFR for languages standards, the ACTFL proficiency guidelines, and the *TOEFL*® family of assessments. Our initial review of the CLB noted that discussion of the context of language use primarily focuses on differentiating *nondemanding*

(common everyday activities) and *demanding* (educational and work-related) contexts and was generally aligned with the CEFR's approach; consequently, we did not include it in our summary. Below, we briefly summarize relevant information from each of the four approaches reviewed in depth.

Social-Ecological Model of Concentric Circles

One way to consider the TLU domain of everyday life is through the lens of ecological models that specify a set of nested social contexts. The social-ecological model (Bronfenbrenner, 1979) originated as a model of human development and describes an ecological system composed of five socially organized subsystems that support human development. It is conceptualized as a set of concentric circles that are centered on the individual (the microsystem); extends to family, peers, and other intimates (the mesosystem); then to neighbors, extended family, and less intimate others (the exosystem); and beyond that to a context that reflects norms from cultural values, customs, and laws (the macrosystem). Given that this model was conceived in the context of development, changes that occur in individuals or the environments within these subsystems over time are accounted for in a fifth subsystem (the chronosystem). According to Bronfenbrenner (1979), human development occurs through progressively more complex interactions between the individual and the people, objects, and symbols in the individual's environment. These interactions are called *proximal processes*. Together, process, person, and context form the core of the ecological model.

Although originally conceived for general child development, this model has been applied to many other fields, including L2 development. Van Lier (2000) related the model to Vygotsky's sociocultural theory (Vygotsky, 1978). Learners develop by engaging in different learning contexts, or proximal processes, analogous to Vygotsky's zone of proximal development (ZPD). These subsystems may roughly translate to a variety of TLU domains or subdomains, each engaging the learner in a different set of proximal processes. As proficiency increases, test takers are able to interact with the increasingly less familiar, moving from their immediate social network to the broader community or culture and social norms, and from concrete ideas to more abstract concepts.

The Council of Europe Framework of Reference Standards

The CEFR standards describe four broad domains of language use: personal, public, educational, and occupational. The personal domain involves "family relations and individual social practices," whereas the public domain involves "ordinary social interaction (business and administrative bodies, public services, cultural and leisure activities of a public nature, relations with media, etc.)" (CEFR, 2009, p. 15). The educational domain relates to "the learning/training context . . . where the aim is to acquire specific knowledge or skills," and the occupational domain focuses on "a person's activities and relations in the exercise of his or her occupation" (Council of Europe, 2009, p. 15).

In practice, these domains may overlap in various ways. For retail workers, the occupational and public domains may largely overlap. For teachers, the educational and occupational domains may overlap. Even in these cases, however, distinctions between these broad domains may be useful to maintain. For

example, the communicative skills needed by retail workers to interact in the occupational domain differ somewhat from the skills needed to interact in the public domain, given the differences between the roles and responsibilities of employees and customers. The communicative skills required by students in training courses—even teacher training courses—differ somewhat from those required by the teachers of those training courses.

The American Council on the Teaching of Foreign Languages Proficiency Guidelines

The ACTFL® proficiency guidelines (ACTFL, 2012) do not formally define language use domains or subdomains, although they provide descriptions of relevant contexts of language use at each level of proficiency. The notion of "everyday contexts" is elaborated in terms of topics of communication related to survival in the target language culture, such as communicating basic personal information, basic objects, and a limited number of activities, preferences, and immediate needs as well as responding to simple, direct questions or requests for information.

The guidelines note that everyday tasks and communicative functions might be expressed in different forms depending on whether speech or writing is presentational (one-way, noninteractive) or interpersonal (i.e., interactive, two-way communication). For example, for writing, tasks and communicative functions may include lists, short messages, postcards, and simple notes (presentational) or they may include instant messaging, e-mail communication, and texting (interpersonal).

The TOEFL Family of Assessments Approach to Domain Definition

The TOEFL family of assessments includes the *TOEFL iBT®*, *TOEFL ITP®*, *TOEFL Junior®*, and *TOEFL Primary®* assessments. Although these assessments are designed to evaluate English proficiency in the context of English-medium education (i.e., academic TLU domain), their overall approach to conceptualizing the TLU domain was considered for how it may be adapted for our purposes. The TOEFL family of assessments' conceptualization of the academic TLU domain includes subdomains that include social-interpersonal, academic navigational, and academic content (see So et al., 2015). Two of these subdomains—social-interpersonal and academic-navigational—have potential relevance to the domain of everyday language use.

In the TOEFL Junior test, communicating in English for social and interpersonal purposes for adolescents encompasses uses of language for establishing and maintaining personal relationships. For example, students participate in casual conversations with their friends in school settings where they have to both understand other speaker(s) and respond appropriately. Students sometimes exchange personal correspondence with friends or teachers. The topics may include familiar ones, such as family, routine daily activities, and personal experiences. The tasks in this domain tend to involve informal registers of language use.

A second use is communicating for navigational purposes, such as communicating with peers, teachers, and other school staff about school- and course-related materials and activities but not about academic content. For example, students communicate about homework assignments to obtain and clarify details. In some cases, they need to extract key information from school-related announcements. That is, students need to communicate to navigate school or course information. The second subdomain captures this specific purpose of communication.

Although the TLU domain targeted by the TOEFL Junior test pertains to young learners, language activities are generally meaning focused and intended to replicate a variety of real-life communication contexts. Language activities are typically organized around a theme (e.g., my weekend) to allow learners to use learned expressions in a variety of settings relevant to young learners (e.g., plan a weekend with a classmate, survey the class on favorite weekend activities). The language use contexts replicated in the English as a foreign language (EFL) classroom are largely social, meaning that learners primarily use language to communicate with people around them (e.g., family, friends, classmates, teachers) on familiar topics (e.g., myself, animals, people) and to obtain basic information from familiar sources (e.g., stories, announcements, directions).

Summary

Although our review did not identify any formal attempt to define the more general-purpose TLU domain of English for everyday adult life, the approach advocated by the authors of the CEFR standards was useful. Specifically, this approach suggested that language is primarily used in personal and public contexts at lower proficiency levels and branches out into specific-purpose domains (academic or occupational) at intermediate to advanced levels. Given the purpose of the assessment—measuring English proficiency at beginning to low-intermediate levels—personal and public contexts should be well represented in the domain definition for everyday adult life. Figure 4 provides a visual representation of this domain.

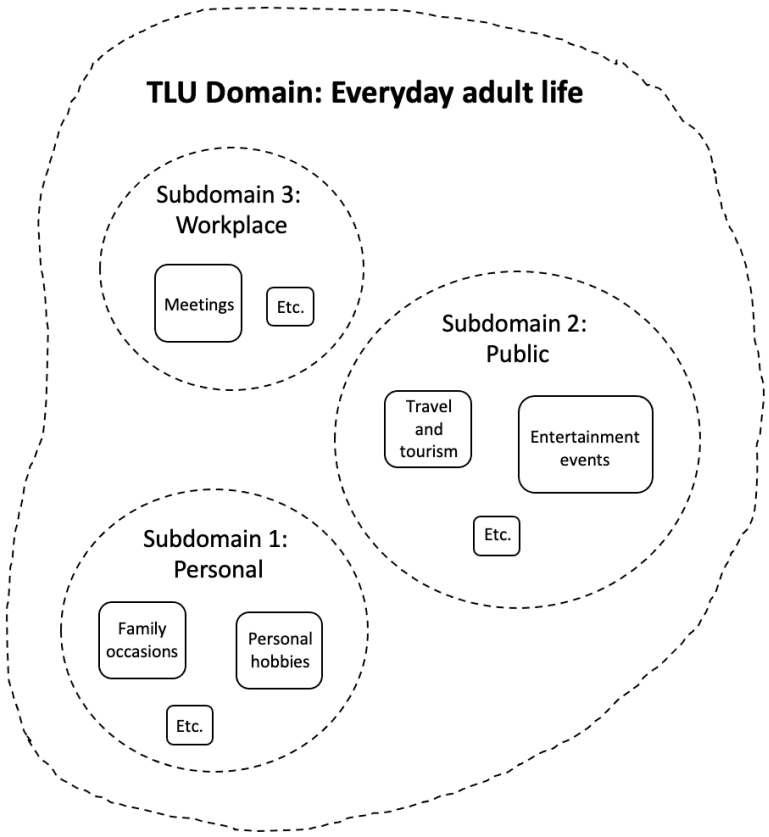

Figure 4. Components of the target language use domain of everyday adult life.

As shown in Figure 4, the TLU domain of everyday adult life is a more general-purpose domain that emphasizes tasks and contexts that are expected to be familiar to adults and young adults. TLU subdomains include personal, public, and some more general and familiar aspects of occupational or workplace contexts. Within each subdomain, there are settings that are expected to be more familiar. An example of a familiar setting within the personal subdomain might be family occasions or settings that relate to personal hobbies and interests. Within the public subdomain, familiar settings may include travel and tourism, entertainment events, and shopping. Only the most general workplace settings (i.e., those that require no industry-specific experience to understand) would be considered relevant to the workplace subdomain.

In addition to specifying the subdomains and settings typical of the TLU domain of everyday adult life, it is important to consider other contextual features of the setting that may need to be represented in language use tasks included in the assessment. In their framework of language task characteristics, Bachman and Palmer (2010) elaborated characteristics of the setting, rubric, input, expected response, and relationship between input and expected response that should be considered when describing or developing language use tasks for the purpose of assessment. Several of these characteristics are worth noting, as the degree to which they are represented in assessment tasks may constrain or facilitate generalization about language proficiency to the TLU domain of everyday adult life. For productive

language use (i.e., speaking and writing), care should be taken to identify the role of the test taker and his or her intended audience in order to simulate the interactional nature of everyday adult life. As researchers have observed, English communication often occurs between L2 users of English who use English as a lingua franca (McNamara, 2011), and so the intended audience in the TLU domain of everyday adult life may include both native (L1) and nonnative (L2) users of English. In addition, the topical characteristics of tasks should reflect the subdomains (personal, public, workplace) included in the TLU domain.

DEFINING ENGLISH READING, LISTENING, SPEAKING, AND WRITING PROFICIENCY

Our review of the components or elements of English language proficiency is influenced by the module-based design of the redesigned TOEIC Bridge tests, as score users may be interested in more selective (e.g., comprehension skills only, speaking ability only) or more comprehensive (i.e., four skills) information about language proficiency for the purpose of decision-making. The conceptualization of overall language ability as multicomponential, consisting of four modalities (reading, listening, speaking, writing) and linguistic elements (e.g., grammar, vocabulary, phonology, sociopragmatics) reflects the view of many researchers (Purpura, 2004).

Listening comprehension has generally been conceptualized through the use of cognitive processing models of listening comprehension or component models of listening ability. Cognitive processing models attempt to identify the phases of cognitive processing and resources involved between the reception of an acoustic (and potentially visual) signal and a listener's response (e.g., Bejar et al., 2000; Field, 2013; Rost, 2005). Component models of listening ability are ontological representations that are influenced by models of communicative competence (e.g., Bachman, 1990), and typically include the higher order components of language competence and strategic competence (Buck, 2001; Weir, 2005). In Buck's model—largely based on Bachman and Palmer's (1996) framework of communicative competence—language competence consists of declarative and procedural knowledge related to listening, including grammatical, discourse, pragmatic, and sociolinguistic knowledge. Strategic competence includes cognitive and metacognitive strategies that are related to listening.

Reading comprehension is typically conceptualized as the process or product of a reader's interaction with a text (Alderson, 2000; Koda, 2013). The process-oriented view conceptualizes reading comprehension as the process of a reader interacting with a text, while the product-oriented view focuses on the product of this interaction, typically demonstrated by answering comprehension questions that require readers to recall the product (i.e., the aspect of comprehension elicited by the question) from memory (Koda, 2013). In both views, the reader may comprehend a text to a greater or lesser degree, and the notion of the processing demands of comprehension questions may be useful to differentiate the degree of overall comprehension. In a comprehensive review of the construct of reading comprehension, researchers have also elaborated the "reader purpose" view, largely complementary to both processing and product perspectives (Jamieson et al., 2000). In the reader purpose view, reading comprehension is conceptualized as the interaction between reader linguistic and processing abilities,

reader purpose, and text characteristics for a given reading task. Thus, a construct definition for reading should elaborate the range of relevant skills and strategies needed by the reader given the purposes and text characteristics involved for the targeted reading tasks. In a meta-analysis of the relationship between reading component variables and passage-level reading comprehension, Jeon and Yamashita (2014) found that L2 grammar knowledge, L2 vocabulary knowledge, and L2 decoding were the strongest predictors of L2 reading comprehension.

Models of speaking proficiency—much like those of listening comprehension—can be characterized as cognitive processing or component models. Cognitive processing models of speech production typically involve four phases: conceptualization, formulation, implementation, and self-monitoring. One of the best known models in second language speech production is Levelt's (1989) parallel model, which hypothesizes knowledge bases that inform the conceptualization phase (e.g., discourse and background knowledge) and formulation phase (lexical, grammatical, and phonological knowledge). Component models of speaking ability have been much more widely utilized in language assessment and generally correspond to models of communicative competence such as that of Bachman and Palmer (2010), who suggested that language ability reflects an interaction between strategic competencies and language knowledge. In Bachman and Palmer's model, language knowledge is composed of organizational knowledge (grammatical and textual knowledge) and pragmatic knowledge (functional and sociolinguistic knowledge), and strategic competencies involve goal setting, appraising, and planning. Some researchers have attempted to refine component models that are perceived to lack sensitivity to contextual features by emphasizing the importance of pragmatic competencies (e.g., Purpura, 2004), or greater representation of contextual facets of tasks in the construct definition (e.g., Xi, 2015).

Writing proficiency is often broadly conceptualized using process-oriented cognitive models (Weigle, 2002), and it is considered in second or foreign language contexts using task-based approaches that elaborate important features of writing tasks such as subject matter, discourse mode (genre, audience, purpose), and stimulus materials (Weigle, 2013). In this task-oriented approach, writing ability is essentially defined by the ability to produce written texts in accordance with the purpose of the task (e.g., to inform, to persuade), follow conventions of the genre (e.g., explanatory writing, transactional writing), and consider the needs of the intended audience (e.g., laypersons, academic specialists). The underlying linguistic knowledge and resources needed to demonstrate writing ability may vary by domain, task, and proficiency levels but often refer to elements such as content, organization, vocabulary use, mechanics, and grammar (e.g., Jacobs et al., 1981; Weir, 1990).

Two overarching themes are present in much of the research related to one or more of these four skills as well as broader conceptualizations of language ability (e.g., Bachman, 1990). The first theme is that language is used with communicative goals in mind. For the modalities of reading and listening, the communicative goal is to understand written or spoken texts with particular characteristics (e.g., a particular genre) for a strategic purpose (e.g., for implied meaning, for the main idea). For speaking and writing, the communicative goal is achieved by successfully performing a specific communicative

task (e.g., making a request, describing an activity). The second theme is that linguistic knowledge and subcompetencies (i.e., language knowledge and skills) are used to achieve communicative goals. Most components of linguistic knowledge and subcompetencies are utilized across modalities of communication and have been articulated in more general models of communicative competence or language ability (e.g., Bachman, 1990). These components include lexical, grammatical, discourse, phonological, and orthographic knowledge of language, as well as pragmatic and strategic competencies.

We incorporated these two themes (communicative goals; linguistic knowledge and subcompetencies) into our initial construct definitions for each of the four skills. For example, the initial construct definition for *speaking proficiency in everyday adult life* included a list of important communicative goals for speakers in the TLU domain (e.g., expressing an opinion) and a broad set of linguistic skills and subcompetencies needed to realize various communicative goals (e.g., the ability to use high-frequency vocabulary appropriate to a task, or lexical knowledge and use).

An important aspect of conceptualizing English language proficiency that is often overlooked is whether the underlying proficiency model emphasizes native-like competence or communicative effectiveness (Hu, 2017). This aspect is particularly relevant for conceptualizing and evaluating speaking proficiency, where emphasis may be placed on the accuracy of form in relation to the norms of a particular variety of English (i.e., emphasizing native-like competence) or on the comprehensibility and communicative impact of speech (i.e., communicative effectiveness). Given the recognition that a speaker's or writer's audience in the domain of English for everyday adult life may include native or nonnative speakers of English, an underlying proficiency model based on communicative effectiveness (as opposed to native-like competence) will inform the construct definition, development, and scoring of the redesigned TOEIC Bridge tests.

Defining Beginner to Low-Intermediate English Proficiency

In the third phase of the domain analysis, we closely examined descriptors of language proficiency standards relevant to the modalities (reading, listening, speaking, writing) and components of linguistic knowledge and subcompetencies identified in the previous phase. This analysis served two purposes. First, one of the mandates for test design—as documented in the initial logic model—was the need to map test scores to language proficiency standards to enhance the interpretation of test scores. Incorporating information from language proficiency standards during the test design stage provides stronger evidence of alignment (Council of Europe, 2009). The second purpose of this analysis was to produce artifacts that could inform task and scoring rubric design. Whereas the prior review of theory and research literature helped inform the construct definition for each test section, it provided minimal guidance on the types of communication goals and linguistic knowledge and subcompetencies expected of L2 users at different levels of proficiency (e.g., communicative goals appropriate for a low beginner versus a high beginner).

With this background in mind, we identified levels of the CEFR standards (Council of Europe, 2018), the CLB (Centre for Canadian Language Benchmarks, 2012), and the ACTFL proficiency guidelines (ACTFL, 2012) that were relevant to the range of beginner to low-intermediate English proficiency. Given the mandate to target the assessment of proficiency from CEFR Levels Pre-A1 to B1, we reviewed relevant descriptors across this range. Based on a study that mapped ACTFL proficiency levels to CEFR levels (Bärenfänger & Tschirner, 2012), we reviewed descriptors associated with ACTFL proficiency levels up to the intermediate high level for speaking and writing, and up to the advanced low level for reading and listening. We also reviewed descriptors associated with CLB Levels 1 to 6 based on Vandergrift's (2006) proposed alignment between CLB and CEFR levels. The CEFR and CLB include overall descriptor scales for each modality as well as more detailed scales that describe more specific activities, competencies, or strategies associated with each modality. Since the CEFR includes a wide range of descriptor scales, we restricted our review to scales relevant to the initial construct definition (see Appendix A for a full list of the CEFR descriptor scales that were reviewed).

For each modality (reading, listening, speaking, writing), we aggregated information across standards and relevant descriptor scales that aligned with CEFR Levels Pre-A1, A1, A2, and B1. For example, for the speaking beginner level (CEFR A1, CLB 1-2, ACTFL novice high), we summarized information in relevant descriptors as they pertained to communication goals, topics, characteristics of the input, and linguistic skills and subcompetencies (lexical knowledge, grammatical knowledge, discourse knowledge, phonological knowledge, pragmatic competence). The summary produced for the speaking beginner level is reproduced in Appendix B.

Although the overall structure of our construct definitions was not affected by the analysis of language proficiency standards, the analysis helped us refine some of the language in our construct definitions. The analysis allowed us to cross-validate our lists of communication goals by comparing them to communicative activities highlighted within and across standards. We also refined some of the language used to describe different linguistic skills and subcompetencies based on standards-based descriptors of these skills at the low-intermediate level. Thus, the analysis did not have a major impact on the components of language ability that were included in the construct definitions (e.g., communicative goals, various linguistics skills and subcompetencies) that were theoretically derived; rather, it helped us refine or cross-validate our expectations of how these components would be realized for beginner to low-intermediate learners.

CONSTRUCT DEFINITION FOR AN ASSESSMENT OF BEGINNING TO LOW-INTERMEDIATE ENGLISH LANGUAGE PROFICIENCY FOR EVERYDAY ADULT LIFE

In this section, we present the proposed construct definition for each of the proposed redesigned TOEIC Bridge tests. The construct for each test section is based on the interactionalist approach to construct definition (described in the Background subsection) and reflects a theoretical approach in which language proficiency is demonstrated by using linguistic knowledge and subcompetencies to achieve communication goals. This overall approach of focusing on communication goals and linguistic knowledge and subcompetencies in context (for each of the four language skills) was based on the reviews described in the subsections Defining the Target Language Use Domain of Everyday Life and Defining English Reading, Listening, Speaking, and Writing Proficiency.

The construct definition for each test begins with a broad statement about what the test intends to measure and then lists the communication goals relevant to the use of English at beginning to low-intermediate levels in the context of everyday adult life. This statement is followed by an elaboration of the specific linguistic knowledge and subcompetencies needed to achieve the communication goals. The categories of linguistic knowledge and subcompetencies (i.e., lexical, grammatical, discourse, phonological, and orthographic knowledge; pragmatic and strategic competence) are generally consistent across all four tests. For each test section, the communication goals and linguistic knowledge and subcompetencies listed also reflect our principled analysis of relevant language proficiency standards (CEFR, CLB, ACTFL). As previously described in the subsection Defining Beginner to Low-Intermediate English Proficiency, this analysis helped refine specific elements of each construct definition and produced artifacts that were used to guide the test development process.

The redesigned TOEIC Bridge tests are a measure of the ability of beginning and low-intermediate learners of English to communicate in personal, public, and general workplace contexts and to comprehend and produce basic spoken and written texts commonly occurring in everyday adult life. The construct definitions for each test section (listening, reading, speaking, and writing) are found in Appendix C.

CONCLUDING COMMENTS

This paper described the process used to produce a construct definition for a new suite of language proficiency tests, the redesigned TOEIC Bridge tests. The process followed a mandate-driven approach to ECD. This approach began by defining the mandate for test design, including test purpose and intended uses, stakeholders, and a logic model that specified assessment components, hypothesized actions (intended uses), and hypothesized intermediate and long-term effects (impact or consequences of test use). Based on this mandate, a domain analysis was conducted that further elaborated the TLU domain (i.e., English for everyday adult life) and targeted language proficiency competencies (i.e., reading, listening, speaking, and writing skills). In order to facilitate alignment between the assessment and language proficiency standards, produce artifacts that could support the next stages of the test development process (i.e., domain modeling and the CAF), and further refine the initial construct definition based on the targeted proficiency levels, we analyzed relevant descriptors from three language proficiency standards (CEFR, CLB, and ACTFL).

The outcome of this work is a proposed construct definition for each test that is based on theory, research, and relevant language proficiency standards. The construct definition reflects an interactionalist approach that specifies characteristics of the TLU domain (e.g., setting, audience, communication goals) and relevant linguistic skills and subcompetencies. These construct definitions provide a basis for the next steps in the ECD process—domain modeling and development of the CAF—as well as justification for the intended meaning of test scores and intended uses of the test. In addition, the construct definitions provide the basis for subsequent evaluations of interpretations and uses—validity research—based on the actual ensuing assessment.

REFERENCES

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. **https://doi.org/10.1017/CBO9780511732935**

American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines 2012*.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–72). University of Ottawa Press. **https://doi.org/10.2307/j.ctt1ckpccf.9**

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford.

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

Bärenfänger, O., & Tschirner, E. (2012). *Assessing evidence of validity of assigning CEFR ratings to the ACTFL Oral Proficiency Interview (OPI) and the Oral Proficiency Interview by computer (OPIc)*. Language Testing International. **http://www.global8.or.jp/OPIc%20CEFR%20Study%20Final%20Report%20pdf.pdf**

Bejar, I., Douglas, D., Jamieson, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 Listening framework: A working paper* (TOEFL Monograph Series MS-19). ETS.

Bennett, R. E. (2010). Cognitively Based Assessment of, for, and as Learning (*CBAL®*): A preliminary theory of action for summative and formative assessment. *Measurement: Interdisciplinary Research and Perspectives, 8*(2–3), 70–91. **https://doi.org/10.1080/15366367.2010.508686**

Bronfenbrenner, U. (1979). *The ecology of human development*. Harvard University Press.

Buck, G. (2001). *Assessing listening*. Cambridge University Press. **https://doi.org/10.1017/CBO9780511732959**

Centre for Canadian Language Benchmarks. (2012). *Canadian language benchmarks: English as a second language for adults*.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383. **https://doi.org/10.1191/0265532203lt264oa**

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. Bachman and A. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). Cambridge University Press. **https://doi.org/10.1017/CBO9781139524711.004**

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language.* Routledge.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.*

Council of Europe. (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors.*
https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Douglas, D. (2000). *Assessing language for specific purposes.* Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18*(2), 171–185. **https://doi.org/10.1177/026553220101800204**

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening* (pp. 77–151). Cambridge University Press.

Fulcher, G. (2013). Test design and retrofit. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Blackwell. **https://doi.org/10.1002/9781405198431.wbeal1199**

Hines, S. (2010). *Evidence-centered design: The TOEIC Speaking and Writing tests.* ETS.

Hu, G. (2017). The challenges of world Englishes for assessing English proficiency. In E. Low & A. Pakir (Eds.), *World Englishes: Rethinking paradigms.* Taylor and Francis.

Jacobs, H. L., Zingram, D. R., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Series Report No. 16). ETS.
https://www.ets.org/Media/Research/pdf/RM-00-03-Jamieson.pdf

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, *64*(1), 160–212. **https://doi.org/10.1111/lang.12034**

Kane, M. T. (2011). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17.
https://doi.org/10.1177/0265532211417210

Kenyon, D. (2014, May 29–June 1). *From test development to test use consequences: What roles does the CEFR play in a validity argument?* [Invited keynote presentation]. European Association of Language Testing and Assessment 11th Annual Conference, University of Warwick, Coventry, United Kingdom.

Knoch, U., & Macqueen, S. (2016). Language assessment for the workplace. In D. Tsagari & J. Baneerjee (Eds.), *Handbook of second language assessment* (pp. 291–307). De Gruyter Mouton.

Koda, K. (2013). Assessment of reading. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell. **https://doi.org/10.1002/9781405198431.wbeal0051**

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. MIT Press.

McNamara, T. (2011). Managing learning: Authority and language assessment. *Language Teaching, 44*(4), 500–515. **https://doi.org/10.1017/S0261444811000073**

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design* (Research Report No. RR-03-16). ETS. **https://doi.org/10.1002/j.2333-8504.2003.tb01908.x**

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement, 1*(1), 3–62. **https://doi.org/10.1207/S15366359MEA0101_02**

Mislevy, R. J., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). Routledge.

Norris, J. M. (2013, October 25). *Reconsidering assessment validity at the intersection of measurement and evaluation* [Invited plenary address]. East Coast Organization of Language Testers Annual Conference. Georgetown University, Washington, DC, United States.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Sage Publications.

Purpura, J. (2004). *Assessing grammar*. Cambridge University Press. **https://doi.org/10.1017/CBO9780511733086**

Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2015). In S. Lange, T. M. Haladyna, & M. Raymond (Eds.), *Handbook of test development* (2nd ed., pp. 40–63). Routledge.

Rost, M. (2005). L2 listening. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 503–507). Routledge. **https://doi.org/10.4324/9781410612700**

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

Vandergrift, L. (2006). *New Canadian perspectives: Proposal for a common framework of reference for languages in Canada*. Canadian Heritage.

Van Lier, L. (2000). From input to affordance: Social-interactive learning from an ecological perspective. In J. P. Lantolf (Ed.), *Sociocultural theory and second language learning* (pp. 245–259). Oxford University Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wei, H., Mislevy, R. J., & Kanal, D. (2008). *An introduction to design patterns in language assessment* (PADI Technical Report 18). SRI International.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. **https://doi.org/10.1017/CBO9780511732997**

Weigle, S. C. (2013). Assessment of writing. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell. **https://doi.org/10.1002/9781405198431.wbeal0056**

Weir, C. J. (1990). *Communicative language testing*. Prentice Hall.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. **https://doi.org/10.1057/9780230514577**

Xi, X. (2015, March 16–20). *Language constructs revisited for practical test design, development and validation* [Paper presentation]. 37th Annual Language Testing Research Colloquium, Toronto, Canada.

APPENDIX A. COMMON EUROPEAN FRAMEWORK OF REFERENCE DESCRIPTOR SCALES INCLUDED IN THE REVIEW

Communicative language activity	Descriptor scales included
Reading comprehension	• Overall reading comprehension • Reading correspondence • Reading for orientation • Reading for information and argument • Reading instructions • Identifying cues and inferring
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 • Identifying cues and inferring
Spoken production	• Overall spoken production • Sustained monologue: describing experience • Sustained monologue: giving information • Sustained monologue: putting a case • Public announcements
Spoken interaction	• Informal discussion • Obtaining goods and services • Information exchange • Phonological control
Written production	• Overall written production • Creative writing • Written reports and essays
Written interaction	• Overall written interaction • Correspondence • Notes, message, and forms
Other (interaction strategies, linguistic, sociolinguistic, pragmatic)	• Online conversations and discussion • General linguistic range • Vocabulary range • Grammatical accuracy • Vocabulary control • Thematic development • Coherence and cohesion • Propositional precision • Spoken fluency • Sociolinguistic appropriateness

APPENDIX B. SUMMARY OF SCALE DESCRIPTORS RELEVANT TO THE SPEAKING CONSTRUCT DEFINITION AT COMMON EUROPEAN FRAMEWORK OF REFERENCE LEVEL A1 (AND CLB LEVELS 1 TO 2, AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGE LEVEL NOVICE HIGH) FROM LANGUAGE PROFICIENCY STANDARDS

Category	Summary
Communication goals	• Ask and respond to simple, direct questions and statements (CEFR, CLB, ACTFL) • Description (CEFR, CLB) • Read a short, prepared/rehearsed statement (CEFR) • Use and respond to basic courtesy formulas and greetings (CEFR, CLB) • Give brief, common, routine instructions (CLB) • Express basic ability or inability (CLB) • Limited number of activities and preferences (ACTFL); Express likes and dislikes (CLB)
Topics	• People, places (CEFR) • Areas of immediate need or very familiar topics, such as asking for assistance, or the time, price, or an amount (CEFR, CLB, ACTFL); Very simple warnings and cautions (CLB) • Very basic personal information: description, occupation, surroundings (CEFR, CLB, ACTFL) • Basic everyday, routine communication (CLB) • Straightforward social situations (ACTFL) • Basic objects (ACTFL)
Characteristics of the input	• Slower speech rate (CEFR) • Questions and instructions addressed carefully and slowly; short, simple directions (CEFR) • Allow rephrasing and repair (CEFR, ACTFL) • Sympathetic or supportive interlocutor (CEFR, CLB, ACTFL)
Linguistic skills and subcompetencies	
Lexical knowledge and use	• Common, familiar words (CLB); money, prices, amounts (CLB); sizes, colors, numbers (CLB); concrete objects (CLB); likes and dislikes (CLB); numbers, quantity, cost, time (CEFR) • Formulaic expressions (CLB); common greetings, introductions, and leave-takings (CEFR, CLB) • May significantly impede communication (CLB) • Numbers and dates, name, nationality, address, age, date of birth, etc. (CEFR) • Basic vocabulary repertoire of isolated words and phrases related to particular concrete situations (CEFR)
Grammatical knowledge and use	• Simple phrases (CEFR) • Imperative forms (CLB); both positive and negative commands (CLB) • Tend to use present tense (CLB, ACTFL) • Little or no control over basic grammar structures and tenses (CEFR, CLB, ACTFL) • May significantly impede communication (CLB)

Category	Summary
Discourse knowledge and use	• Mainly isolated words or phrases, no or little evidence of connected discourse (CEFR, CLB) • Link words or simple phrases with very basic linear connectors such as "and" or "then" (CEFR) • Short conversational openings and closings (CEFR, CLB)
Phonological knowledge and use	• Not adequate to sustain simple conversations (CLB, ACTFL) • Slow speech rate with frequent pauses, hesitations, repetitions; rephrasing and repair (CEFR, CLB, ACTFL) • Pronunciation difficulties may significantly impede communication (CLB) • Use alphabet to spell out words, such as name (CLB)
Pragmatic competence	• Use appropriate courtesy words (CLB)

APPENDIX C. CONSTRUCT DEFINITIONS FOR THE REDESIGNED TOEIC BRIDGE TESTS: LISTENING, READING, SPEAKING, AND WRITING

LISTENING

The redesigned 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. This includes the ability to understand high-frequency vocabulary, formulaic phrases, and the main ideas and supporting details of clearly articulated speech across familiar varieties of English on familiar topics. Test takers can comprehend simple greetings, introductions, and requests; instructions and directions; descriptions of people, objects, situations; personal experiences or routines; and other basic exchanges of information.

Communication Goals

In English, test takers can understand commonly occurring spoken texts, demonstrating the ability to

- understand simple descriptions of people, places, objects, and actions;
- understand short dialogues or conversations on topics related to everyday life (e.g., making a purchase); and
- understand short spoken monologues as they occur in everyday life (e.g., an announcement in a public area) when they are spoken slowly and clearly.

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower intermediate English language learners need the ability to

- understand common vocabulary and formulaic phrases (lexical knowledge);
- understand simple sentences and structures (grammatical knowledge);

- understand sentence-length speech and some common registers (discourse knowledge);
- 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);
- infer implied meanings, speaker roles, or context in short, simple spoken texts (pragmatic competence); and
- understand the main idea and stated details in short spoken texts (listening strategies).

READING

The redesigned TOEIC Bridge Reading test measures the ability of beginning and lower intermediate English language learners to understand short written English texts in personal, public, and familiar workplace contexts and across a range of formats. This includes the ability to understand high-frequency vocabulary, formulaic phrases, and the main ideas and supporting details of short written texts dealing with familiar topics. Test takers can comprehend simple texts such as signs, lists, menus, schedules, advertisements, narrations, routine correspondence, and short descriptive texts.

Communication Goals

In English, test takers can understand commonly occurring written texts, demonstrating the ability to

- understand nonlinear written texts (e.g., signs, schedules);
- understand written instructions and directions;
- understand short, simple correspondence; and
- understand short informational, descriptive, and expository written texts about people, places, objects, and actions.

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower intermediate English language learners need the ability to

- understand common vocabulary (lexical knowledge);
- understand simple sentences and structures (grammatical knowledge);
- understand the organization of short written texts in a variety of formats (discourse knowledge);
- recognize simple mechanical conventions of written English (orthographic knowledge);
- infer implied meanings, including context or writer's purpose, in short, simple written texts (pragmatic competence); and
- understand the main idea and stated details in short written texts; infer the meaning of unknown written words through context clues (reading strategies).

SPEAKING

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. This includes the ability to communicate immediate needs, provide basic information, and interact on topics of personal interest with people who are speaking clearly. Test takers can answer simple questions on familiar topics and use phrases and sentences to describe everyday events. They can provide brief reasons for and explanations of their opinions and plans and narrate simple stories.

Communication Goals

In spoken English, perform simple communication tasks, demonstrating the ability to

- ask for and provide basic information;
- describe people, objects, places, activities;
- express an opinion or plan and give a reason for it;
- give simple directions;
- make simple requests, offers, and suggestions; and
- narrate and sequence simple events.

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower intermediate English language learners need the ability to

- use high-frequency vocabulary appropriate to a task (lexical knowledge);
- use common grammar structures to contribute to overall meaning (grammatical knowledge);
- use simple transitions to connect ideas (e.g., so, but, after—discourse knowledge);
- pronounce words in a way that is intelligible to proficient speakers of English; use intonation, stress, and pauses to pace speech and contribute to comprehensibility (phonological knowledge); and
- produce speech that is appropriate to the communication goal (pragmatic competence).

WRITING

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. This includes the ability to use high-frequency vocabulary and basic grammar structures to produce phrases, sentences, and paragraphs on subjects that are familiar or of personal interest. Test takers can write notes and messages relating to matters of immediate need. They can write simple texts, such as personal letters describing experiences and giving simple opinions.

Communication Goals

In written English, perform simple communication tasks, demonstrating the ability to

- ask for and provide basic information;
- make simple requests, offers, and suggestions, express thanks;
- express a simple opinion and give a reason for it;
- describe people, objects, places, activities; and
- narrate and sequence simple events.

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower intermediate English language learners need the ability to

- use high-frequency vocabulary appropriate to a task (lexical knowledge);
- write a sentence using simple word order, such as subject-verb-object, interrogatives, imperatives; use common grammatical structures to contribute to meaning (grammatical knowledge);
- arrange ideas using appropriate connectors (e.g., for example, *in addition, finally*); sequence ideas to facilitate understanding (discourse knowledge);
- control mechanical conventions of English (spelling, punctuation, and capitalization) to facilitate comprehensibility of text (orthographic knowledge); and
- produce text that is appropriate to the communication goal (pragmatic competence).

DEVELOPMENT OF THE REDESIGNED *TOEIC BRIDGE*® TESTS

Philip Everson, Trina Duke, Pablo Garcia Gomez, Elizabeth Carter Grissom, Elizabeth Park, and Jonathan Schmidgall

The test design process for the redesigned *TOEIC Bridge*® tests was a collaboration among researchers, content developers, psychometricians, and the business directors of the *TOEIC*® program following a process of evidence-centered design (ECD). ECD can be viewed as a methodology that comprises best practices for the creation and ongoing development of an assessment. It clarifies what is being measured by a test and supports inferences made on the basis of evidence derived from the test. ECD systematizes test design by specifying a process with five stages or layers, including domain analysis, domain modeling, construction of an assessment framework, assessment implementation, and assessment delivery (Mislevy & Yin, 2012). As shown in Figure 1, these stages concretize what we want to be able to say about test takers based on observations we make on their performance on the test tasks.

Layer	Role	Key entities or Components	Explanation of key entity or component
1. Domain analysis	Gather information about what is to be assessed	Analysis and summary of theory, research, and expert judgment as it pertains to what is to be assessed	Language framework, proficiency guidelines, etc.
2. Domain modeling	Incorporation of information from stage one into three components; sketch of potential variables and substantive relationships	Proficiency paradigm	Substantive construct expressed as claims
		Evidence paradigm	Observations required to support claims
		Task paradigm	Types of situations that provide opportunities for test takers to show evidence of their proficiencies
3. Construction of conceptual assessment design framework	Development of a final blueprint; provide technical detail required for implementation including statistical models, rubrics, specifications, and operational requirements	Student model	Statistical characterization of the abilities to be assessed
		Evidence model	1. Rules for scoring test tasks
			2. Rules for updating variables in the student model
		Task model	Detailed description of assessment tasks
		Presentation model	Specification of how the assessment elements will look during testing
		Assembly model	Specification of the mix of tasks on a test for a particular student
4. Assessment implementation	Operational item writing and form assembly	Task materials, work products, operational data	Rendering protocols for tasks, tasks as displayed, etc.
5. Assessment delivery	Test administration and scoring	Tasks as presented, work products as created, scores as evaluated	Actual rendering of task materials in assessment, score reports, etc.

Figure 1. Layers of the evidence-centered design process. Green indicates evidence-centered design steps addressed in this paper.

This research memorandum is concerned primarily with the development of the ECD steps shaded in green in Figure 1:

- Task models
- Presentation models
- Assembly models

The other steps in the process that precede and follow these three are discussed in Lin et al. (2019) and Schmidgall et al. (2019).

Task modeling begins with the development of prototype tasks. Multiple tasks were developed for each of the four assessments. In many cases, two or more versions of the same prototype task were developed, where the specifications for the versions varied in some important way—for instance, different response times or different levels of specificity in the directions. Prototype tasks were evaluated through small-scale user-acceptance testing and larger scale piloting. Through pilot testing, developers were able to finalize task specifications and, for speaking and writing, finalize the rubrics used to score productive tasks.

To a certain extent, task modeling overlaps with the evidence paradigm and the task paradigm from the domain modeling stage of ECD. The domain definitions for the redesigned TOEIC Bridge tests were based on the proposed construct definition that was a result of the domain analysis stage, described in detail in Schmidgall et al. (2019). The domain definitions include communication goals, and the communication goals are, for the most part, definitions of task paradigms. They state, at an abstract level, the kinds of situations that allow test takers to show evidence of ability. In the case of the listening domain of the redesigned TOEIC Bridge test, the domain description includes the communication goals (among others) of "understand short, simple descriptions" and "understand short conversations." These communication goals define the kinds of tasks that would be appropriate to include in an operational assessment aligned with the domain definitions.

If task models are concerned with representing as fully as possible specific communication goals, or evidence paradigms, as they occur in the real world, presentation models focus on the task types as test items and evaluate the tasks from the point of view of the test taker. Primary questions include the following: Is the task accessible? Do test takers know what they are supposed to do? If the task is timed, do test takers have adequate time to consider and complete the task? Are all the tools available in the testing platform easy to access and use? These questions are particularly important for an assessment like the redesigned TOEIC Bridge tests because directions and collateral material are in English, and the test takers are beginning to intermediate English learners.

After pilot testing, test developers were able to create draft test blueprints for each of the four assessments. Pilot testing provided evidence for which prototype tasks or versions of tasks produced usable evidence to support the claims derived from the domain model. The draft test blueprints were used to create the forms to be administered in the field test.

Task Modeling

The process of designing task prototypes for the redesigned TOEIC Bridge suite of assessments began with discussions of the program requirements that were necessary to make the final product useful in the marketplace and that affected test design. These program requirements informed the initial domain analysis and construct definitions for the redesigned TOEIC Bridge assessments as described in Schmidgall et al. (2019) but led to additional considerations for task modeling that initiated the process of operationalizing the construct definition.

The following is a partial list of the business requirements that were most relevant to assessment design:

- The redesigned TOEIC Bridge tests will measure all four language skills—listening, reading, speaking, and writing—and provide scores and feedback on each.
- Each of the four assessments will focus on representative communication skills at the A1, A2, and B1 levels of the Common European Framework of Reference (CEFR).
- The tests will be module based so that score users can require and test takers can take different combinations of skills.
- The listening and reading assessments will be administered on paper but designed so that future computer-delivered versions will be possible.
- The speaking and writing assessments will be computer based.
- The listening and reading assessments will be machine scored.
- The speaking and writing assessments will be scored by human raters.
- The combined testing time for the listening and reading assessments should not exceed the testing time of the existing TOEIC Bridge test.
- Accents from the United States, Canada, the United Kingdom, and Australia will be used in the listening and speaking stimulus materials.
- The assessments will include, where possible, contemporary means of communication, such as e-mail and instant messages.
- The assessment design will promote meaningful mapping to the CEFR.
- The assessments will provide meaningful feedback to teachers and learners in the form of proficiency descriptors.

Some of these requirements were motivated by the desire that the redesigned TOEIC Bridge assessments be consistent in important respects with other components of the TOEIC family of assessments—for instance, the inclusion of varied accents in the listening assessment. Others were motivated by considerations of the overall cost structure of the operational assessment, such as the use of human raters to score the speaking and writing tests. They all were taken into account in task design so that the final assessment design was helpful to end users. The requirements that the redesigned TOEIC Bridge

assessments be meaningfully mapped to the CEFR and other internationally recognized language standards and that they provide appropriate feedback to teachers and learners made following an ECD process especially important.

A second, and equally important, set of guidelines for prototype task development was the product of the domain analysis ECD step, as described in Schmidgall et al. (2019). The first product was the definition of the assessments' overall target language use (TLU) domain. The TLU was defined as "everyday adult life" and included three subdomains: the personal sphere, the public sphere, and the workplace sphere. Building on the overall definition of the TLU domain of everyday adult life, the test designers then created domain definitions for each of the four skills—listening, reading, speaking and writing—with explicit communication goals and underlying competencies that support the successful completion of the communication goals. These domain definitions also incorporated information from a principled review of the language proficiency standards expected to be most relevant to score users, including the CEFR standards, Canadian Language Benchmarks, and American Council on the Teaching of Foreign Language's proficiency guidelines. The review of language proficiency standards also produced summaries of the language activities, strategies, and competencies relevant to the range of proficiency levels targeted by the test (i.e., CEFR A1 to B1) that informed test development. Figures 2–5 show the four domain definitions that guided task development for each section of the redesigned TOEIC Bridge test.

Listening Domain Definition

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. This includes the ability to understand high-frequency vocabulary, formulaic phrases, and the main ideas and supporting details of clearly articulated speech across familiar varieties of English on familiar topics. Test takers can comprehend simple greetings, introductions, requests, instructions, and directions; descriptions of people, objects, situations, personal experiences, or routines; and other basic exchanges of information.

Communication Goals

In English, test takers can understand commonly occurring spoken texts, demonstrating the ability to

- understand simple descriptions of people, places, objects, and actions
- understand short dialogues or conversations on topics related to everyday life (e.g., making a purchase)
- understand short spoken monologues as they occur in everyday life (e.g., an announcement in a public area) when they are spoken slowly and clearly

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower-intermediate English language learners need the ability to

- understand common vocabulary and formulaic phrases (lexical knowledge)
- understand simple sentences and structures (grammatical knowledge)
- understand sentence-length speech and some common registers (discourse knowledge)
- 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)
- infer implied meanings, speaker roles, or context in short, simple spoken texts (pragmatic competence)
- understand the main idea and stated details in short, spoken texts (listening strategies)

Figure 2. Listening domain definition.

Reading Domain Definition

The TOEIC Bridge Reading test measures the ability of beginning and lower-intermediate English language learners to understand short written English texts in personal, public, and familiar workplace contexts and across a range of formats. This includes the ability to understand high-frequency vocabulary, formulaic phrases, and the main ideas and supporting details of short, written texts dealing with familiar topics. Test takers can comprehend simple texts such as signs, lists, menus, schedules, advertisements, narrations, routine correspondence, and short descriptive texts.

Communication Goals

In English, test takers can understand commonly occurring written texts, demonstrating the ability to

- understand nonlinear written texts (e.g. signs, schedules)
- understand written instructions and directions
- understand short, simple correspondence
- understand short informational, descriptive, and expository written texts about people, places, objects, and actions

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower-intermediate English language learners need the ability to

- understand common vocabulary (lexical knowledge)
- understand simple sentences and structures (grammatical knowledge)
- understand the organization of short written texts in a variety of formats (discourse knowledge)
- recognize simple mechanical conventions of written English (orthographic knowledge)
- infer implied meanings, including context or writer's purpose in short, simple written texts (pragmatic competence)
- understand the main idea and stated details in short, written texts; infer the meaning of unknown written words through context clues (reading strategies)

Figure 3. Reading domain definition.

Speaking Domain Definition

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. This includes the ability to communicate immediate needs, provide basic information, and interact on topics of personal interest with people who are speaking clearly. Test takers can answer simple questions on familiar topics and use phrases and sentences to describe everyday events. They can provide brief reasons for and explanations of their opinions and plans and narrate simple stories.

Communication Goals

In spoken English, perform simple communication tasks, demonstrating the ability to

- ask for and provide basic information
- describe people, objects, places, activities
- express an opinion or plan and give a reason for it
- give simple directions
- make simple requests, offers, and suggestions
- narrate and sequence simple events

Linguistic Knowledge and Subcompetencies

To achieve these goals, beginning and lower-intermediate English language learners need the ability to

- use high-frequency vocabulary appropriate to a task (lexical knowledge)
- use common grammar structures (grammatical knowledge)
- use simple transitions to connect ideas, e.g., *so, but, after* (discourse knowledge)
- pronounce words in a way that is intelligible to native speakers and proficient nonnative speakers of English; use intonation, stress, and pauses to pace speech and contribute to comprehensibility (phonological knowledge)
- produce speech that is appropriate to the communication goal (pragmatic competence)

Figure 4. Speaking domain definition.

Writing Domain Definition

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. This includes the ability to use high-frequency vocabulary and basic grammar structures to produce phrases, sentences, and paragraphs on subjects that are familiar or of personal interest. Test takers can write notes and messages relating to matters of immediate need. They can write simple texts such as personal letters describing experiences and giving simple opinions.

Communication Goals

In written English, perform simple communication tasks, demonstrating the ability to

- ask for and provide basic information
- make simple requests, offers, and suggestions; express thanks
- express a simple opinion and give a reason for it
- describe people, objects, places, activities
- narrate and sequence simple events

Linguistic Knowledge and Subcompetencies

- To achieve these goals, beginning and lower-intermediate English language learners need the ability to
- use high-frequency vocabulary appropriate to a task (lexical knowledge)
- write a sentence using simple word order, such as SVO (subject/verb/object); interrogatives; imperatives; use common grammatical structures to contribute to meaning (grammatical knowledge)
- arrange ideas using appropriate connectors (e.g., *for example, in addition, finally*); sequence ideas to facilitate understanding (discourse knowledge)
- use mechanical conventions of English (spelling, punctuation, and capitalization) to facilitate comprehensibility of text (orthographic knowledge)
- produce text that is appropriate to the communication goal (pragmatic competence)

Figure 5. Writing domain definition.

Task Design

It should be noted that at the beginning of the task model development process, the communication goals to be measured on the assessment are aspirational. That is, test designers may be more or less successful in creating tasks that are authentic and valid representations of the communication goals. Communication tasks that are carried out within the subdomains of everyday adult life (personal, public, and familiar workplace contexts) at beginning to low intermediate levels of proficiency cannot always be exactly replicated in a language test. To the extent possible, real-world tasks are used or approximated in the TOEIC family of tests to maximize the validity of test scores and proficiency descriptors.

For each language skill, the domain definitions outline the claims to be made about test-taker abilities. More tasks were developed for pilot testing than final versions of the tests would contain, and specific questions were posed so that pilot test results would better inform later decisions. These decisions included task choice as well as specific task characteristics such as the presentation of items, preparation and response times, and rubric refinement. It should be noted that the pilot forms were not intended to be draft versions of operational forms but merely the delivery of individual tasks that designers considered to be likely candidates for operational use.

Listening Task Modeling

In developing prototype tasks for the redesigned TOEIC Bridge Listening assessment, 12 different tasks or task variants were considered before the pilot. Of these, five were chosen for piloting based on the criteria of most efficient representation of the construct, as shown in Table 1. The pilot tasks included the following (tasks are identified here by the shorthand name used during design discussions):

1. Photographs. The test taker looks at a photograph, listens to four 1-sentence options, and chooses the option that best describes the content of the photograph.
2. Four Pictures (Listening). The test taker listens to a one-sentence description of a person, place, or object and then selects from four graphic options the picture that is consistent with the stimulus.
3. Question-Response. The test taker listens to the first half of a conversational exchange and then selects from four options the response appropriate to the exchange.
4. Conversations. The test taker listens to a short conversation, and comprehension is assessed by a set of multiple-choice questions. In a variant of this task, the spoken stimulus is supplemented by a short, very simple text graphic, such as a schedule or address, and one or more of the multiple-choice questions requires the test taker to connect information in the audio stimulus to information in the text.
5. Talks. The test taker listens to a monologue, and comprehension is assessed by sets of multiple-choice questions. As with conversations, some monologues are supplemented by simple graphical information that must be synthesized with the spoken information to answer one or more questions.

Table 1 shows how the listening tasks were expected to align with the CEFR levels that the redesigned TOEIC Bridge tests (A1-B1) were meant to measure, the communication goals outlined in the listening domain definition, and the relevant linguistic subskills. All the listening prototype tasks are shown in the table to align with more than one CEFR level because specific items within the task type, depending on content or context or other features, may align at different levels. For instance, in the Question-Response item type, the exchange being tested may be an extremely common formula and align with CEFR level A1. Another exchange may be less common, require more contextual knowledge by the listener, and align with level A2.

TABLE 1

Alignment of Listening Prototype Tasks With Domain Definition

Domain definition	Description	Photo	Four Pics	Q/R	Cons	Talks
CEFR level	Corresponds to A1	✓	✓	✓		
	Corresponds to A2	✓	✓	✓	✓	✓
	Corresponds to B1			✓	✓	✓
Communication goals from domain definition	Understand simple descriptions of people, places, objects, and actions	✓	✓			
	Understand short dialogues or conversations on topics related to everyday life (e.g., making a purchase)			✓	✓	
	Understand short spoken monologues as they occur in everyday life (e.g., an announcement in a public area) when they are spoken slowly and clearly					✓
Linguistic knowledge and subcompetencies	Understand common vocabulary and formulaic phrases (lexical knowledge)	✓	✓	✓	✓	✓
	Understand simple sentences and structures (grammar knowledge)	✓	✓	✓	✓	✓
	Understand sentence-length speech and some common registers (discourse knowledge)	✓		✓	✓	✓
	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)	✓	✓	✓	✓	✓
	Infer implied meanings, speaker roles, or context in short, simple spoken texts (pragmatic competence)				✓	✓
	Understand the main idea and stated details in short, spoken texts (strategic competence)	✓	✓	✓	✓	✓

Note. Photo = Photographs; Four Pics = Four Pictures (Reading); Q/R = Question-Response; Cons = Conversations.

Reading Task Modeling

Eight different task types were modeled for the redesigned TOEIC Bridge Reading test. Of these, four were chosen for piloting based on considerations of construct coverage and practical issues of timing.

1. Four Pictures (Reading). The test taker reads a phrase or short sentence, then selects from four graphic options the one that best represents the content of the stimulus.
2. Sentence Completion. The test taker completes a cloze item based on a single-sentence assessing vocabulary and relatively simple grammatical structures.
3. Text Completion. The test taker completes a series of cloze items in the context of a multisentence-length paragraph. Sets include items that test vocabulary, appropriate word forms, and discourse knowledge (by selecting a sentence to be inserted into the paragraph).
4. Reading Comprehension. The test taker reads a 30- to 140-word stimulus and shows comprehension by answering two to three multiple-choice questions. The stimulus may be based on a range of genres, including websites and text message chains.

Table 2 shows how these prototype task models aligned with the CEFR levels the redesigned TOEIC Bridge assessments were intended to measure (A1–B1), the communication goals outlined in the reading domain definition, and the reading enabling skills or linguistic subskills. All the reading prototype tasks are shown as aligning with more than one CEFR level because they could be adapted to different levels of reading ability. For instance, the Sentence Completion task can be used to assess vocabulary knowledge. If the word tested is very common, then the task may be successfully completed by readers at the A1 level; if the word tested is less common, the task may align with A2 readers' skills.

TABLE 2

Alignment of Reading Prototype Tasks With Domain Definition

Domain definition	Description	Four Pics	Sent. Comp.	Text Comp.	Reading Comp.
CEFR level	Corresponds to A1	✓	✓		
	Corresponds to A2	✓	✓	✓	✓
	Corresponds to B1		✓	✓	✓
Communication goals	Understand nonlinear written texts (e.g., signs, schedules)				✓
	Understand written instructions and directions		✓	✓	✓
	Understand short, simple correspondence		✓	✓	✓
	Understand short informational, descriptive, and expository written texts about people, places, objects, and actions		✓	✓	✓
Linguistic knowledge and subcompetencies	Understand common vocabulary (lexical knowledge)	✓	✓	✓	✓
	Understand simple sentences and structures (grammatical knowledge)		✓	✓	✓
	Understand the organization of short written texts in a variety of formats (discourse knowledge)			✓	✓
	Recognize simple mechanical conventions of written English (orthographic knowledge)	✓	✓	✓	✓
	Infer implied meanings, including context or writer's purpose, in short, simple written texts (pragmatic competence)				✓
	Understand the main idea and stated details in short, written texts; infer the meaning of unknown written words through context clues (strategic competence)				✓

Note. Four Pics = Four Pictures (reading); Sent. Comp. = Sentence Completion; Text Comp. = Text Completion; Reading Comp. = Reading Comprehension

Speaking Task Modeling

Approximately 12 tasks were modeled, of which 10 were selected for piloting the TOEIC Bridge Speaking assessment. The prototype items ranged from a direct measure of an enabling skill—pronunciation—to several different task models intended to capture evidence of the ability to complete communication goals at varying levels of complexity. Because speaking task response time is much shorter than writing response time, more speaking prototype tasks and their variants can be piloted within practical administration time than writing tasks. The following items were administered in the speaking pilot:

1. Read Aloud. The test taker is given a short text to read aloud. The task assesses the subcompetencies of pronunciation and intonation.
2. Describe a Picture. The test taker is instructed to describe a photograph. The task assesses description of people, places, and objects.
3. Tell a Story. The test taker is presented with a series of four pictures that graphically convey a simple narrative. The test taker is instructed to tell the story out loud. The task assesses narration.
4. Respond to Questions. The test taker is asked two related questions about personal experiences. (What time do you get up? What do you eat for breakfast?) The task assesses asking for and providing straightforward information.
5. Respond to Questions With Information Provided. The test taker is given a brief text, such as an advertisement or schedule, with information in telegraphic form. The test taker then responds to three specific questions that can be answered with information from the text.
6. Give Two Reasons. The test taker is asked for a preference on a relatively concrete and immediate topic and to give two reasons for the preference. The task is intended to assess giving and supporting an opinion.
7. Express an Opinion. The test taker is presented with a prompt on a relatively abstract topic that requires the construction of an argument for support.
8. Listen-Speak. The test taker listens to a short (40–60 word) informative stimulus— for example, an announcement. The test taker is then required to tell a third person the important information in the stimulus. The task is designed to assess the communication goals of giving directions and narrating.
9. Ask/Invite/Request. The test taker is given a short text, such as a ticket to a sporting event or a receipt from a purchase and is instructed to role play a specific communication goal, such as inviting someone to do something or asking for help with a problem. The task is intended to assess making simple requests, offers, or suggestions.
10. Suggest a Solution. The test taker is given an audio stimulus of a telephone message in which the caller presents a problem and asks the test taker to respond with a solution. The task is intended to assess the communication goal of offering a suggestion.

Table 3 shows how the speaking prototype task models aligned with the CEFR levels the redesigned TOEIC Bridge assessments were intended to measure (A1–B1), the communication goals outlined in the speaking domain definition, and the relevant enabling skills or linguistic subskills. The Read Aloud task type did not align with any of the communication goals in the domain definition. It was intended to provide relevant information about a subskill, pronunciation and intonation, especially for test takers at the A1 level with very limited ability to communicate through speaking. Several of the communication goals in the table are aligned with multiple prototype tasks. At this point in the development process, it was not clear which task type would be most useful in an operational assessment, and the test designers expected to use data from pilot testing to make further decisions about the prototype tasks.

TABLE 3

Alignment of Speaking Prototype Tasks With Domain Definition

Domain definition	Description	RA	Desc Pic	Story	Resp	Resp With Info	Give 2	Op	L-S	Ask/ Inv/ Req	Sol
CEFR level	Corresponds to A1	✓	✓								
	Corresponds to A2				✓	✓	✓		✓	✓	✓
	Corresponds to B1			✓			✓	✓			
Communication goals	Ask for and provide basic information				✓	✓			✓	✓	
	Describe people, objects, places, activities		✓	✓							
	Narrate and sequence simple events			✓							
	Give simple directions								✓		
	Make simple requests, offers, and suggestions									✓	✓
	Express an opinion or plan and give a reason for it						✓	✓			

Domain definition	Description	RA	Desc Pic	Story	Resp	Resp With Info	Give 2	Op	L-S	Ask/ Inv/ Req	Sol
Linguistic knowledge and sub-competencies	Use high-frequency vocabulary		✓	✓	✓	✓	✓	✓	✓	✓	✓
	Use common grammatical structures		✓	✓	✓	✓	✓	✓	✓	✓	✓
	Use simple transitions to connect ideas			✓		✓	✓	✓	✓	✓	✓
	Pronounce words in a way that is intelligible to native speakers and proficient nonnative speakers of English	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Produce speech that is appropriate to the communication goal		✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. RA = Read Aloud; Desc Pic = Describe a Picture; Story = Tell a Story; Resp = Respond to Questions; Resp With Info = Respond to Questions With Information Provided; Give 2 = Give Two Reasons; Op = Opinion; L-S = Listen-Speak; Ask/Inv/Req = Ask/Invite/Request; Sol = Propose a Solution.

Writing Task Modeling

The test designers created drafts of a large number of writing task models and variants of the models, of which six were selected for piloting. For some task types, the pilot forms included some variations of directions and response times. The following six task types were included in the pilot:

1. Scrambled Sentence. The test taker is presented with four to six segments of a sentence in random order and must put them in appropriate order. This task was designed to assess the linguistic subskill of using common grammatical structures.
2. Write a Sentence Based on a Picture. The test taker is presented with a picture and two words and must write a sentence using both words that is consistent with the picture.
3. Respond to an E-mail. The test taker reads an e-mail message and then writes a response. The prompt explicitly gives the test taker two functions to be included in the response (i.e., "ask two questions," or "give two pieces of information").
4. Respond to an Instant Message. The test taker reads an instant message that requires a short narration in past, present, or future tenses in response.
5. Blog Post. The test taker is instructed to write a short narrative about a specific personal topic (for example, describe a time when you helped a friend).
6. Opinion. The test taker is instructed to write a 100–150-word text giving an opinion with support on a specific topic.

Table 4 shows how the writing prototype task models aligned with the CEFR levels the redesigned TOEIC Bridge assessments were intended to measure (A1–B1), the communication goals outlined in the writing domain definition, and the relevant enabling skills or linguistic subskills. One task type, Scrambled Sentences, did not align with any of the communication goals in the domain definition but only with a linguistic subskill. It was intended to provide relevant information about test takers who may be at a very basic level of writing development. Several of the communication goals in the table are aligned with multiple prototype tasks. At this point in the development process, it was not clear which task type would be most useful in an operational assessment, and the test designers expected to use data from pilot testing to make further decisions about the prototype tasks.

TABLE 4

Alignment of Writing Prototype Tasks With Domain Definition

Domain definition	Description	Scramb Sent	Pic-Sent	E-mail	IM	Blog Post	Op
CEFR level	Corresponds to A1	✓	✓				
	Corresponds to A2		✓	✓	✓		
	Corresponds to B1			✓		✓	✓
Communication goals	Ask for and provide basic information			✓			
	Describe people, places, objects, activities		✓		✓	✓	
	Narrate and sequence simple events				✓	✓	
	Make simple requests, offers, and suggestions			✓			
	Express a simple opinion and give a reason for it			✓			✓
Linguistic knowledge and subcompetencies	Use high-frequency vocabulary appropriate to a task (lexical knowledge and use)		✓	✓	✓	✓	✓
	Write a sentence using simple word order, such as SVO, interrogatives, imperatives; use common grammatical structures to contribute to meaning (grammatical knowledge and use)	✓	✓	✓	✓	✓	✓
	Arrange ideas using appropriate connectors (e.g., *for example, in addition, finally*); sequence ideas to facilitate understanding (discourse knowledge and use)		✓	✓	✓	✓	✓
	Use mechanical conventions of English (spelling, punctuation, and capitalization) to facilitate comprehensibility of text (orthographic knowledge and use)		✓	✓	✓	✓	✓
	Produce text that is appropriate to the communication goal (pragmatic competence)		✓	✓	✓	✓	✓

Note. Scramb Sent = Scrambled Sentence; Pic-Sent = Write a Sentence Based on a Picture; E-mail = Respond to an E-mail; IM = Instant Messaging; Blog Post = Blog Post; Op = Write an Opinion.

ITEM PROTOTYPING USABILITY STUDY

After the prototype pilot tasks had been selected, a small-scale usability study was undertaken. The usability study focused on speaking and writing tasks because they required navigating a computer interface and included task types that were likely to be unfamiliar to the testing population. The aims of the study were threefold: (a) to understand how beginning to intermediate learners of English would react to the new item types; (b) to evaluate the clarity of item directions; and (c) to understand potential challenges that learners of English, particularly beginning to intermediate learners, may have in navigating a computer-based assessment. By conducting a small-scale usability study, test designers were able to refine the new item types for speaking and writing administration to a larger number of pilot participants in the fall.

The usability study was conducted at the ETS headquarters in New Jersey in August 2017. Four English language learners of beginning to intermediate proficiency each completed approximately 20 speaking and writing tasks. Immediately after completing the tasks, the participants were interviewed. For all of the structured interviews, an interpreter was available to provide real-time translation for the participant and interviewer.

The feedback to the new item types was generally positive. Participants reported that they understood item directions. However, some issues were raised regarding navigating from screen to screen and the difficulty of some of the audio stimulus components of two item types, Listen and Retell and Respond to Questions with Information Provided.

Based on this feedback, variants of these two speaking tasks were included in the pilot forms: One version included a transcript of the audio on screen so that participants could read along with the audio, and the other version did not.

PILOT TESTING

In September 2017, pilot tests were administered to 464 participants from Brazil ($n = 84$), Japan ($n = 257$), Korea ($n = 57$), and Taiwan ($n = 66$). The pilot administration included the assessment of all four skills (i.e., listening, reading, speaking, and writing). For the skills of listening and reading, one pilot form was created consisting of 50 multiple-choice listening items followed by 50 multiple-choice reading items. For the skills of speaking and writing, two pilot forms were created. One pilot form consisted of 10 speaking tasks followed by nine writing items, and another pilot form consisted of nine speaking tasks followed by seven writing items. The pilot forms were intended not to be draft operational forms but rather to produce information about the performance of tasks and to be as similar as possible to one another in overall administration time.

The listening and reading pilot form was paper based, and the speaking and writing pilot forms were computer based. All 464 participants took the listening and reading pilot form, and 436 participants were randomly assigned to one of the two of the speaking and writing pilot forms.

Results of Pilot Testing

The goals of pilot testing for listening and reading and for speaking and writing were somewhat different. Because there were existing TOEIC Bridge Listening and Reading assessments, the TOEIC program and the test designers were concerned that the overall difficulty of redesigned TOEIC Bridge Listening and Reading tests be approximately the same as the existing assessments. That is, they did not want the new assessments to be so difficult that they would be discouraging for the current test-taking population nor so easy as to not give that population meaningful information. To that end, the listening and reading pilot forms included a subset of items from the existing assessments so that meaningful comparisons could be made between the old item types and the new. Because there were no existing TOEIC Bridge Speaking and Writing assessments, the comparative difficulty of the tasks was not a concern, and the purpose of the pilot was to collect information on whether or not beginning to intermediate learners understood the tasks and produced responses that could be meaningfully and reliably scored by human raters and whether the draft rubrics were as useful as possible.

After finishing the speaking and writing pilot test, test takers in Japan ($n = 30$) and Brazil ($n = 5$) completed surveys administered in their local language that asked them to provide feedback on the usability of the test (e.g., clarity of directions, adequacy of preparation and response time), various perceptions of the test (e.g., authenticity, difficulty), and task-specific questions (e.g., usefulness of a visual stimulus). Although a majority of participants indicated that the English directions were not difficult to understand for all of the tasks, a relatively high proportion (>40%) indicated that the directions for several of the speaking test tasks were difficult to understand (Ask, Invite, Request; Respond to Questions With Information Provided). In almost all instances, a majority of participants believed the preparation and response time provided for pilot test tasks was "OK," although a larger proportion (>40%) believed preparation and response times were too short for several speaking and writing tasks (Ask, Invite, Request; Respond to Questions With Information Provided; Give a Reason). Participants' perceptions of task difficulty and authenticity largely aligned with expectations, as tasks designed to target higher levels of proficiency were viewed as more difficult. Finally, task-specific questions helped identify features that could be refined for the next phase of testing.

Pilot Results for Listening

Overall, the results of the pilot for the listening prototype tasks were positive. All of the prototype items performed within the range of acceptable reliability and difficulty. The following items were of particular note:

- The Four-Picture items were comparable in difficulty to the easiest items on the existing TOEIC Bridge Listening test and differentiated among students at the beginning level.
- The Question-Response items were piloted in two versions. In one, similar to an existing TOEIC Bridge Listening item type, the options were audio only; in the other, the options were presented as audio and as text in the test book. The two versions of the task performed similarly, and it was decided to include the text and audio version in field study.

- The pilot Short Conversation two-item sets were not obviously more difficult than Short Conversation items from the existing TOEIC Bridge Listening test with one item per stimulus.
- The pilot Short Talk (monologue) two-item sets were not obviously more difficult than Short Talks from the existing TOEIC Bridge Listening test with one item per stimulus.

Based on the results of the pilot and in consultation with the program's psychometricians, Table 5 shows the form blueprint that was developed for the listening test.

TABLE 5

Redesigned TOEIC Bridge Listening Field Test Blueprint

Test part	Question type
Part 1	Four Pictures 6 questions
Part 2	Question-Response 20 questions
Part 3	Conversations 5 sets, 2 items per set, 10 item
Part 4	Talks 6 sets, 2–3 items per set, 14 items

Pilot results for reading. As with listening, the analysis of item difficulty and discrimination in the reading pilot was encouraging. The most relevant findings included the following:

- The Four-Picture reading items were of similar difficulty to the Four-Picture listening items.
- The pilot Sentence Completion items covered a wide range of difficulty.
- The Text Completion items were more difficult on average than the Sentence Completion items.
- The new genres of stimuli for reading sets—text messages, FAQs, website material with reader comments—all performed well, were in a similar range of difficulty as the legacy TOEIC Bridge Reading sets and represented a more up-to-date range of real-world texts.

Based on the results of the pilot and in consultation with the relevant psychometricians, Table 6 shows the form blueprint that was developed for reading.

TABLE 6

Redesigned TOEIC Bridge Reading Field Test Blueprint

Test part	Question type
Part 1	Sentence Completion 15 items
Part 2	Text Completion 5 sets, 3 items per set, 15 items
Part 3	Reading Comprehension 8 sets, 2–3 items per set, 20 items

Scoring the Constructed Responses From Pilot Testing

For the speaking and writing task types, rubric refinement was an integral component of test design. The test design team developed rubrics for evaluating spoken and written responses in tandem with designing the new task types. Each piloted task type was accompanied by its own holistic rubric, with 0 to 3 or 0 to 4 score points per rubric. Once the constructed responses from the pilot tests were available, the test design team analyzed the spoken and written responses to hone the rubrics that had been created during the task design phase.

The rubrics were further refined during the rangefinding process by a group of senior-level test developers. The three or more members of the rangefinding team each applied the rubrics to score between 5% and 10% of responses for each piloted item. The team members independently assigned scores to pilot test responses. Disagreements about scores were resolved by discussing team members' rationales for scoring until consensus among the team members was reached. Based on these discussions, the rangefinding team revised rubrics for clarity so as to minimize potential confusion that might lead to low agreement among raters. Samples for rater training were selected through the rangefinding process, and these trainings samples were annotated with scoring rationales.

Two main considerations guided the revision of rubrics during pilot scoring: to make the rubrics as responsive as possible to actual observed differences in performance among the pilot population and to make the rubrics easy to use for raters. The team decided that each task type required its own rubric with different specific details of task completion included. However, because individual operational raters would almost certainly be rating different task types in one scoring session, the team felt that the rubrics would be easiest to use if they were as parallel in construction and wording as possible. To that end, the team decided that each rubric should be structured with a general and repeated statement of task completion for each score point and the details of task-specific completion in bullets below the heading. Figures 6 and 7 show how this was put in practice for a speaking task (Describe a Picture) and a writing task (Blog Post).

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following. • The response describes the important parts of the picture. • Delivery is generally intelligible but may require some listener effort. • Choice of vocabulary and use of structures are appropriate, though minor errors that do not affect meaning may be present.
2	The response is partially effective at addressing the prompt and exhibits one or more of the following. • The response is connected to the picture, but the meaning is obscured in places. • Delivery is sometimes unintelligible and requires listener effort. • Choice of vocabulary and use of structures are limited, and errors interfere with comprehensibility.
1	The response does not effectively address the prompt and exhibits one or more of the following. • The response is not connected to the picture. • Delivery is mostly unintelligible. • Severely limited choice of vocabulary and/or use of structures obscure meaning. • The response may consist of isolated words or phrases.
0	No response OR no English in the response.

Figure 6. Sample TOEIC Bridge Speaking rubric. Scoring guide for Describe a Picture.

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following. • The response fully addresses the topic and task. • Overall meaning is clear, though minor grammatical errors that do not obscure meaning may be present. • The story is told in a logical sequence, and any connectors are used appropriately. • The choice of vocabulary is appropriate to the topic of the prompt.
2	The response is partially effective at addressing the prompt and is marked by one or more of the following. • The response partially addresses the topic or partially completes the task. • Use of language structures contributes to meaning, though grammatical errors may occasionally obscure meaning. • The logical sequence of the story is mostly clear. • The choice of vocabulary is sometimes limited or inappropriate to the topic.
1	The response does not effectively address the prompt and exhibits one or more of the following. • The response is unsuccessful at addressing the topic or task, though it may contain some related words. • The response is off topic or seriously underdeveloped. • Frequent and serious grammatical errors interfere with the comprehensibility of most of the response. • The choice of vocabulary is limited (use of isolated words), inaccurate, and/or it relies on repetition of the prompt in most of the response.
0	No response OR no English in the response.

Figure 7. Sample TOEIC Bridge Writing rubric. Scoring guide for Blog Post.

After rangefinding and rubric refinement, a larger group of 17 test developers with extensive experience in scoring constructed responses for assessments of English language proficiency was trained using the samples selected during rangefinding. All pilot test responses were scored by at least two raters. Exact agreement rates by item ranged between 64% and 93%, with most items (58%) ranging between 70% and 79% exact agreement. When scores were discrepant—that is, when raters disagreed by more than one score point—a member of the design team reviewed the response and provided the final, adjudicated score. Fewer than 1% of pilot test responses caused discrepant scores.

Pilot Results for Speaking

After scoring the pilot speaking responses, test developers evaluated the effectiveness of each of the pilot task types. Three task types were considered for operational use with very few modifications:

- Two different versions of the Read a Short Text Aloud task type were piloted, one with a 20- to 25-word stimulus and one with a 30- to 45-word stimulus. Pilot results indicated that the longer stimulus provided more differentiation among responses.
- Directions for the Describe a Picture item type were elaborated to give test takers more explicit help in creating a response. Also, detailed specifications for the picture content were created to ensure that the task was approachable for beginning to intermediate learners.
- Directions for Tell a Story were elaborated to give test takers more support in creating full responses.

Three task types were revised post pilot:

- The Listen and Retell task type was revised to be less challenging than the pilot version by reducing the number of important points the test taker should include in the response.
- The Answer Questions With Information Provided item type was revised to allow for question formation as well as suggestions.
- Aspects of the Opinion and Give Two Reasons task types were combined in a post-pilot version of the task intended to elicit opinion. In this task, the test taker is given a table or list with short, telegraphic information about two options, alternatives, or choices that can be compared and contrasted and uses this information to express an opinion, make a recommendation, give advice, or state a preference and support it.

In addition to the task modifications described above, the test developers gave the speaking task types the names that would be used publicly when describing the test contents. The goal was to give names to items that were connected to the communication goal being assessed and to be as transparent as possible. Table 7 presents the form blueprint that was created for the speaking field test forms.

TABLE 7

Redesigned TOEIC Bridge Speaking Field Test Blueprint

Question type	Description of questions	Timing
Read a Short Text Aloud (two questions)	The test taker reads a short paragraph aloud.	Prepare 25 seconds Respond 30 seconds
Describe a Photograph (two questions)	The test taker describes a photograph presented on the computer screen.	Prepare 30 seconds Respond 30 seconds
Listen and Retell	The test taker listens to a talk on an everyday social or workplace-related situation and communicates the main ideas and important details of the talk to a third person.	Prepare 10 seconds Respond 30 seconds
Short Interaction	The test taker reads a brief text and asks for information and/or makes requests, offers, suggestions, and invitations.	Prepare 20 seconds Respond 30 seconds
Tell a Story	The test taker narrates a story based on a picture sequence.	Prepare 45 seconds Respond 60 seconds
Make and Support a Recommendation	The test taker reads a short, simple piece of information showing options, alternatives, or choices that can be compared and contrasted and uses this information to express an opinion, make a recommendation, give advice, or state a preference and support it.	Prepare 45 seconds Respond 60 seconds

Note. Total test time, with directions and audio stimuli, 12 minutes, 45 seconds.

Pilot Results for Writing

For both speaking and writing, the test designers were also members of the rating team, and thus for each item type, they had rated many, if not all, the responses. Based on their own experience rating and on the group's ratings for each pilot task type item, designers evaluated the pilot tasks. In general, they focused on these questions:

- Did a significant number of pilot participants misunderstand the task and produce responses unrelated to the targeted communication goal?
- Were a reasonable number of responses successful according to the a priori criteria represented in the draft rubric (i.e., was the task too difficult for most of the pilot population)?

- Did the responses allow differentiation between pilot participants? Did the rubric produce a range of ratings?
- Did raters find the rubrics easy to use, and could they score efficiently?
- Did two or more tasks focus on the same communication goal? Did one produce better evidence for the relevant claim(s)?

Test designers made no changes to two of the pilot writing tasks: Write a Sentence Based on a Picture and Blog Post.

After evaluating the pilot responses, test designers made the following changes to task designs:

- In the pilot version of the Scrambled Sentence task, test takers were required to type out the unscrambled sentence. The pilot responses made clear that many participants' limited keyboarding skills were interfering with their ability to respond. The task was redesigned so that test takers can drag and drop the sentence elements into the appropriate order.
- Elements of different pilot tasks were combined to create a task that assessed the communication goals of providing basic information, asking for basic information, and making simple requests, offers, suggestions, or invitations. A short message (20–45 words) prompts the response. The stimulus includes two questions that require two or three details in response. The stimulus is preceded by a lead-in enumerating the communication goals test takers should produce (i.e., ask one question and make one suggestion in response to the message).
- Features of the E-mail pilot item and the Opinion pilot item were combined in the Respond to an Extended Message item type. The stimulus is presented as an e-mail inquiry and asks two separate questions that are thematically related. The first question requires straightforward information as a response. The second question requires a brief opinion.

In addition to the task modifications described above, test developers also gave the task types the names that would be used publicly in referring to the tests. The goal was to use task names that were as transparent as possible. After analysis of the pilot responses, a form blueprint for the writing field test was developed (see Table 8).

TABLE 8

Redesigned TOEIC Bridge Writing Field Test Blueprint

Question type	Description of task	Timing
Build a Sentence (three questions)	Arrange a set of words or phrases in the appropriate order to form a grammatically correct sentence or question.	Respond 3 minutes for three questions
Write a Sentence (three questions)	The test taker sees a photograph with two key words or phrases below it. Using both of the key words or phrases, the test taker writes one grammatically correct sentence that describes the picture. The test taker can change the forms of the words and can use them in any order.	Respond 3 minutes for three questions
Respond to a Brief Message	The test taker reads a brief message (e.g., an instant message) from an acquaintance and composes a response that completes two communication goals (e.g., providing basic information; asking for basic information; or making simple requests, offers, suggestions, and invitations).	Respond 8 minutes
Write a Narrative	The test taker reads a prompt on the screen that specifies a category of past experience and writes a narrative based the prompt.	Respond 8 minutes
Respond to an Extended Message	The test taker reads an extended message (e.g., an e-mail) from a person or entity, and writes a response.	Respond 10 minutes

Note. Total test time, with directions: 33 minutes.

FIELD STUDY

After task types were selected for use in the field test, task-level specifications could be finalized. Written specifications for each of the tasks had been evolving since the initial modeling phase before piloting. In preparation for the field test, task-level specifications were needed to generate relatively large (more than 100 items) usable pools of listening and reading items from which forms could be constructed. Fewer items were needed for the speaking and writing field tests than for the listening and reading field tests. However, because speaking and writing forms are not equated, it was important that the items representing task types across forms be as similar in construction as possible. The most effective way to create similar items is to create and use very explicit task-level specifications. Task-level specifications, in the form of item shells, included the claim about the test taker for which the task provides evidence, a description of the task, the scoring criteria, the fixed elements of the task (those aspects of the task that are the same in every instance), and the variable elements (the things that change to make two items representing the same task different from one another). Figure 8 presents an item shell for the redesigned TOEIC Bridge Write a Sentence task type.

Claims	Task	Fixed elements	Variable elements	List of variants
The test taker can write sentences to describe people, objects, places, and activities.	Test takers will see a photograph with two key words or phrases below it. Using both of the key words or phrases, test takers will write one grammatically correct sentence that describes the picture. The form of the key words or phrases can be manipulated. **Scoring:** • 0–3 holistic scale • criteria: use of key words, consistency with picture, correctness of grammar.	**Directions** Write ONE sentence based on the picture. Use the TWO words or phrases under the picture. You can change the forms of the words, and you can use them in any order. You have 90 seconds to write. **Stimuli:** **One color photograph** The picture should not rely on the text and must have a clear focus. A description of the photograph should not require specialized vocabulary. **Two key words or phrases** Key words or phrases are located below the photograph, and key words require no or limited form transformation and are presented in canonical word order (SVO). **Response:** • Response length: a sentence • Response time: 90 seconds per item	**1. Photograph contexts** **2. Key words/phrases** **3. Actions**	1. **Examples of contexts:** activities, dining out, entertainment, family and friends, business, health, housing, offices, news, school, shopping, travel 2. **Parts of speech:** adjective adverb coordinating conjunction noun preposition verb 3. **Examples of actions:** Activities (participating in hobbies, playing sports) Dining (eating, drinking, ordering) Entertainment (playing music, visiting museums) Health (visiting doctor, attending a class) Household tasks (cleaning, repairing, moving, cooking) Shopping (buying groceries, selecting clothes) Travel (taking trains, waiting in airports, buying tickets, checking schedules, looking at maps or documents)

Figure 8. Redesigned TOEIC Bridge Write a Sentence item shell.

Field Test Usability Study

Prior to the field test, researchers again conducted cognitive interviews to evaluate the usability of the refined computer-based speaking and writing field test forms for low-proficiency adult learners of English ($n = 9$). Again, the study identified general usability issues (e.g., some participants began speaking before recording started, timing directions for writing tasks were sometimes misunderstood) and item-specific concerns (e.g., some participants did not notice the two words under the picture for the Picture Sentence task in the writing test) and provided suggestions for remediating these issues and concerns. This phase of development corresponds to the presentation model identified in the ECD framework.

Field Test Results for Listening and Reading

In May–June of 2017, two parallel forms of the redesigned TOEIC Bridge Listening and Reading assessments were administered to a total of 2,484 test takers in six countries (Japan, Korea, Taiwan, Brazil, Mexico, and Colombia). Each participant took either Form 1 ($N = 1{,}220$) or Form 2 ($N = 1{,}264$). The results of the field test confirmed the listening and reading test designs. Item types and individual items were within the expected range of difficulty for the TOEIC Bridge population, and the assessments reliably distinguished four levels of performance for both listening and reading, which served as the basis for the development of level descriptors (see below). Further detailed discussion of the field test results for the listening and reading assessments are available in Lin et al. (2019).

Field Test Results for Speaking and Writing

The May–June field test also included administration of two parallel computer-based forms of speaking and writing assessments. The same six countries participated (Form 1, $N = 1{,}228$; Form 2, $N = 1{,}174$). Assessment developers and experienced TOEIC test Speaking and Writing raters used the rubrics finalized as part of the post-pilot item specification process to identify benchmark responses, training responses, and calibration sets for each of the speaking and writing task types. TOEIC test raters, who are expected to be the operational redesigned TOEIC Bridge test raters, were trained and scored the bulk of the field test responses. All field test responses were double scored. A detailed discussion of the speaking and writing field test results is available in Lin et al. (2019).

Rater Survey

Raters who scored the speaking field test and writing field test were invited to participate in an online survey. Raters of the speaking test ($n = 156$) and writing test ($n = 41$) who responded were asked to indicate (using 5-point Likert-type scales) the extent to which they agreed or disagreed with the statements "It was easy to form judgments" and "I felt confident in my scores" for each of the scoring rubrics they used. Overall, a high percentage of raters agreed (i.e., agree or strongly agree) with the statements for each of the scoring rubrics. The percentage of raters who agreed with these statements ranged from 64% to 87% across scoring rubrics for the speaking test and 71% to 97% for the writing test. Raters were also asked to estimate the approximate percentage of test takers who (a) did not seem to

have adequate time to provide responses and (b) did not seem to understand task directions. A large percentage of raters (> 75%) indicated that most test takers had adequate time to provide a response and seemed to understand directions, with several exceptions. Results of the survey suggested that at least 35% of raters believed that test takers could use more time to provide responses for the Tell a Story task on the speaking test and for the Blog and E-mail tasks on the writing test. At least 35% of raters also believed that test takers seemed to have some confusion about the directions for the Give Reasons task on the speaking test and for the Blog task on the writing test.

Test-Taker Survey for Speaking and Writing

Test takers who completed the speaking and writing field test were invited to complete a follow-up survey in their local language, and responses were obtained from participants in Brazil (n = 268), Colombia (n = 18), Japan (n = 1251), Korea (n = 323), Mexico (n = 48), and Taiwan (n = 333). In the surveys, participants again provided feedback on the usability of the test (e.g., clarity of directions, adequacy of preparation and response time) and their various perceptions of the test (e.g., authenticity, difficulty) and answered task-specific questions (e.g., usefulness of a visual stimulus). A majority of participants indicated that the English directions were not difficult to understand for most tasks (ranging from 62% to 87% across tasks and forms for speaking and 74% to 87% for writing) with the exception of the Short Interaction task on the speaking test (48%). For most tasks, a majority of participants indicated that the preparation and/or response times provided were good (ranging from 61% to 74%) with the exception of several speaking test tasks where at least 30% of participants believed preparation times were insufficient (Short Interaction, Tell a Story, Listen and Retell) and/or response times were insufficient (Tell a Story). Participants' perceptions of task difficulty and authenticity were largely aligned with expectations, as tasks designed to target higher levels of proficiency were viewed as more difficult. Finally, participant responses to task-specific questions indicated that features of the field test tasks were largely functioning as intended.

Final Adjustment to Item Presentations

Based on feedback from the field test usability study, the rater survey, the survey of speaking and writing test takers, and the field test results, final adjustments were made to speaking and writing task types.

The order of the speaking items was changed to reflect increasing difficulty of tasks as observed in the field test results (see Figure 9).

	Field test		Operational test
1	Read a Short Text Aloud	1	Read a Short Text Aloud
2	Read a Short Text Aloud	2	Read a Short Text Aloud
3	Describe a Photograph	3	Describe a Photograph
4	Describe a Photograph	4	Describe a Photograph
5	Short Interaction	5	Listen and Retell
6	Tell a Story	6	Short Interaction
7	Listen and Retell	7	Tell a Story
8	Make and Support a Recommendation	8	Make and Support a Recommendation

Figure 9. Redesigned TOEIC Bridge Speaking field test and operational test item order.

Adjustments were also made to the preparation and response times of several speaking and writing task types (see Tables 9 and 10). With these changes, the task design and test blueprint design processes were complete.

TABLE 9

Redesigned TOEIC Bridge Speaking Field Test and Operational Test Item Timing

Speaking		Field test prep (seconds)	Field test response (seconds)	Operational prep (seconds)	Operational response (seconds)
1	Read a Short Text Aloud	20	30	25[a]	30
2	Read a Short Text Aloud	20	30	25[a]	30
3	Describe a Photograph	30	30	30	30
4	Describe a Photograph	30	30	30	30
5	Listen and Retell	10	30	15[a]	30
6	Short Interaction	20	30	30[a]	30
7	Tell a Story	45	45	45	60[a]
8	Make and Support a Recommendation	45	60	60[a]	60

[a] Cells shaded in gray indicate a change from field test to operational test timing.

TABLE 10

Redesigned TOEIC Bridge Writing Field Test and Operational Test Item Timing

Writing		Field test response (seconds)	Operational response (seconds)
1	Build a Sentence	1	1
2	Build a Sentence	1	1
3	Build a Sentence	1	1
4	Write a Sentence	1	1.5[a]
5	Write a Sentence	1	1.5[a]
6	Write a Sentence	1	1.5[a]
7	Respond to a Brief Message	8	8
8	Write a Narrative	8	10[a]
9	Respond to an Extended Message	10	10

[a] Cells shaded in gray indicate a change from field test to operational test timing.

PROFICIENCY DESCRIPTORS

The last assessment design task for test developers in test design prior to moving to support the operational test was the creation of proficiency descriptors for each of the domains of listening, reading, speaking, and writing. One of the business requirements for the redesigned TOEIC Bridge tests was to provide meaningful feedback to teachers and learners in the form of proficiency descriptors. With this end in mind, the test development team created descriptions of what test takers can do using English. The team used the following to inform the construction of the proficiency descriptors: results of the initial ECD test design process, recommendations from psychometric analysis of the field test, findings from a mapping study, results from a survey of field test participants, and for speaking and writing tasks, review of the field test responses.

The test development team began this phase of the project by revisiting the findings from the initial domain analysis and modeling and also reviewing the domain definitions and the TLU. The purpose of returning to the initial domain analysis and modeling was to ensure the resulting descriptors would be aligned with the TLU/domain definition, in accordance with ECD. Next, the test development team revisited the results from the task modeling phase of the project, reviewing the task specifications, the claims for each task type derived from the domain model, and the rubrics for the speaking and writing tasks, again in keeping with ECD to ensure alignment with the domain model.

Psychometric analysis of field test scores indicated test takers could be grouped into four distinct score ranges for each skill assessed (Lin et al., 2019). For each of these score ranges, the most common patterns in field test participants' performances were identified and examined by the test development team. For listening and reading, the average percent of items answered correctly by task type was used to identify patterns, and for speaking and writing, the average item score by task type was used. To draft the descriptors, the test development team linked these patterns back to the task claims from the domain model, and for speaking and writing, the patterns were linked back to the rubrics.

To validate the proficiency descriptors, the test development team compared the drafts to the results of a study mapping the TOEIC Bridge field test scores onto international standards of language proficiency (Schmidgall et al., 2019). The test development team also compared the proficiency descriptors to the results of a can-do survey conducted with the field test participants (Schmidgall et al., 2019). For speaking and writing, responses from the field test were reviewed and compared to the drafted proficiency descriptors. Finally, the validated drafts were reviewed by subject matter experts, researchers, product managers, marketing, and ETS partners prior to the finalization of the proficiency descriptors.

CONCLUSION

This paper described the process of developing the task types, presentation models, and assembly models of the four parts of the redesigned TOEIC Bridge tests: listening, reading, speaking, and writing. Design for each part began with consideration of the business requirements for the assessment program and well-defined domain definitions. Task models, or prototype tasks, were developed, tried out in cognitive labs, and piloted. The results of the pilot informed modifications of the prototype tasks and tentative selection of task types for the field test. Further cognitive labs preceded the field test. Field test data, supplemented by surveys of the raters of constructed response tasks, were used to set the final specifications for all task types and the operational assembly models.

ECD is often presented as a systematic approach to test development that emphasizes how a test may be used to elicit evidence of the ability to be assessed from test-taker performance. Another benefit of such a systematic approach is the collection of documentation throughout the test development process to justify design decisions by test developers. With this in mind, we described the various sources of data we obtained throughout the test development process (e.g., cognitive labs, surveys, item performance) and how each influenced item and test design decisions.

REFERENCES

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

Mislevy, R. J., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). Routledge.

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**

APPENDIX A. REDESIGNED TOEIC BRIDGE SPEAKING AND WRITING TESTS SCORING GUIDES

TABLE A1

Redesigned TOEIC Bridge Speaking Test: Scoring Guide for Read a Short Text Aloud

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The entire text is read aloud AND other-language influence does not affect overall intelligibility. • At the word level, pronunciation is mostly intelligible, but there may be some minor lapses. • At the phrase and sentence level, intonation and stress are mostly appropriate, though the response may include some lapses and/or some other language influence.
2	The response is partially effective at addressing the prompt and exhibits one or more of the following: • A section of the text is NOT read aloud. • At the word level, pronunciation is sometimes unintelligible and requires some listener effort. • At the phrase and sentence level, intonation and stress are somewhat appropriate, but lapses and/or other language influence are present.
1	The response does not effectively address the prompt and exhibits one or more of the following: • Most of the text is NOT read aloud. • The response is off topic. • Speech is mostly unintelligible and requires significant listener effort to understand.
0	No response OR no English in the response.

TABLE A2

Redesigned TOEIC Bridge Speaking Test: Scoring Guide for Describe a Photograph

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The response describes the important parts of the picture. • Delivery is generally intelligible but may require some listener effort. • Choice of vocabulary and use of structures are appropriate, though minor errors that do not affect meaning may be present.
2	The response is partially effective at addressing the prompt and exhibits one or more of the following: • The response is connected to the picture, but the meaning is obscured in places. • Delivery is sometimes unintelligible and requires listener effort. • Choice of vocabulary and use of structures are limited, and errors interfere with comprehensibility.
1	The response does not effectively address the prompt and exhibits one or more of the following: • The response is not connected to the picture. • Delivery is mostly unintelligible. • Severely limited choice of vocabulary and/or use of structures obscure meaning. • The response may consist of isolated words or phrases.
0	No response OR no English in the response.

TABLE A3

Redesigned TOEIC Bridge Speaking Test: Scoring Guide for Listen and Retell

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The response adequately communicates two main ideas of the talk. • Delivery is generally intelligible but may require some listener effort. • The choice of vocabulary and use of structures fulfills the demands of the task.
2	The response is partially effective at addressing the prompt and exhibits one or more of the following: • The response communicates at least one of the main ideas of the talk, but the response is incomplete or one of the main ideas is inaccurate. • Delivery is sometimes unintelligible and/or sometimes requires listener effort. • The choice of vocabulary and use of structures are limited and interfere with overall comprehensibility.
1	The response is not effective at addressing the prompt and exhibits one or more of the following: • The response may consist of isolated words or phrases or is off topic. • Delivery is mostly unintelligible and/or requires significant listener effort.
0	No response OR no English in the response.

TABLE A4

Redesigned TOEIC Bridge Speaking Test: Scoring Guide for Short Interaction

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The required task (request, offer, suggestion, invitation) and details from the prompt are successfully communicated. • Delivery is generally intelligible but may require some listener effort. • Choice of vocabulary and use of structures fulfill the demands of the prompt. • Minor errors do not obscure overall meaning.
2	The response is partially effective at addressing the prompt and exhibits one or more of the following: • The required task (request, offer, suggestion, invitation) is attempted and relevant details are included, but the response is incomplete. • Delivery is sometimes unintelligible and requires listener effort. • Choice of vocabulary and use of structures are limited and sometimes affect meaning.
1	The response does not effectively address the prompt and exhibits one or more of the following: • The required task (request, offer, suggestion, invitation) is not attempted. • The response is off topic. • Delivery is mostly unintelligible. • The choice of vocabulary and use of structures are severely limited (use of isolated words). Meaning is obscured.
0	No response OR no English in the response.

TABLE A5

Redesigned TOEIC Bridge Speaking Test: Scoring Guide for Tell a Story

Score	Response description
4	The response successfully addresses the prompt and is marked by all of the following: • The presentation of a cohesive story is based on the main features of the picture sequence. • The delivery is generally intelligible and does not interfere with meaning. • The choice of vocabulary and use of structures and connecting language fulfill the demands of the task.
3	The response is mostly effective at addressing the prompt. The response consists of a mostly cohesive story based on the picture sequence, although part of the story may be incomplete or unclear because • delivery is occasionally unintelligible or requires listener effort, and/or • choice of vocabulary and use of structures and connecting language occasionally interfere with overall comprehensibility.
2	The response is partially effective at addressing the prompt. Parts of the picture sequence may be conveyed, but the story is mostly incomplete or unclear because • parts of the narrative sequence are missing, and/or • unintelligible delivery interferes with parts of the narrative sequence, and/or • choice of vocabulary and use of structures and connecting language are limited and interfere with overall comprehensibility.
1	The response does not effectively address the prompt and exhibits one or more of the following: • The response may be only tangentially related to the picture sequence. • The response is off topic. • Delivery is mostly unintelligible and/or requires significant listener effort throughout. • Choice of vocabulary and use of structures are severely limited (use of isolated words) and may significantly interfere with comprehensibility.
0	No response OR no English in the response.

TABLE A6

Redesigned TOEIC Bridge Speaking Test: Scoring Guide for Make and Support a Recommendation

Score	Response description
4	The response successfully addresses the prompt and is marked by all of the following: • All of the information from the stimulus is clearly and appropriately communicated. • A recommendation is made and is adequately supported. • Delivery is intelligible but may require some listener effort. • Choice of vocabulary and use of structures fulfill the demands of the task.
3	The response is mostly effective at addressing the prompt and exhibits one or more of the following: • Most of the information provided in the stimulus is appropriately communicated. • A recommendation is made, but support is limited. • Delivery is mostly intelligible, though listener effort is required at times. • Choice of vocabulary and use of structures are fairly effective, though they interfere with comprehensibility at times.
2	The response is partially effective at addressing the prompt and exhibits one or more of the following: • Information from the stimulus is presented, but only limited original language is used. • A recommendation may be made, but support is missing. • Delivery is sometimes unintelligible and may require listener effort. • Choice of vocabulary and use of structures are limited and often obscure meaning.
1	The response does not effectively address the prompt and exhibits one or more of the following: • Information from the stimulus is missing. • The response is off topic. • The response is limited to reading aloud the prompt, the directions, or the information in the stimulus without adding original language. • Delivery may be mostly unintelligible and require listener effort. • The response contains errors that obscure meaning most of the time.
0	No response OR no English in the response.

TABLE A7

Redesigned TOEIC Bridge Writing Test: Scoring Guide for Write a Sentence

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The response is consistent with the picture. • Forms of both key words are used appropriately in one sentence. • No grammatical errors are present.
2	The response is partially effective at addressing the prompt and is marked by one or more of the following: • The response is consistent with the picture. • Forms of both key words are present, though they may be in different sentences, or the form of the word(s) may not be accurate. • Minor grammatical errors are present but do not obscure meaning.
1	The response does not effectively address the prompt and exhibits one or more of the following: • The response is not consistent with the picture. • The response omits one or both key words. • Significant grammatical errors are present that obscure meaning.
0	No response OR no English in the response.

TABLE A8

Redesigned TOEIC Bridge Writing Test: Scoring Guide for Respond to a Brief Message

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The response is clear and fully responsive to the stimulus message. • The choice of vocabulary is effective, with allowance for slight inaccuracies that do not obscure meaning. • The use of grammatical structures fulfills the demands of the task. A few minor errors may be present but do not obscure meaning.
2	The response is partially effective at addressing the prompt and is marked by one or more of the following: • The response attempts both tasks, though one or both tasks may not be successful. • The response is somewhat clear. • Errors in use of grammar and choice of vocabulary appear throughout the response and may occasionally obscure meaning.
1	The response does not effectively address the prompt and exhibits one or more of the following: • The response attempts only one of the required tasks, or the response does not attempt any of the required tasks. • The response contains very little original language and may contain words or phrases related to or copied from the stimulus. • The response is mostly incoherent. • Errors in grammar and usage frequently obscure meaning.
0	No response OR no English in the response. There may be keystroke characters that convey no meaning.

TABLE A9

Redesigned TOEIC Bridge Writing Test: Scoring Guide for Write a Narrative

Score	Response description
3	The response successfully addresses the prompt and is marked by all of the following: • The response fully addresses the topic and task. • Overall meaning is clear, though minor grammatical errors that do not obscure meaning may be present. • The story is told in a logical sequence, and any connectors are used appropriately. • The choice of vocabulary is appropriate to the topic of the prompt.
2	The response is partially effective at addressing the prompt and is marked by one or more of the following: • The response partially addresses the topic or partially completes the task. • Use of language structures contributes to meaning, though grammatical errors may occasionally obscure meaning. • The logical sequence of the story is mostly clear. • The choice of vocabulary is sometimes limited or inappropriate to the topic.
1	The response does not effectively address the prompt and exhibits one or more of the following: • The response is unsuccessful at addressing the topic or task, though it may contain some related words. • The response is off topic or seriously underdeveloped. • Frequent and serious grammatical errors interfere with the comprehensibility of most of the response. • The choice of vocabulary is limited (use of isolated words), inaccurate, and/or it relies on repetition of the prompt in most of the response.
0	No response OR no English in the response.

TABLE A10

Redesigned TOEIC Bridge Writing Test: Scoring Guide for Respond to an Extended Message

Score	Response description
4	The response successfully addresses the prompt and is marked by all of the following: • The requested information, opinion, and support for that opinion are present and elaborated clearly. • The response is well organized, well developed, and coherent. • Tone and register are appropriate for the intended audience. • Command of the language demonstrates appropriate use of structures, syntactic variety, and vocabulary, though there may be minor lexical or grammatical errors.
3	The response is mostly effective at addressing the prompt and is marked by one or more of the following: • The requested information, opinion, and support for that opinion are present, though they may not be clear or relevant. • Organization, development, and/or coherence are generally appropriate for the task. • Tone and register are not fully appropriate for the intended audience. • Command of the language demonstrates mostly appropriate use of structures, syntactic variety, and vocabulary, though some lexical and/or grammatical errors occasionally obscure meaning.
2	The response unsuccessfully addresses the prompt and is marked by one or more of the following: • Parts of the requested information, opinion, or support for that opinion are missing or inappropriate/incoherent. • Organization, development, and/or coherence is generally inappropriate for the task. • Tone and register are not appropriate for the intended audience. • Command of the language is limited. Use of structures, syntactic variety, and/or vocabulary obscure meaning.
1	The response fails to address the prompt and exhibits one or more of the following: • The requested information, opinion, and support for that opinion are not present. • The response is off topic. • Organization, development, and coherence are inadequate. • Serious and frequent errors in structure and vocabulary (use of isolated words) severely limit comprehensibility.
0	No response OR no English in the response.

APPENDIX B. REDESIGNED TOEIC BRIDGE PROFICIENCY DESCRIPTORS

REDESIGNED TOEIC BRIDGE LISTENING TEST PROFICIENCY DESCRIPTORS

The performance descriptors outline the types of general skills and abilities in understanding spoken English that are typical of test takers who have achieved similar scores. The descriptor associated with the score will help test takers understand the strengths and weaknesses of their listening ability in English. Each test taker will receive a description of listening proficiency in English on his or her score report.

Listening Score: Scaled Score 15

Test takers in this score range can understand a few words, very simple phrases, and some short sentences that are spoken clearly and very slowly. Some test takers can recognize individual words such as numbers or days of the week. Some test takers can understand highly predictable questions and statements when they are spoken one phrase at a time. Some test takers may be able to understand a limited range of sentences consisting of very simple grammatical structures and very common vocabulary on very familiar topics.

Listening Score Range: Scaled Score 16 to 25

Test takers in this score range typically can understand some slowly spoken words, simple phrases, and short sentences on familiar topics. Test takers in this score range can understand short pieces of speech that are spoken clearly and very slowly. Generally they can understand short formulaic phrases, simple sentences, and simple grammatical structures when common vocabulary is used. They can understand short, predictable messages and instructions on familiar topics. They can understand unconnected speech, one sentence at a time.

Listening Score Range: Scaled Score 26 to 38

Test takers in this score range typically can understand short spoken sentences and a limited range of grammatical structures; they can understand short conversational exchanges on familiar topics. Typically test takers in this score range can understand speech that is clear and slow. They can usually understand key words, formulaic phrases and expressions, and relatively short, sentence-length speech. Generally test takers in this score range can understand spoken language on familiar topics and routines. They understand simple descriptions and information about people, family, shopping, location, and employment. Most of the time, test takers can understand simple sentences and simple grammatical structures, and they may inconsistently understand some complex sentences and structures. Test takers in this score range can occasionally understand implied meanings.

Listening Score Range: Scaled Score 39 to 50

Test takers in this score range typically can understand short spoken conversations and monologues made up of connected sentences and some complex structures. They can understand some implied meaning and some abstract ideas. Typically test takers in this score range can usually understand a range of common vocabulary and some complex sentences and grammatical structures. Generally test takers in this score range can understand concrete topics and some abstract ideas related to work and other familiar contexts. In this score range, test takers understand some implied meanings and can connect facts in conversations and short spoken monologues with information in a short written text. They can comprehend formal and informal spoken language if the topics are relevant and familiar.

REDESIGNED TOEIC BRIDGE READING TEST PROFICIENCY DESCRIPTORS

The performance descriptors outline the types of general skills and abilities in understanding written English that are typical of test takers who have achieved similar scores. The descriptor associated with the score will help test takers understand the strengths and weaknesses of their reading ability in English. Each test taker will receive a description of reading proficiency in English on his or her score certificate.

Reading Score Range: Scaled Score 15 to 18

In this score range, test takers may succeed in identifying a limited number of words and phrases related to very familiar needs. The words that test takers are likely to identify are very common words and/or phrases that are strongly supported by context. Some test takers may understand simple instructions such as "Stop," "No exit," especially with illustrations to help understanding. Occasionally they may be able to understand material longer than a single phrase. Some test takers may only be able to recognize letters of the alphabet.

Reading Score Range: Scaled Score 19 to 33

Test takers in this score range can typically identify familiar words and phrases in very short texts written with very common vocabulary and basic grammatical structures. They can understand some simple language on familiar topics. Typically test takers in this score range can understand very short texts on familiar topics. They can understand some simple phrases and sentences, especially those supported by visual cues and common formats. For example, they can identify some details of written language on signs and in schedules. They may understand the overall meaning of simple texts by recognizing common words and phrases.

Reading Score Range: Scaled Score 34 to 44

Test takers in this score range can typically understand short texts written with common vocabulary and basic grammatical structures. They can understand simple language used to describe familiar topics. Typically test takers in this score range understand writing that is short and simple. They can understand the overall meaning of written language in a variety of formats such as e-mails, letters, and web pages.

They are developing familiarity with the basic organization of texts in English and can sometimes use this knowledge to support their understanding. They can usually understand both the overall meaning and the purpose of written communication on familiar topics such as family, shopping, and employment. Most of the time, test takers can understand simple sentences and simple grammatical structures, and they may occasionally understand a limited range of complex sentences.

Reading Score Range: Scaled Score 45 to 50

Test takers in this score range can understand short written texts in personal, public, and familiar workplace contexts and across a range of formats. Typically test takers in this score range can understand a variety of common texts such as web pages, letters, and articles written in formal and informal styles. They are familiar with the basic organization of short texts in English and can use this knowledge to support their understanding. They can understand vocabulary related to concrete topics as well as some abstract topics related to everyday life. They are familiar with a variety of grammatical structures and are developing the ability to understand complex sentences and structures. They can connect information across sentences. They can understand overall meaning, purpose, and many details. They can sometimes understand meaning that is implied rather than directly stated.

REDESIGNED TOEIC BRIDGE SPEAKING TEST PROFICIENCY DESCRIPTORS

Speaking Score Range: Scaled Score 15 to 22

Test takers in this score range are developing the ability to produce words and short phrases. Test takers in this score range can occasionally use simple words or phrases to identify people, objects, places, and activities that are highly familiar. They are developing the ability to read short texts aloud.

Speaking Score Range: Scaled Score 23 to 36

Test takers in this score range can typically use spoken English to perform very familiar and routine social interactions. They can use common and some high-frequency words and simple phrases, and they have limited control of simple structures. Listener effort is typically needed to understand the test taker's meaning due to issues with pronunciation, intonation, word stress, choice of vocabulary, and use of grammatical structures.

- Test takers in this score range can occasionally ask for and provide basic information.
- Test takers in this score range are developing the ability to describe people, objects, places, and activities.
- Test takers in this score range can sometimes express basic preferences, likes, and dislikes about very familiar topics.
- Test takers in this score range can occasionally give a basic description of simple and very familiar events.

Speaking Score Range: Scaled Score 37 to 42

Test takers in this score range can typically use spoken English to perform simple communication tasks involving familiar everyday activities, experiences, wants, and needs. They can use phrases, short sentences, and some longer sentences. They have some control over simple grammatical structures and vocabulary. At times, listener effort may be needed to understand the test taker's meaning due to occasional issues with pronunciation, intonation, word stress, choice of vocabulary, and use of grammatical structures.

- Test takers in this score range can typically ask for and provide simple and direct information.
- Test takers in this score range can usually give basic descriptions of people, objects, places, and activities, though meaning may be obscured at times.
- Test takers in this score range can typically sequence simple events to tell a story, but part of the story may be unclear. The test taker can use simple, linear connectors such as *and* or *then*.
- Test takers in this score range can typically ask and answer questions and make simple requests, offers, and suggestions, but attempts may be incomplete or unclear at times.
- Test takers in this score range can sometimes express a simple opinion or recommendation, but they may only be able to provide limited support for the recommendation.

Speaking Score Range: Scaled Score 43 to 50

Test takers in this score range can typically use spoken English to perform a variety of communicative tasks relevant to everyday life and the speaker's areas of interest. When needed, they can combine sentences to produce connected discourse. Their use of common vocabulary is appropriate. They have good control of simple sentence structures and some control of more complicated sentence structures. Some errors may occur that do not affect meaning. Pronunciation, intonation, and word stress are generally intelligible but may require some listener effort.

- Test takers in this score range can ask for and provide basic information.
- Test takers in this score range can describe objects and people performing activities.
- Test takers in this score range can express thanks and make simple requests, offers, and suggestions.
- Test takers in this score range can narrate and sequence simple events.
- Test takers in this score range can express a simple opinion and give a reason for it.

REDESIGNED TOEIC BRIDGE WRITING TEST PROFICIENCY DESCRIPTORS

Writing Score Range: Scaled Score 15 to 19

Test takers in this score range are developing the ability to write simple words and phrases in order to provide basic personal information such as name, address, age, etc. They typically know the alphabet and can copy words.

- Some test takers in this score range can communicate very simple information about themselves.
- Some test takers in this score range can use simple words or phrases to identify people, objects, places, and activities.

Writing Score Range: Scaled Score 20 to 31

Test takers in this score range can typically write phrases and simple sentences and make use of a limited range of very common vocabulary about very familiar subjects. They can use writing to meet some limited, basic, and practical communication needs, though their writing is sometimes unclear. They have limited control of simple grammatical structures and may have difficulty with word order and word forms.

- Test takers in this score range can communicate very basic information about themselves.
- Test takers in this score range can sometimes give a basic description of people, objects, places, and activities.
- Test takers in this score range are developing the ability to narrate events relating to daily life. They can include some relevant details. They can sequence words and phrases with basic connectors such as *and* or *then*.

Writing Score Range: Scaled Score 32 to 42

Test takers in this score range can typically write phrases and sentences about familiar topics, such as family, people, places, and work. They generally have adequate control of simple grammatical structures and an adequate range of common vocabulary that allow them to meet basic communication needs. Typically there are minor errors in their writing, and some errors may obscure meaning at times.

- Test takers in this score range can typically ask for and provide basic information. However, some important details may be missing or otherwise inappropriate for the task.
- Test takers in this score range can sometimes make simple requests, offers, and suggestions in familiar, everyday situations.
- Test takers in this score range can express basic preferences, likes, and dislikes about familiar topics. However, they may be unable to give a clear reason for their preference.

- Test takers in this score range can usually describe people, objects, places, and activities, though errors may obscure meaning at times.
- Test takers in this score range can describe a simple series of events using a logical sequence. However, the story may be incomplete or underdeveloped. Errors may obscure meaning at times.

Writing Score Range: Scaled Score 43 to 50

Test takers in this score range can typically write sentences, paragraphs, and short essays about familiar topics that contain both abstract and concrete ideas. They generally have good control of common grammatical structures and a good range of common vocabulary that allow them to communicate moderately complex messages. They can connect sentences to form paragraphs that are organized and coherent. Typically there are some minor errors in their writing when expressing complex thoughts or unfamiliar topics.

- Test takers in this score range can ask for and provide basic information.
- Test takers in this score range can describe objects and locations as well as people performing activities.
- Test takers in this score range can express thanks and make simple requests, offers, and suggestions.
- Test takers in this score range can narrate and sequence simple events and routines.
- Test takers in this score range can express a simple opinion and give a reason for it.

FIELD STUDY STATISTICAL ANALYSIS FOR THE REDESIGNED *TOEIC BRIDGE®* TESTS

Peng Lin, Jaime Cid, and Jiayue Zhang

The *TOEIC Bridge®* tests are English language proficiency tests for nonnative speakers of English designed to measure language proficiency at the beginning and the lower-intermediate levels. Test takers may be students of English or those who need to use English for work or travel. From its inception through 2018, the original TOEIC Bridge test consisted of two separate timed sections: listening and reading, with 50 items in each section. The listening section was paced by audio recording.

In 2016, based on feedback received from clients, ETS decided to redesign the original TOEIC Bridge test. The redesigned TOEIC Bridge tests were launched in June 2019. Two changes to the test occurred. First, the redesigned TOEIC Bridge tests focus on communication in the context of everyday adult life (personal, public, and familiar workplace contexts) for the beginning to lower-intermediate English language learners. Second, the redesigned TOEIC Bridge tests also measure speaking and writing communication skills. Unlike the original TOEIC Bridge test, the redesigned tests are a module-based assessment with four modules: listening, reading, speaking, and writing. It is possible to take a single module or any combination of the modules as needed. The redesigned tests measure English language listening, reading, speaking, and writing proficiency of test takers at the levels of A1, A2, and B1 of the Common European Framework of Reference (CEFR). The CEFR describes a progression of language proficiency in listening, reading, speaking, and writing on a six-level scale clustered in three bands: A1–A2 (basic user), B1–B2 (independent user), and C1–C2 (proficient user; Council of Europe, 2001).

A variety of item types of the redesigned TOEIC Bridge tests were evaluated by content experts (see Everson et al., 2019). An item-level pilot study was administered in September 2017 in Japan, Korea, Taiwan, and Brazil to help specify both the appropriate item types and the appropriate number of items within each item type for all four skills (tests). Observations from the pilot study (e.g., item difficulty, format appropriateness, and testing time) were used to refine the item and test specifications for the redesigned TOEIC Bridge tests.

In April 2018, a field study was launched in three Asian countries (Japan, Korea, and Taiwan) and three non-Asian countries (Colombia, Brazil, and Mexico), in which the original Bridge test was well adopted. After the data collection was completed, statistical analyses were conducted to evaluate the statistical properties of the redesigned TOEIC Bridge tests (e.g., difficulty and discrimination of the items, correlation among different parts of the test, reliability, interrater reliability for speaking and writing). The purpose of this report is to document the results of the statistical analyses of the listening, reading, speaking, and writing tests of the field study. These results contributed to the conceptual assessment framework and assessment implementation layers of the evidence-centered design test development process that was utilized for the development of the redesigned TOEIC Bridge tests (see Mislevy & Yin, 2012). Although not part of

this report, the results from the statistical analyses of the field study informed the final decisions on the reporting scales of the redesigned TOEIC Bridge tests and the performance proficiency levels for listening, reading, speaking, and writing. The reported score scales of all four tests were set to range from 15 to 50 in increments of 1.

Background: Field Study Test Specifications

The redesigned TOEIC Bridge Listening and Reading tests contain only multiple-choice items that are scored dichotomously. As shown in Table 1, the listening test consists of four parts and the reading test consists of three parts. Unlike the original TOEIC Bridge test, which had two subscore areas for the listening section and three for the reading section, four ability measures were developed for each test (i.e., listening and reading) of the redesigned TOEIC Bridge test. The four abilities are reported to test takers using a percentage correct score. Table 2 presents the number of items associated with the abilities in the listening and reading tests of the field study. The position and the number of items associated with each ability may vary across operational forms. The redesigned TOEIC Bridge Speaking test consists of six constructed-response item types. The redesigned TOEIC Bridge Writing test consists of four constructed-response item types and one multiple-selection multiple-choice item type (Build a Sentence). See Tables 3 and 4 for details.

TABLE 1

Parts of the Redesigned TOEIC Bridge Listening and Reading Tests

Part	Number of items
Listening	
Part 1. Four Pictures	6
Part 2. Question Response	20
Part 3. Conversations	10
Part 4. Talk	14
Reading	
Part 1. Sentence Completion	15
Part 2. Text Completion	15
Part 3. Reading Comprehension	20

TABLE 2

Ability Measures of the Redesigned TOEIC Bridge Listening and Reading Tests

Ability	Number of items
Listening	
Appropriate Response	20
Short Dialogue or Conversation	32
Short Monologue	12
Main Idea or Stated Fact	23
Reading	
Vocabulary	14
Grammar	13
Main Idea or Stated Fact	16
Short Informational Written Texts	20

Note. The listening and reading tests each have 50 items. The sum of items for all abilities is greater than 50 as some items contribute to more than one ability.

TABLE 3

Item Types of the Redesigned TOEIC Bridge Speaking Test

Item	Item type	Score scale
1–2	Read a Short Text Aloud	0–3
3–4	Describe a Photograph	0–3
5	Listen and Retell	0–3
6	Short Interaction	0–3
7	Tell a Story	0–4
8	Make and Support a Recommendation	0–4

TABLE 4

Item Types of the Redesigned TOEIC Bridge Writing Test

Item	Item type	Score scale
1–3	Build a Sentence	0–2
4–6	Write a Sentence	0–3
7	Respond to a Brief Message	0–3
8	Write a Narrative	0–3
9	Respond to an Extended Message	0–4

Field Study Test Data Collection

Two parallel test forms for listening and reading (Form LR1 and Form LR2) and two for speaking and writing (Form SW1 and Form SW2) were assembled and administered in the field study. All items were new with no previous statistics available. The two listening and reading forms shared 20 common items in listening and 20 in reading (i.e., 40% of the total items in the test). No items were common between the two speaking and writing forms.

The test was administered in two separate sessions: one for listening and reading and one for speaking and writing. In each session, the two forms were randomly administered to test takers in order to make the test-taking groups of the two forms approximately equivalent. For listening and reading, the scores of the two forms were equated through common items and converted to scale scores. For speaking and writing, the scores were made comparable between forms through well-defined and articulated scoring rubrics and quality control procedures. Thus, the scale scores from the two forms can be deemed comparable within each test (i.e., listening, reading, speaking, and writing) of the field study.

In total, 2,368 test takers from six countries (Japan, Korea, Taiwan, Colombia, Brazil, and Mexico) participated and took all four tests in the field study. Although an effort was made to recruit test takers from all the ability scale ranges of the target population (i.e., A1, A2, and B1), the small samples collected from some countries precluded a balanced ability distribution in all countries. In addition, the number of test takers from Colombia and Mexico was noticeably below the targeted numbers. Tables 5 and 6 summarize demographic compositions of the field study sample by country and by gender. Approximately half of the test takers were from Japan.

TABLE 5

Country Distributions of Test Takers in Field Study

Country	*N*	Percentage
Brazil	251	11
Colombia	18	1
Japan	1,250	53
Korea	391	17
Mexico	49	2
Taiwan	409	17
Total	2,368	100

TABLE 6

Country Distributions of Test Takers in Field Study

Gender	*N*	Percentage
Female	1,118	47
Male	1,249	53
Unidentified	1	
Total	2,368	100

Performance by Country and Gender

Table 7 provides the mean and standard deviation of the scale scores of the field study for the four tests by country. Recall that the redesigned TOEIC Bridge test has the scale scores of all four tests reported on a scale from 15 to 50 in increments of 1. On average, Japanese test takers were the most able group among the six countries for both listening and reading. This finding is different from what was observed in the original TOEIC Bridge test in operational settings, where Korean test takers performed better than Japanese test takers in both listening and reading. Therefore, the field study sample may not have been representative of the operational test-taking population. Scaled scores of Taiwanese test takers were close to those of Japanese test takers in listening but were noticeably lower in reading. For speaking, Colombian and Mexican test takers scored higher than the other countries, and for writing, Japan, Colombia, and Mexico were the countries that had the highest scaled scores. Test takers from Brazil produced the lowest scaled score means in all four tests. When interpreting these results, it is important to note that Colombia and Mexico had considerably smaller samples of test takers than the other countries.

TABLE 7

Mean and Standard Deviation of the Test Scores by Country

Country	*N*	Listening *M*	Listening *SD*	Reading *M*	Reading *SD*	Speaking *M*	Speaking *SD*	Writing *M*	Writing *SD*
Brazil	251	21.95	9.92	25.39	9.93	24.02	11.08	26.96	11.79
Colombia	18	32.06	9.32	39.78	8.24	38.94	8.79	41.17	8.30
Japan	1,250	36.35	8.35	42.58	7.44	36.37	8.54	41.47	8.04
Korea	391	31.63	9.81	32.93	10.51	32.25	9.71	35.85	9.93
Mexico	49	31.35	12.39	36.10	10.72	38.55	9.76	41.67	8.62
Taiwan	409	36.09	10.20	39.35	10.75	35.99	11.12	38.29	11.78
Total	2,368	33.86	10.24	38.45	10.63	34.38	10.29	38.46	10.54

The mean and standard deviation of scale scores of the field study by gender are provided in Table 8. As can be seen from the table, on average, female test takers performed better than male test takers in all four tests in all countries. We observed this same trend with the original TOEIC Bridge test in operational settings.

TABLE 8

Mean and Standard Deviation of the Test Scores by Gender

Country	*N*	Listening *M*	Listening *SD*	Reading *M*	Reading *SD*	Speaking *M*	Speaking *SD*	Writing *M*	Writing *SD*
Female	1,118	34.89	10.44	39.25	10.42	35.27	10.64	39.41	10.42
Male	1,249	32.94	9.97	37.73	10.77	33.59	9.90	37.61	10.59
Unidenti-fied	1	38.00	–	38.00	–	23.00	–	32.00	–
Total	2,368	33.86	10.24	38.45	10.63	34.38	10.29	38.46	10.54

Statistical Analysis Results

The analyses presented in the next sections are based on the combined field study samples, with a total of 2,050 test takers, from the countries with large numbers of test takers—Japan, Korea, and Taiwan—in operational administrations. These countries comprised 83% of the field study sample. The summary statistics of the scaled scores, including mean, standard deviation, minimum, maximum, and the correlation among tests, are presented in Table 9. On average, the reading and writing tests yielded the

highest means, while the listening and speaking tests yielded the lowest means. The Pearson correlation coefficients for the four tests in Table 10 show that the four sets of test scores were moderately correlated. These correlations are similar to the ones reported for the *TOEIC*® Listening, Reading, Speaking, and Writing tests (e.g., Liu & Costanzo, 2013).

TABLE 9

Summary Statistics of Test Scores for the Field Study Sample (Japan, Korea, and Taiwan)

Test	Listening	Reading	Speaking	Writing
N	2,050	2,050	2,025[a]	2,050
Mean	35.40	40.09	35.52	39.76
SD	9.21	9.56	9.45	9.54
Minimum	15	15	15	15
Maximum	50	50	50	50

[a] Twenty-five test takers with some nonscorable responses in the speaking test did not have speaking scores.

TABLE 10

Correlation Coefficients for the Four Tests of the Field Study Sample (Japan, Korea, and Taiwan)

Correlation	Listening	Reading	Speaking	Writing
Listening	1	.78	.68	.66
Reading		1	.66	.74
Speaking			1	.71
Writing				1

Table 11 provides the mean and standard deviation of scale scores for the two listening and reading forms (LR1 and LR2) and for the two speaking and writing forms (SW1 and SW2) for test takers from Japan, Korea, and Taiwan. One can see that the mean and standard deviation of the scale scores of the two forms within each test were very close, which indicates that the groups taking LR1 and LR2 in listening and reading and SW1 and SW2 in speaking and writing were approximately equivalent. It also indicates that the approaches to making speaking and writing scores comparable across forms appeared successful in the field study.

TABLE 11

Mean and Standard Deviation of the Test Scores by Form

Form	*N*	Listening *M*	Listening *SD*	Reading *M*	Reading *SD*	Speaking *M*	Speaking *SD*	Writing *M*	Writing *SD*
LR1	1,018	35.28	9.01	40.45	9.51				
LR2	1,032	35.51	9.41	39.74	9.61				
SW1	1,028					35.90	9.32	39.83	9.31
SW2	1,022					35.14	9.57	39.69	9.76

Note. LR = listening and reading; SW = speaking and writing.

Due to differences in test format across tests (i.e., multiple-choice items for listening and reading and constructed-response items for speaking and writing), we conducted separate statistical analyses for listening and reading and for speaking and writing. In the following sections, the statistical analysis results are presented in the same order.

LISTENING AND READING

Reliability and Standard Error of Measurement

The reliabilities of the listening and reading tests in the field study were estimated using an internal consistency method (reliability coefficient called alpha). This method assesses the consistency of test takers' responses to all of the items in the test, part, or ability. The reliability estimate ranges from 0 to 1. The higher the reliability coefficient is for the test, part, or ability, the higher the consistency of test takers' responses is to the items of the test, part, or ability. The standard error of measurement (SEM)—another indicator of score consistency—estimates the average variation expected in a test taker's score from one test form to another.

The reliability coefficients and the SEMs for the total test and different parts and abilities of the two listening and reading field study forms are reported in Tables 12 and 13, respectively. The reliabilities of listening in Form LR1 and Form LR2 were .88 and .89, respectively. Reading produced reliabilities of .93 in both forms. In listening, Four Pictures (Part 1) with six items and Question Response (Part 2) with 20 items produced the lowest reliability and highest reliability, respectively, in both forms. Likewise, Reading Comprehension (Part 3) with 20 items produced the highest reliability in reading. The reliabilities of both listening and reading of the field study forms were relatively higher than the average reliabilities

of listening (.83) and reading (.85) of the original TOEIC Bridge test forms. The reliabilities of three of the four abilities in listening and all four abilities in reading were above .75 in both forms. To increase the reliabilities of Short Monologue in listening, which were .68 and .71, respectively, in the two LR forms, it was decided to add two more items to this ability in the operational forms. The SEM of total score was very close in the two forms in both listening and reading, with listening yielding a slightly higher total SEM than reading (3.0 vs. 2.5).

TABLE 12

Reliability and SEM for All Scores of Listening Test by Form

Part/ability	LR1– number of items	LR1– reliability	LR1– SEM	LR2– number of items	LR2– reliability	LR2– SEM
Part						
Part 1. Four Pictures	6	.43	0.73	6	.47	0.66
Part 2. Question Response	20	.77	1.67	20	.78	1.64
Part 3. Conversations	12	.70	1.34	12	.68	1.30
Part 4. Talk	12	.68	1.43	12	.71	1.42
Ability						
Appropriate Response	20	.77	1.68	20	.78	1.64
Short Dialogue or Conversation	32	.84	2.16	32	.85	2.10
Short Monologue	12	.68	1.43	12	.71	1.42
Main Idea and/or Stated Fact	23	.80	1.94	22	.82	1.85
Total (scale score)	50	.88	3.07	50	0.89	3.09

Note. LR = listening and reading; SEM = standard error of measurement. Form LR1: N = 1,018. Form LR2: N = 1,032. The sum of items for all abilities in a form might be greater than 50 as some items contribute to more than one ability.

TABLE 13

Reliability and SEM for All Scores of Reading Test by Form

Part/ability	LR1– number of items	LR1– reliability	LR1– SEM	LR2– number of items	LR2– reliability	LR2– SEM
Part						
Part 1. Sentence Completion	15	.77	1.44	15	.80	1.42
Part 2. Text Completion	15	.83	1.24	15	.77	1.43
Part 3. Reading Comprehension	20	.87	1.65	20	.89	1.66
Ability						
Vocabulary	14	.78	1.22	13	.78	1.21
Grammar	13	.78	1.32	14	.77	1.43
Main Idea or Stated Fact	16	.86	1.49	16	.87	1.48
Short Informational Written Texts	20	.87	1.65	20	.89	1.66
Total (scale score)	50	.93	2.49	50	.93	2.51

Note. LR = listening and reading; SEM = standard error of measurement. Form LR1: N = 1,018. Form LR2: N = 1,032. The sum of items for all abilities in a form might be greater than 50 as some items contribute to more than one ability.

Please note that the magnitude of reliability depends not only on the internal consistency of the items in the test but also on the homogeneity of the test takers. The reliabilities for the field study forms in this report may not be directly comparable to what one would observe in operational settings, as the field study sample may have not been representative of the operational test-taking population.

Item Difficulty

In this study, item difficulty was evaluated by examining two types of statistical indices: *p* value (defined as the proportion of the test takers who answered an item correctly in a given population) and delta (defined as 13 minus 4z, where z is the standard normal deviate corresponding to proportion correct). The *p* value ranges from 0 to 1. The higher the *p* value is, the easier the item is. The *p* value is dependent on the ability levels of the sample taking the test. That is, the *p* value of the same items based on a more able group will be higher than those based on a less able group. Therefore, *p* values are not directly comparable across forms taken by different groups of test takers to reflect the difficulty of items in different forms. The equated deltas provide a difficulty metric that accounts for the different ability levels

across groups that take different forms. Delta values typically range from 6.0 for a very easy item (i.e., approximately 95% correct) to 20 for a very difficult item (i.e., approximately 5% correct); the mean of 13.0 corresponds to 50% correct. To compare the item difficulty between the two field study forms with the original TOEIC Bridge test, equated deltas, which transfer the observed delta of the field test forms to the scale of the original TOEIC Bridge test, were calculated based on a single group design. Specifically, a group of 300 test takers took both the field study Form LR1 and an operational form of the original TOEIC Bridge test. The equated deltas of items in Form LR1 were calculated based on the equated deltas of items in the operational form of the original TOEIC Bridge test. Form LR1 was then used as the reference form and the delta value of LR2 items were adjusted to the scale of LR1 through the anchor items.

The *p* value and equated delta of items for listening are summarized in Tables 14 and 15. As can be seen, on average, the listening tests of the two field study forms were comparable in difficulty. The mean of the equated delta in both forms was 11.7. Among the four parts in listening, Four Picture items were, on average, the easiest, and the Talk items were the most difficult. Among the four abilities in the listening test, Appropriate Response and Short Dialogue or Conversation were relatively easier than Short Monologue and Main Idea or Stated Fact.

TABLE 14

Item Difficulty in Listening Test of Form LR1

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Four Pictures	.88	.09	9.5	2.2
Part 2. Question Response	.76	.14	11.7	2.0
Part 3. Conversations	.75	.08	12.0	1.0
Part 4. Talk	.70	.10	12.7	1.3
Ability				
Appropriate Response	.76	.14	11.7	2.0
Short Dialogue or Conversation	.76	.12	11.8	1.7
Short Monologue	.70	.10	12.7	1.3
Main Idea or Stated Fact	.72	.10	12.4	1.2
All 50 items	.76	.13	11.7	1.9

TABLE 15

Item Difficulty in Listening Test of Form LR2

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Four Pictures	.90	.06	9.3	1.6
Part 2. Question Response	.77	.14	11.6	1.9
Part 3. Conversations	.74	.16	12.0	1.8
Part 4. Talk	.68	.11	12.8	1.2
Ability				
Appropriate Response	.77	.14	11.6	1.9
Short Dialogue or Conversation	.76	.15	11.8	1.9
Short Monologue	.68	.11	12.8	1.2
Main Idea or Stated Fact	.70	.14	12.6	1.6
All 50 items	.76	.15	11.7	2.0

The results for reading are summarized in Table 16 and Table 17. As is shown, on average, the reading tests of the two field study forms were comparable in difficulty. The mean of the equated delta of reading tests of the two forms were 11.3 and 11.4, respectively. Among the three parts in the reading test, Text Completion, on average, was easier than Sentence Completion and Reading Comprehension in Form LR1. But this was not the case in Form LR2, where Text Completion, on average, was as difficult as Sentence Completion and easier than Reading Comprehension. Among the four abilities in the reading test, Vocabulary was easier than the remaining three abilities, which had similar difficulties in both forms.

TABLE 16

Item Difficulty in Reading Test of Form LR1

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Sentence Completion	.72	.18	11.8	2.9
Part 2. Text Completion	.83	.08	10.0	1.5
Part 3. Reading Comprehension	.72	.14	11.9	2.3
Ability				
Vocabulary	.80	.16	10.3	2.6
Grammar	.74	.13	11.5	2.2
Main Idea or Stated Fact	.72	.11	11.9	2.0
Short Informational Written Texts	.72	.14	11.9	2.3
All 50 items	.75	.15	11.3	2.5

TABLE 17

Item Difficulty in Reading Test of Form LR2

Part/ability	*p* value *M*	*p* value *SD*	Equated delta *M*	Equated delta *SD*
Part				
Part 1. Sentence Completion	.77	.08	11.1	1.6
Part 2. Text Completion	.76	.12	11.1	2.0
Part 3. Reading Comprehension	.72	.11	11.9	1.9
Ability				
Vocabulary	.82	.06	10.2	1.2
Grammar	.70	.10	10.7	1.7
Main Idea or Stated Fact	.72	.11	11.9	1.9
Short Informational Written Texts	.72	.11	11.9	1.9
All 50 items	.75	.11	11.4	1.9

Tables 18 and 19 provide a comparison of the equated deltas of the two field study forms with those of the operational forms of the original TOEIC Bridge test since 2013. As can be seen from the tables, the average equated delta of the listening test of both field study forms (mean equated delta = 11.7) was slightly higher than the maximum delta (11.6) of the operational forms. The listening test of both field study forms (mean equated delta = 11.7) was on average more difficult than the listening test of the original TOEIC Bridge test (mean equated delta = 11.0). The 0.7 difference between the equated delta of the field study forms and the average of the original TOEIC Bridge test forms indicates that the listening test of the field study forms was approximately 6% (3 items or points for 50 questions) more difficult than the original TOEIC Bridge test. On the other hand, the equated delta of the reading test of the two field study forms (mean equated delta = 11.3 and 11.4, respectively) was lower than the average of the operational forms (mean equated delta = 11.9) but still within the range of operational forms (11.0 – 12.8). The 0.55 average difference between the equated delta of the field study forms and the average of the original TOEIC Bridge test operational forms indicates that the reading test of the field study forms was approximately 5% (2.5 items or points for 50 questions) easier than the original TOEIC Bridge test. In addition, Tables 18 and 19 also suggest that the difficulty of listening and reading tests may be closer to one another in the redesigned TOEIC Bridge test (mean equated delta = 11.7 for listening vs. mean equated delta = 11.4 for reading in field study) than in the original TOEIC Bridge test (mean equated delta = 11.0 for listening vs. mean equated delta = 11.9 for reading). In summary, although on average the field study forms were more difficult (listening) or easier (reading) than the original TOEIC Bridge test, their average difficulty was very close to the operational historical ranges in both measures. Therefore, we expect that the difficulty of the redesigned TOEIC Bridge test in the operational samples will be comparable to that of the original TOEIC Bridge test.

TABLE 18

Equated Delta of Listening and Reading of Field Study Forms

Test/field study form	*M*	*SD*
Listening–Form LR1	11.7	1.9
Listening–Form LR2	11.7	2.0
Reading–Form LR1	11.3	2.5
Reading–Form LR2	11.4	1.9

Note. LR = listening and reading. The original TOEIC Bridge test includes operational forms since 2013.

TABLE 19

Equated Delta of the Original TOEIC Bridge Test

Test	*M*	Minimum	Maximum
Listening	11.0	10.4	11.6
Reading	11.9	11.0	12.8

Item Discrimination

Item discrimination was evaluated by the biserial correlation coefficient. The biserial correlation is the relationship between test takers' scores on a particular item (i.e., 0 for an incorrect response or 1 for a correct response) with the corresponding total score (i.e., sum of item scores for a test). The biserial correlation indicates how well an item serves to discriminate between low- and high-ability test takers. Tables 20 and 21 present the summary statistics of the biserial correlations of items in listening and reading, respectively, in the two field study forms. For listening, the biserials were comparable between the two field study forms and across parts and abilities within the forms. For reading, the biserials, on average, were comparable between the two field study forms but were varied across parts and abilities within the forms. The overall biserial of the listening and reading of the two field study forms were both higher than the average biserial of the original TOEIC Bridge forms listening and reading, .45 and .46, respectively.

TABLE 20

Biserial Correlations of Items in the Listening Test of Form LR1 and Form LR2

Part/ability	Form LR1 *M*	Form LR1 *SD*	Form LR2 *M*	Form LR2 *SD*
Part				
Part 1. Four Pictures	.46	.10	.47	.06
Part 2. Question Response	.51	.09	.52	.09
Part 3. Conversations	.50	.09	.52	.10
Part 4. Talk	.49	.08	.52	.09
Ability				
Appropriate Response	.51	.09	.52	.09
Short Dialogue or Conversation	.51	.09	.52	.09
Short Monologue	.49	.08	.52	.09
Main Idea or Stated Fact	.49	.09	.52	.10
All 50 items	.50	.09	.51	.09

Note. LR = listening and reading.

TABLE 21

Biserial Correlations of Items in Reading Test of Form LR1 and Form LR2

Part/ability	Form LR1 *M*	Form LR1 *SD*	Form LR2 *M*	Form LR2 *SD*
Part				
Part 1. Sentence Completion	.53	.12	.55	.09
Part 2. Text Completion	.60	.09	.53	.08
Part 3. Reading Comprehension	.61	.11	.61	.09
Ability				
Vocabulary	.56	.14	.56	.08
Grammar	.57	.10	.50	.09
Main Idea or Stated Fact	.62	.09	.62	.09
Short Informational Written Texts	.61	.11	.61	.09
All 50 items	.58	.11	.57	.09

Note. LR = listening and reading.

Speededness

The TOEIC Bridge Listening test is paced by an audio recording, and thus speededness could not be assessed in the traditional way. For the reading test of the field study, four types of statistics frequently used to evaluate the speededness of a test are presented in Table 22: (a) percentage of test takers completing the whole test, (b) percentage of test takers completing 75% of the test, (c) the number of items reached by 80% of the test takers, and (d) the ratio of not reached variance to the total score variance (i.e., the speededness index). If nearly all of the test takers complete 75% of the items in a test and if nearly all of the items are reached by at least 80% of the test takers and if the speededness index is .15 or less, the test is usually considered relatively unspeeded. The statistics in Table 22 indicate that the reading test of both field study forms was not speeded.

TABLE 22

Statistics of Speededness for Reading

Speededness	Form LR1	Form LR1	Original TOEIC Bridge Test		
			M	Minimum	Maximum
Percent completing test	97.8%	97.0%	95.0%	92.3%	97.3%
Percent completing 75%	99.8%	99.6%	99.6%	98.7%	99.9%
Number of items reached by 80%	50	50	49.9	48	50
Speededness index	0.01	0.02	0.05	0.02	0.1

Note. LR = listening and reading. The original TOEIC Bridge test includes operational forms since 2013.

SPEAKING AND WRITING

Difficulty

The difficulty of the two field study speaking and writing forms (SW1 and SW2) was evaluated at item level. The means and standard deviations of item scores and total scores (scale score) of the speaking and writing tests of the field study are presented in Tables 23 and 24. Missing responses were excluded when calculating the statistics of item scores. Therefore, the sample sizes were slightly different across items within the same form. In general, the higher the score mean was (relative to its possible score range), the easier the item was for a given group of test takers. The means of the total score of the two forms were comparable to one another in both speaking and writing. Overall, in speaking, the difficulty of items was comparable between the two forms. Among the six item types of the speaking test, Read a Short Text Aloud was relatively easier than the other item types, and Short Interaction was the most difficult item type. In the writing test, one can see a larger variability in difficulty within the same item type. For example, on both forms, the first Build a Sentence item was the easiest and the third Build a Sentence item was the hardest. On average, the Respond to a Brief Message item and the Respond to an Extended Message item were relatively easier than the other item types; Write a Narrative was the hardest. Overall, these results indicate that item types in both the speaking and the writing tests can have different levels of item difficulty. These findings were shared with test developers so they could make appropriate adjustments to the test design and were taken into account when making final decisions on the reporting scales of the redesigned TOEIC Bridge test.

TABLE 23

Item Difficulty for Speaking Tests of Form SW1 and Form SW2

Item	Item type	Score scale	Form SW1–*N*	Form SW1–*M*	Form SW1–*SD*	Form SW2–*N*	Form SW2–*M*	Form SW2–*SD*
1	Read a Short Text Aloud	0–3	1,012	2.58	0.58	983	2.55	0.59
2	Read a Short Text Aloud	0–3	1,016	2.60	0.57	986	2.61	0.58
3	Describe a Photograph	0–3	1,012	2.37	0.59	977	2.37	0.62
4	Describe a Photograph	0–3	1,011	2.37	0.62	986	2.31	0.61
7	Listen and Retell	0–3	921	2.16	0.74	913	2.15	0.70
5	Short Interaction	0–3	927	1.93	0.84	880	1.75	0.62
5	Short Interaction	0–3	927	1.93	0.84	880	1.75	0.62
6	Tell a Story	0–4	1,005	2.60	0.79	971	2.62	0.82
8	Make and Support a Recommendation	0–4	969	2.73	0.89	936	2.67	0.84

Note. SW = speaking and writing.

TABLE 24

Item Difficulty for the Writing Tests of Form SW1 and Form SW2

Item	Item type	Score scale	Form SW1–*N*	Form SW1–*M*	Form SW1–*SD*	Form SW2–*N*	Form SW2–*M*	Form SW2–*SD*
1	Build a Sentence	0–2	1,014	1.65	0.48	1,018	1.96	0.20
2	Build a Sentence	0–2	1,023	1.67	0.47	1,022	1.38	0.49
3	Build a Sentence	0–2	1,024	1.22	0.42	1,013	1.16	0.37
4	Write a Sentence	0–3	999	2.22	0.77	987	2.00	0.79
5	Write a Sentence	0–3	1,019	2.23	0.71	1,001	1.57	0.71
6	Write a Sentence	0–3	1,019	2.15	0.71	1,000	2.15	0.73
7	Respond to a Brief Message	0–3	1,010	2.11	0.80	997	2.35	0.80
8	Write a Narrative	0–3	996	2.00	0.76	976	2.13	0.77
9	Respond to an Extended Message	0–4	1,001	2.73	0.93	977	2.98	0.88

Note. SW = speaking and writing.

Item Total Correlations

In order to evaluate the contribution of items to the total scores, Pearson product-moment correlations were calculated between items and the total scores. Items with a high correlation with the total test score are deemed better at discriminating high-ability test takers from low-ability test takers and, therefore, contribute more to the overall test reliability. Observed score correlation coefficients between item score and total raw score (sum of the item scores) are presented in Tables 25 and 26 for the speaking and writing tests, respectively. In the speaking test, the correlations were moderate to high (from .52 to .80). On both forms, Item 8 (Make and Support a Recommendation) and Items 1 and 2 (Read a Short Text Aloud) yielded the highest and lowest correlations, respectively. In the writing test, the item total correlation ranged from .28 to .79 on both forms. The correlations for Items 1 through 6, especially for Item 3 (one of the Write a Sentence items) were noticeably lower than those for Items 7 through 9. As expected, the item total correlations of the Build a Sentence item type (Items 1, 2, and 3) were, on average, lower than those of the other item types on both forms because of their narrow score range (0 – 2) and extreme difficulty (i.e., items too easy or too difficult). Item 9 (Respond to an Extended Message) was the item that produced the highest item total correlation on both forms.

TABLE 25

Item Total Correlation for Speaking Forms SW1 and SW2

Form	Total score	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
SW1	Speaking raw score	.52	.53	.61	.58	.60	.73	.68	.80
SW2	Speaking raw score	.57	.58	.61	.59	.54	.77	.66	.79

Note. SW = speaking and writing.

TABLE 26

Item Total Correlation for Writing Forms SW1 and SW2

Form	Total score	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9
SW1	Writing raw score	.40	.38	.30	.41	.48	.41	.69	.72	.79
SW2	Writing raw score	.28	.51	.34	.34	.38	.52	.67	.76	.79

Note. SW = speaking and writing.

Test Reliability and Standard Error of Measurement

The traditional coefficient alpha index—a measure of internal consistency—was used to estimate the reliability of speaking and writing tests. Table 27 displays the internal consistency reliability coefficients and SEM of test scores for the two forms of the speaking and writing tests. Although SEM was comparable between the two forms for both speaking and writing, respectively, speaking yielded smaller SEM than writing. The reliability estimates for the two speaking forms were .83 and .86, respectively. The reliability estimates for the two writing forms were .73 and .75, respectively. Based on feedback provided by test developers regarding conceptual communalities for some item types and the different levels of item difficulty observed in the field study, we also evaluated the consistency of test-taker performance on individual items within three levels of item difficulty: low, medium, and high. The coefficient alpha calculated based on the alphas within each classification is known as stratified coefficient alpha. Table 27 shows that in the speaking test the coefficient alpha and stratified alpha were quite comparable on both forms. However, in the writing test, the stratified alpha was higher than the coefficient alpha, especially on Form SW2. These findings informed the final decisions regarding the test design and the reporting scales of the redesigned TOEIC Bridge test. Specifically, the findings were used to determine the appropriate weights for each of the item types and evaluate their impact on the reliability of total scores.

TABLE 27

Reliability and SEM of Speaking and Writing of Forms SW1 and SW2

Alpha	Form SW1–reliability	Form SW1–SEM	Form SW2–reliability	Form SW3–SEM
Coefficient alpha speaking	.83	3.82	.86	3.54
Coefficient alpha writing	.73	4.79	.75	4.85
Stratified alpha speaking	.84	3.73	.87	3.45
Stratified alpha writing	.78	4.37	.82	4.14

Note. SW = speaking and writing; SEM = standard error of measurement.

Interrater Agreement

Because all items in the TOEIC Bridge Speaking and Writing tests (except for Writing Items 1, 2, and 3, which are multiple choice items) were scored by two independent raters in the field study, it was important to evaluate the consistency of ratings given by different raters. The agreement rates between the first and second ratings are presented in Tables 28 and 29. In the speaking test, the percentage of exact agreement ranged from 57% to 81% on both forms, indicating that, for all items, more than half of the test takers received the same ratings from the first and second raters. The percentage of discrepancy was below 1% for most speaking items. Item 6 (Tell a Story) yielded the highest percentage of discrepancy on both forms (2.85% and 2.65% on forms SW1 and SW2, respectively), indicating that only a few test takers were given a score with a difference of two or more points from the two raters. This finding was consistent with the interrater correlation for speaking items, which ranged from .56 to .89. The interrater agreement observed in writing items, on average, was higher than that of the speaking items. In writing, the percentage of exact agreement ranged from 63% to 91%. The highest discrepancy rate (1.31%) was produced by Item 9 (Respond to an Extended Message). The interrater correlation for writing items ranged from .77 to .92.

TABLE 28

Interrater Agreement and Reliability of Speaking for Forms SW1 and SW2

Item	Form SW1–*N*	Form SW1–exact %	Form SW1–adj. %	Form SW1–dis. %	Form SW1–interrater	Form SW2–*N*	Form SW2–exact %	Form SW2–adj. %	Form SW2–dis. %	Form SW2–interrater
1	1,019	73	27	0.2	.62	1,005	69	30	0.3	.67
2	1,019	73	27	0.1	.59	1,006	72	28	0.4	.67
3	1,019	66	33	0.2	.56	1,006	68	32	0.2	.70
4	1,019	67	33	0.5	.59	1,006	62	37	0.4	.59
5	1,019	78	21	0.8	.87	1,006	74	25	1.1	.78
6	1,019	60	37	2.9	.67	1,006	57	41	2.7	.70
7	1,019	81	19	0.1	.89	1,006	79	21	0.0	.87
8	1,019	63	36	1.3	.82	1,006	64	35	0.9	.83

Note. SW = speaking and writing; exact % = the percentages of exact agreement between two ratings; adj. % = the percentages of adjacent ratings between the two raters; dis. % = the percentages of ratings with a discrepancy of 2 or more score points.

TABLE 29

Interrater Agreement and Reliability of Writing for Forms SW1 and SW2

Item	Form SW1–*N*	Form SW1–exact %	Form SW1–adj. %	Form SW1–dis. %	Form SW1–interrater	Form SW2–*N*	Form SW2–exact %	Form SW2–adj. %	Form SW2–dis. %	Form SW2–interrater
4	1,008	91	9	0.5	.92	998	84	14	1.9	.84
5	1,021	89	11	0.6	.88	1,007	86	13	0.6	.85
6	1,023	87	12	0.9	.85	1,007	86	14	0.7	.85
7	1,017	74	25	0.5	.80	1,005	80	20	0.3	.84
8	1,002	73	26	0.9	.76	989	73	26	0.7	.77
9	1,011	70	29	1.2	.82	989	63	35	1.3	.78

Note. SW = speaking and writing; exact % = the percentages of exact agreement between two ratings; adj. % = the percentages of adjacent ratings between the two raters; dis. % = the percentages of ratings with a discrepancy of 2 or more score points.

CONCLUSION

The redesigned TOEIC Bridge tests were launched in June 2019. Before the official launch, a field study, with two parallel forms for listening and reading and two for speaking and writing, was administered in April 2018 to evaluate the statistical properties of the redesigned TOEIC Bridge tests. Test takers from six countries (Japan, Korea, Taiwan, Colombia, Brazil, and Mexico) participated in the field study. The statistical analysis in this report focused on the test takers from Japan, Korea, and Taiwan, who comprised 83% of the field study sample.

Overall, the reliabilities of the listening and reading tests in the field study were relatively higher than the average reliabilities of the original TOEIC Bridge Listening and Reading tests. Although on average the field study forms were harder (listening) or easier (reading) than the original TOEIC Bridge test, their average difficulty was very close to the operational historical ranges in both measures. Therefore, we expect that the overall difficulty of the redesigned TOEIC Bridge Listening and Reading tests will not differ much from that of the original TOEIC Bridge Listening and Reading test. The overall item discrimination of both listening and reading of the field study was relatively higher than the original TOEIC Bridge test, with no speediness issues identified.

Because speaking and writing are not part of the original TOEIC Bridge test, it is not possible to compare the statistics of the redesigned TOEIC Bridge tests to those of the original TOEIC Bridge test. The difficulty of items varied across different item types for both the speaking and writing tests. Overall, these results indicate that item types in both speaking and writing tests can have different levels of item difficulty. In speaking, reliability—measured by coefficient alpha and stratified alpha—were quite comparable on both forms (over .80). In writing, although stratified alpha was higher than the coefficient alpha, reliabilities were lower than speaking on both forms (over .70). The interrater agreement rates were found to be reasonably high for both tests.

All in all, these findings not only helped to inform the final decisions regarding the final reporting scales of the redesigned TOEIC Bridge tests, but also allowed test developers make appropriate adjustments to the test design. Given that the findings of this study were based on a field study sample, which may have not been fully representative of the operational test-taking population, additional analyses will be conducted once sufficient operational data are gathered after the redesigned TOEIC Bridge tests are officially launched.

REFERENCES

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Press Syndicate of the University of Cambridge.

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.

Liu, J., & Costanzo, K. (2013). The relationship among TOEIC listening, reading, speaking, and writing skills. In D. E. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies* (Vol. II, pp. 2.1–2.25). ETS.

Mislevy, R. J., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). Routledge.

SECTION II: ACCUMULATING EVIDENCE TO SUPPORT CLAIMS

MAPPING THE REDESIGNED *TOEIC BRIDGE*® TEST SCORES TO PROFICIENCY LEVELS OF THE COMMON EUROPEAN FRAMEWORK OF REFERENCE FOR LANGUAGES

Jonathan Schmidgall

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 seconds to prepare and 30 seconds 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 seconds to prepare and 30 seconds 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 seconds to prepare and 30 seconds 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 seconds to prepare and 30 seconds 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 seconds to prepare and 45 seconds 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 seconds to prepare and 60 seconds 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 seconds 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 seconds 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				
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				

CEFR communicative language activity, strategy, or competency descriptor scales	TOEIC Bridge			
	Reading	Listening	Speaking	Writing
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 hours 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 hrs), 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*® test 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.

	A	B	C	D
1	Item \ Panelist	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

Figure 1. Sample of a completed spreadsheet for Round 1 judgments using the modified Angoff method.

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+**
M	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+
M	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+
M	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+
M	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*® test 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 test and TOEIC test scores. Because these tests had been independently mapped to the CEFR and were part of the TOEIC program 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 occurred 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 test 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 test 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. **https://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 manual* (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"

THE REDESIGNED *TOEIC BRIDGE*® TESTS: RELATIONS TO TEST-TAKER PERCEPTIONS OF PROFICIENCY IN ENGLISH

Jonathan Schmidgall

One of the most critical activities in assessment is establishing the meaning of test scores and communicating it in terms that test takers and score users can understand. The meaning of test scores is elaborated in the definition of the ability to be assessed (i.e., the construct), established by validity research, and may be expanded by research that relates scores to practical information about test takers' abilities. The construct definition elaborates the knowledge, skills, and abilities to be evaluated by the test and is often based on theory and an analysis of the knowledge, skills, abilities, and tasks that commonly occur in real-world language use (Bachman & Palmer, 2010). The construct definition, once articulated and justified by theory and domain analysis, essentially becomes a claim about the meaning of test scores (Mislevy, 2013).

The redesigned *TOEIC Bridge*® tests aim to assess the listening, reading, speaking, and writing proficiency of beginning to lower–intermediate English language learners in the context of everyday life (Schmidgall et al., 2019). For each of the four testing components (listening, reading, speaking, and writing), a construct definition was developed based on a review of theory and influential language proficiency standards. For each test component (language skill), the construct definition elaborates the communication goals to be measured by the test and the linguistic knowledge and subcompetencies that are needed to achieve these goals. For example, in the construct definition of the Listening test section, test takers are expected to understand commonly occurring spoken texts as well as simple descriptions of people, places, objects, and actions (a communication goal). This requires using knowledge of common vocabulary and formulaic phrases (lexical knowledge, a component of linguistic knowledge). According to the construct definition, the ability to achieve each communication goal requires the use of multiple components of linguistic knowledge (e.g., lexical, grammatical, discourse, phonological).

The role of validity research is to investigate the extent to which claims about the meaning and use of test scores are supported by evidence (Schmidgall & Xi, 2020). One common approach in validity research is to examine the strength of the relationship between test scores and other measures of the same construct, or a criterion measure (American Educational Research Association et al., 2014). As Powers and Powers (2015) have pointed out, learner self-assessments provide useful information in a variety of contexts, including general education (Falchikov & Boud, 1989; J. Ross, 2006), personality research (Ackerman et al., 2002), occupational psychology (Mabe & West, 1982), and language learning (Bachman & Palmer, 1989; S. Ross, 1998). One of the potential advantages of self-assessment as a method for evaluating proficiency is that learners may have more complete knowledge of their strengths and weaknesses (Shrauger & Osberg, 1981; Upshur, 1975). However, self-assessments may have important limitations as well. Studies that have compared student self-assessments of language abilities with teacher or peer assessments have generally found that students rated themselves more severely than peers (Matsuno, 2009) and teachers (Iwamoto, 2015), and teacher judgments were more strongly

correlated with language test performance (Johansson, 2013). Thus, as S. Ross (1998) argued in his meta-analysis of self-assessments of language proficiency, self-assessments have been shown to be useful as criterion measures of proficiency, but the accuracy of self-assessments may be influenced by learners' experience with the specific task(s) described in the self-assessment instrument. Essentially, learners are more likely to provide accurate and useful self-assessments for tasks with which they have prior experience. Consequently, self-assessment ratings are likely to be influenced by both the sample of learners (their background and experiences) and the self-assessment instrument itself (its relevance to learners).

To further elaborate the meaning of test scores, research may also be conducted to map test scores to language proficiency standards or external measures of language proficiency (Papageorgiou et al., 2015). In the case of the redesigned TOEIC Bridge tests, influential language proficiency standards were carefully examined during the construct definition and task development phases of test design (see Everson et al., 2019; Schmidgall et al., 2019). This included the Common European Framework of Reference (CEFR) for Languages (Council of Europe, 2018), Canadian Language Benchmarks (CLB; Centre for CLB, 2012), and American Council on the Teaching of Foreign Languages (ACTFL) proficiency guidelines (ACTFL, 2012). These language proficiency standards include descriptors of the types of language knowledge and use that may be expected at different levels of proficiency, and TOEIC Bridge test tasks were designed to target different levels of proficiency (beginner, high beginner, low intermediate) based on a review of relevant descriptors in the CEFR, CLB, and ACTFL proficiency guidelines (Schmidgall et al., 2019). Consequently, the proposed alignment between levels of language proficiency standards and test scores may also inform expectations about additional types of activities that test takers should be able to perform at different score levels. Typically, each level of a language proficiency standard is associated with a wide range of descriptions of both expected language use and specific communicative activities. Although the ability to perform some of these communicative activities may be directly measured by a test, many are not. A mapping study can provide convincing evidence of correspondence between test scores and specific levels of a language proficiency standard (one that is presumably sufficiently backed by research). In such cases, it may be reasonable to expect that test takers at a particular level are able to perform tasks associated with that level even if the tasks are not directly measured by the test. Regardless, one may expect that tasks associated with higher levels of language proficiency based on language proficiency standards would be perceived as increasingly more difficult to perform by TOEIC Bridge test takers.

THE CURRENT STUDIES

The studies described in this paper investigated the meaningfulness of redesigned TOEIC Bridge test scores by comparing performance on the test to self-assessments of language ability. This investigation examined claims about meaningfulness in several ways.

Research Question 1: What is the relationship between redesigned TOEIC Bridge test scores and self-assessments?

First, the strength of the relationship between scores on each TOEIC Bridge test component (listening, reading, speaking, and writing) and self-assessments of these abilities was examined in order to determine the extent to which test scores are related to an external measure of the same ability.

Research Question 2: What activities do test takers at different proficiency levels report being able to perform?

Second, the meaningfulness of scores may be expanded by elaborating the types of activities that test takers at different levels of proficiency report being able to perform. This information may supplement score-based interpretations by elaborating the types of real-world activities that test takers at each proficiency level report being able to perform with a specified degree of confidence.

Research Question 3: To what extent do test takers report being able to perform activities as expected on the basis of their redesigned TOEIC Bridge test scores (proficiency levels)?

Finally, the relative difficulty of different communicative activities for test takers at different levels of proficiency, as indicated in self-assessments, can be examined to determine the extent to which it aligns with theoretical expectations based on test design. The redesigned TOEIC Bridge tests were designed with the expectation that learners at some levels of proficiency should be more (or less) able than test takers at other levels to accomplish particular tasks. Therefore, this analysis may provide evidence to support assumptions about the meaning of test scores based on construct definition and the test design process. Consequently, one might expect that a reasonable percentage of test takers at beginning levels of proficiency, as measured by the test, should report mainly being able to perform tasks associated with only beginner levels of proficiency based on language proficiency standards (i.e., CEFR Levels A1 to A2; CLB Level 1 to 4; ACTFL Novice High), and a higher percentage of test takers at intermediate levels of proficiency as measured by the test should report being able to perform tasks associated with intermediate levels of proficiency based on language proficiency standards.

Two studies were conducted to investigate each of these research questions in relation to the redesigned TOEIC Bridge Speaking and Writing tests (Study 1) and the TOEIC Bridge Listening and Reading test (Study 2). Due to practical constraints, the studies were performed 6 months apart using different samples of test takers.

Study 1: Speaking and Writing

Test takers who participated in the redesigned TOEIC Bridge tests field study (see Lin et al., 2019) were invited to complete an online self-assessment survey approximately 2 months after the field test. In total, 1,659 test takers from Japan and Taiwan were invited, and 1,056 participated, a response rate of 64%. The response rate was higher in Japan (n = 935, response rate of 70%) and lower in Taiwan (n = 121, response rate of 30%). The distribution of TOEIC Bridge Speaking and Writing test scores of respondents from each country was similar to the field study, although Taiwanese respondents were slightly more proficient than the overall sample of Taiwanese participants in the field study. As shown in Table 1, the subgroups varied somewhat in terms of their demographic characteristics: The Taiwanese sample had relatively more female respondents and was relatively younger in terms of average age. The samples also differed in terms of the proportion identifying as employed (full- or part-time) or students: A majority of Japanese respondents indicated they were employed (72% employed, 24% students), whereas a majority of Taiwanese respondents indicated they were students (58% students, 41% employed).

TABLE 1

Demographic Characteristics of Participants in the Speaking and Writing Can-Do Survey

TOEIC Bridge Speaking and Writing can-do survey sample	*n*	TOEIC Bridge Speaking *M* (*SD*)	TOEIC Bridge Writing *M* (*SD*)	% Female	Age in years *M* (*SD*)
Japan	935	36.60 (8.35)	41.63 (7.96)	48	34.1 (11.4)
Taiwan	121	37.89 (10.12)	40.99 (9.16)	67	25.0 (9.1)

The online self-assessment survey consisted of a series of *can-do statements* that described various communicative tasks that involved speaking or writing skills. The speaking section included 24 statements. Seven statements were included based on their relevance to the TOEIC Bridge Speaking test construct definition, which elaborates the communication goals and linguistic skills the test measures (see Schmidgall et al., 2019). For example, one of the communication goals assessed is the ability to ask for and provide basic information; this was included as the can-do statement, "ask for and provide basic information about everyday topics." The remaining 17 statements were based on descriptors drawn from proficiency levels from three different language proficiency standards (ACTFL Novice High to Intermediate High, CEFR A1 to B2, CLB 1 to 6). The writing section also included 24 statements, five based on their relevance to the TOEIC Bridge Writing test construct definition and the remaining 19 based on their relevance to language proficiency standards. The statements were selected from standards in order to represent a range of tasks across proficiency levels (stratified by proficiency level) and distinct activities (to avoid too much overlap between descriptors within each section). In the online survey, items were randomly ordered within each section (speaking and writing), and the order in which each section was presented was counterbalanced.

The survey was originally drafted in English and then translated into participants' first languages (Japanese and Taiwanese Mandarin) prior to administration. After reading each statement (e.g., "When speaking in English, I can ask for and provide basic information about everyday topics"), participants selected a response to indicate their ability to perform the communicative task (1 = *Cannot do at all*; 2 = *Can do with great difficulty*; 3 = *Can do with some difficulty*; 4 = *Can do with little difficulty*; 5 = *Can do easily*). For each language skill, can-do statements were identified and coded to correspond to (a) the communication goals targeted by the relevant TOEIC Bridge test construct definition, and/or (b) communicative activities described in various language proficiency standards, including CEFR, CLB, and ACTFL.

Analysis

After data were collected for the surveys, a validity check was conducted to identify and screen out unmotivated responses from the analysis. The validity check identified participants whose response times suggested they did not read a substantial portion of the items (*speeders*) and participants whose responses across items were unreasonably invariant (*invariant responders*). Speeders (n = 47) were identified by comparing response times in the online survey to benchmarks established by research assistants who were instructed to complete the survey as quickly as possible. Invariant responders (n = 25) were identified by examining the distribution of standard deviations of participants' mean response to items in the survey. After excluding participants whose mean response was at extreme ends of the scale—potentially valid response patterns whose standard deviations would necessarily be small—a cutoff point was identified to indicate unreasonably invariant respondents. In total, 72 participants in the survey were screened out in the validity check, reducing the overall sample to 984 participants for the analysis (Japan = 873; Taiwan = 111).

For each can-do scale (speaking and writing), a scale analysis was conducted for each subgroup (Japan and Taiwan) and the overall sample. Research has shown that self-assessment of language abilities can vary based on background factors (Iwamoto, 2015), and differences between the subgroups potentially include proficiency level, age, and cultural background. The scale analysis included estimates of item difficulty, item–total correlations, and estimates of reliability using Cronbach's alpha. Self-assessment scores for each skill were estimated by calculating the average of responses to individual self-assessment items. To answer the first research question, the relationship between TOEIC Bridge test scaled scores and self-assessment scores for each subpopulation and the overall sample was quantified using Pearson correlations. Correlations may range from −1.00 (perfect negative relationship) to +1.00 (perfect positive relationship) and can be interpreted as the strength of the relationship between two measures. A conventional standard in social science research is to interpret correlations of .50 and above as "large," and correlations between .30 and .50 as "medium" (Cohen, 1988), but this recommendation was updated by Plonsky and Oswald (2014) to .60 and above as large and .40 to .59 as medium based on their broad review of studies in second language research.

To answer the second research question, tables were prepared for each self-assessment scale that displayed, for each TOEIC Bridge test proficiency level (1 to 4), the percentage of participants who indicated they were able to perform each communicative task. Because participants rated the degree of effort needed to perform each task, the ordinal scale of ratings (1 to 5) needed to be transformed to dichotomous ratings (not able to do, able to do). In previous research, different standards have been applied to rescale can-do ratings (Ito et al., 2005; Powers et al., 2009). In line with previous research, we considered two standards for rescaling ratings: defining "likely able to do" by ratings of "with some difficulty," "with little difficulty," and "easily" (less stringent standard) and by ratings of "with little difficulty" and "easily" (more stringent standard). Ultimately, we used the less stringent standard based on two reasons. First, we considered the interpretability of results when using each standard. Second, lower proficiency learners in the Japan and Taiwan testing populations have been historically more likely to focus on listening and reading than on speaking and writing skills. As a result, they may be expected to have comparatively less experience and confidence in their ability to perform speaking and writing tasks, and this is likely to be reflected in their self-assessments. Thus, the tables for speaking and writing still indicate the speaking and writing tasks that participants think they can do, but the results also reflect the comparatively lower degree of confidence that participants may have in these abilities. In other words, when participants reported being able to perform a task (but only with some difficulty), we gave them the benefit of the doubt, classifying them as likely able to do.

After the tables were prepared, the pattern of results was analyzed to determine the extent to which they conformed to expectations in order to answer the third research question. Based on the design of the test and an initial CEFR mapping study (Schmidgall, 2019), test takers at TOEIC Bridge test proficiency Level 1 should be able to perform some tasks associated with CEFR Level Pre-A1. Test takers at proficiency Level 2 should be able to perform tasks associated with CEFR Level Pre-A1 and some tasks associated with CEFR Level A1. Test takers at proficiency Level 3 should be able to perform tasks associated with CEFR Levels Pre-A1 and A1, and some tasks associated with CEFR Level A2. Test takers at proficiency Level 4 should be able to perform tasks associated with CEFR Levels Pre-A1 to A2, and some tasks associated with CEFR Levels B1 and above.

Results of Study 1

Table 2 shows the correlations between TOEIC test Speaking and Writing scores and test takers' assessments of their ability to perform the can-do tasks, as defined by the average of their responses to each can-do scale. All the measures had adequate reliability (internal consistency): The reliability of TOEIC Bridge Speaking and Writing scores using stratified alpha ranged from alpha = .78 to .87 (see Lin et al., 2019), and the reliability of the can-do speaking and writing scales using coefficient alpha ranged from alpha = .97 to .99.

TABLE 2

Correlations Among Speaking and Writing Can-Do Self-Assessments and TOEIC Bridge Scores for the Overall Sample (N = 984)

Measure	*M* (*SD*)	TOEIC Bridge Speaking score	TOEIC Bridge Writing score	Can-do speaking score
TOEIC Bridge score				
Speaking	36.74 (8.48)			
Writing	41.58 (8.06)	.64		
Can-do score				
Speaking	2.85 (0.78)	.51	.40	
Writing	3.12 (0.80)	.49	.46	.82

Note. All correlations are significant at the $p < .001$ level.

As shown in Table 2, TOEIC Bridge Speaking test scores had a medium correlation with self-assessed speaking skills ($r = .51$), and TOEIC Bridge Writing test scores had a medium correlation with self-assessed writing skills ($r = .46$). This relationship was similar for Japanese and Taiwanese participants. Generally, TOEIC Bridge Speaking and Writing test scores had medium correlations with self-assessed speaking and writing skills for the Japanese ($r = .51$ and 44, respectively) and Taiwanese ($r = .54$ and .58) participants.

Tables 3 and 4 show, for each task in the survey, the percentages of test takers at each TOEIC Bridge Speaking and Writing score level who thought they could perform the task easily or with little difficulty. The list of tasks is arranged by easiest to most difficult, as indicated by the mean response on the original rating scale (1 to 5) for each task. The correlation between TOEIC Bridge test scores and ratings for each task is also shown in the table. Correlations ranged from $r = .37$ to .46 (median $r = .425$) for speaking tasks, and from $r = .31$ to .43 (median $r = .38$) for writing tasks. The tables also employ a highlighting convention used in previous research in order to more clearly indicate patterns in overall percentages of test takers who believed they could perform each task across proficiency levels (e.g., Powers, Bravo, et al., 2008; Powers, Kim, & Weng, 2008; Powers et al., 2009). For the convenience of score users, these results may also be summarized to indicate the tasks that test takers report being able to perform (or not perform) at each TOEIC Bridge Speaking and Writing test proficiency level. Following the convention and rationale of previous research, for each TOEIC Bridge Speaking or Writing proficiency level we indicated the tasks that test takers indicated they (a) probably can do, (b) probably can do with difficulty, and (c) probably cannot do (Powers et al., 2009). These can-do table summaries are provided in Appendices A (speaking) and B (writing).

In Tables 3 and 4, as the percentage of test takers who report being able to perform the task increases, the color shading get darker. Thus, when viewed from left to right, the pattern of color shading is a rough visual indicator of the percentages of test takers who report being able to perform each task (i.e., with no, little, or some difficulty), ordered by TOEIC Bridge test proficiency level. When viewed from top to bottom, the pattern of color shading is a rough indicator of the percentages of test takers who can perform each task at each proficiency level, ordered from easiest to most difficult task.

TABLE 3

Percentages of TOEIC Bridge Test Takers, by Speaking Proficiency Level, Who Indicated They Could Perform Various English Speaking Tasks Easily, With Little Difficulty, or With Some Difficulty

		TOEIC Bridge Speaking proficiency level							
		1	2	3	4				
		TOEIC Bridge Speaking scaled score							
ID#	Descriptor ("I can…")	15–22	23–36	37–42	43–50	*M*	*SD*	Correlation with TOEIC Bridge Speaking	Corresponding standard(s)
S15	ask a few simple, formulaic questions in social situations (for example: "How are you?," "Where are you from?," "What do you do for fun?")	71[d]	89[e]	97[f]	99[f]	3.82	0.94	.43	ACTFL Nov-H
S19	give basic personal information in response to a direct question from a supportive listener (for example: your name, where you are from)	58[c]	84[e]	92[f]	99[f]	3.56	0.96	.44	CLB1
S08	read aloud a very short, rehearsed statement	52[c]	82[e]	91[f]	95[f]	3.53	1.00	.40	CEFR A1
S04	give simple directions	49[b]	72[d]	85[e]	93[f]	3.20	0.93	.42	CEFR A2; CLB3; ACTFL Int-M
S09	give a short, rehearsed basic presentation on a familiar subject	43[b]	72[d]	80[e]	88[e]	3.17	1.00	.37	CEFR A2
S20	open a short conversation with someone who is familiar and supportive	44[b]	61[c]	78[d]	94[f]	3.17	1.05	.45	CLB2
S07	use simple phrases and sentences to describe where I live and people I know	42[b]	66[c]	81[e]	92[f]	3.13	0.96	.46	CEFR A1
S17	ask a variety of questions to obtain simple information about everyday things (for example: directions, prices, and services)	38[b]	65[c]	78[d]	93[f]	3.13	0.98	.46	ACTFL Int-M
S06	make simple requests, offers, and suggestions	36[b]	61[c]	80[e]	92[f]	3.03	0.94	.43	CEFR A2+; CLB5; ACTFL Int-M-H
S01	ask for and provide basic information about everyday topics	31[b]	57[c]	74[d]	91[f]	2.97	0.96	.45	CEFR A1; CLB1-2; ACTFL Nov-H
S11	can explain what I like or dislike about something	31[b]	58[c]	72[d]	89[e]	2.93	0.96	.43	CEFR A2+
S21	give simple, common, routine instructions and directions to a familiar person	31[b]	49[b]	66[c]	84[e]	2.83	0.97	.43	CLB3
S22	participate in a very short, simple phone call with a familiar person	22[a]	44[b]	65[c]	84[e]	2.81	1.08	.43	CLB4
S02	describe people, objects, places, and activities	29[a]	51[c]	67[c]	85[e]	2.81	0.92	.39	CEFR A2; CLB2; ACTFL Int-L
S05	narrate and sequence simple events	25[a]	46[b]	65[c]	87[e]	2.80	0.93	.46	CEFR A2+; CLB4

ID#	Descriptor ("I can…")	TOEIC Bridge Speaking proficiency level 1 / TOEIC Bridge Speaking scaled score 15–22	2 / 23–36	3 / 37–42	4 / 43–50	*M*	*SD*	Correlation with TOEIC Bridge Speaking	Corresponding standard(s)
S10	handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself	32[b]	39[b]	56[c]	76[d]	2.64	1.02	.37	CEFR A2
S03	express an opinion or plan and give a reason for it	21[a]	36[b]	57[c]	78[d]	2.62	0.94	.45	CEFR B1; CLB5-6; ACTFL Int-M-H
S13	give detailed accounts of experiences, describing feelings and reactions	18[a]	31[b]	47[b]	72[d]	2.48	0.97	.42	CEFR B1
S16	use simple words and phrases fluently and accurately in social situations	13[a]	25[a]	44[b]	64[c]	2.37	1.01	.37	ACTFL Int-L
S12	narrate a story or relate the plot of a book or film and describe my reactions	18[a]	22[a]	43[b]	65[c]	2.34	0.92	.39	CEFR B1
S18	converse with ease and confidence when dealing with everyday tasks and social situations	13[a]	24[a]	40[b]	64[c]	2.33	0.94	.40	ACTFL Int-H
S23	agree, disagree, and give opinions in small group discussions or meetings	10[a]	22[a]	43[b]	62[c]	2.29	0.96	.40	CLB5
S14	explain a viewpoint on a topical issue giving the advantages and disadvantages of various options	9[a]	22[a]	37[b]	57[c]	2.22	0.92	.37	CEFR B2
S24	give a detailed presentation (~7 min long) about a familiar topic	6[a]	19[a]	33[b]	56[c]	2.14	0.98	.37	CLB6
	Sample size for score interval	77	318	300	289				

Note. ACTFL = American Council on the Teaching Foreign Languages; CEFR = Common European Framework of Reference; CLB = Canadian Language Benchmarks. The pattern of color shading is a rough indicator of the percentages of test takers who can perform each task at each proficiency level.

[a][0–29] [b][30–49] [c][50–69] [d][70–79] [e][80–89] [f][90–100]

TABLE 4

Percentages of TOEIC Bridge Test Takers, by Writing Proficiency Level, Who Indicated They Could Perform Various English Writing Tasks Easily, With Little Difficulty, or With Some Difficulty

		TOEIC Bridge Writing proficiency level							
		1	2	3	4				
		TOEIC Bridge Writing scaled score							
ID#	Descriptor ("I can…")	15–19	20–31	32–42	43-50	*M*	*SD*	Correlation with TOEIC Bridge Writing	Corresponding standard(s)
W20	write basic personal identification information, words, simple phrases, and a few sentences about highly familiar information related to everyday life	62[c]	74[d]	88[e]	97[f]	3.71	0.94	.39	CLB2
W09	write a series of simple phrases and sentences linked with simple connectors like "and," "but," and "because"	48[b]	73[d]	83[e]	97[f]	3.65	0.96	.43	CEFR A2
W08	write very simple messages and personal online postings as a series of very short sentences about hobbies, likes/dislikes, etc., relying on the aid of a translation tool	71[d]	76[d]	82[e]	95[f]	3.60	1.00	.35	CEFR A1
W07	write simple isolated phrases and sentences	38[b]	61[c]	74[d]	92[f]	3.37	0.99	.40	CEFR A1
W06	post simple online greetings, using basic formulaic expressions and emoticons	48[b]	61[c]	76[d]	86[e]	3.31	1.01	.31	CEFR Pre-A1
W21	write 3–5 sentences describing a familiar person	29[a]	54[c]	69[c]	89[e]	3.24	0.98	.39	CLB3
W01	ask for and provide basic information about everyday topics	24[a]	49[b]	70[d]	89[e]	3.17	0.93	.40	CEFR A1
W19	copy numbers, letters, words, short phrases, or sentences from simple lists or very short passages, for personal use or to complete short tasks	48[b]	57[c]	68[c]	86[e]	3.17	0.98	.32	CLB1
W05	make simple requests, offers, and suggestions	38[b]	53[c]	69[c]	88[e]	3.16	0.95	.39	CEFR A2+
W15	write simple sentences on very familiar topics	33[b]	55[c]	64[c]	88[e]	3.16	0.96	.38	CEFR A1; ACTFL Nov-H
W22	complete simple forms that require basic personal information or familiar information and some responses to 15–20 simple questions	38[b]	56[c]	66[c]	87[e]	3.16	0.98	.39	CLB4
W12	write basic emails or letters to request information	24[a]	49[b]	66[c]	85e	3.14	0.99	.39	CEFR B1
W04	narrate and sequence simple events	29[a]	48[b]	66[c]	86[e]	3.09	0.95	.40	CEFR A2+

ID#	Descriptor ("I can…")	TOEIC Bridge Writing proficiency level: 1 / TOEIC Bridge Writing scaled score: 15–19	2 / 20–31	3 / 32–42	4 / 43-50	*M*	*SD*	Correlation with TOEIC Bridge Writing	Corresponding standard(s)
W11	write very short, basic descriptions of events, past activities, and personal experiences	24[a]	52[c]	63[c]	85[e]	3.08	0.97	.38	CEFR A2+
W17	write short, simple communications, compositions, and requests for information about personal preferences, daily routines, common events, and other personal topics	24[a]	51[c]	59[c]	84[e]	3.07	0.98	.37	ACTFL Int-M
W02	describe people, objects, places, and activities	33[b]	50[c]	64[c]	84[e]	3.06	0.93	.35	CEFR A2
W10	engage in basic social communication online (e.g., writing a simple message on a virtual card for special occasions, sharing news, and making/confirming arrangements to meet)	24[a]	50[c]	63[c]	78[d]	3.02	1.02	.35	CEFR A2
W16	write statements and formulate questions based on familiar topics	24[a]	37[b]	53[c]	79[d]	2.90	0.98	.40	CEFR A2; ACTFL Int-L
W13	make personal online postings about experiences, feelings, and events and respond individually to the comments of others in some detail, though my vocabulary may be limited	10[a]	39[b]	52[c]	76[d]	2.88	0.99	.36	CEFR B1
W03	express a simple opinion and give a reason for it	10[a]	37b	51[c]	79[d]	2.88	0.96	.40	CEFR A2+
W18	write compositions and simple summaries related to work or school experiences	19[a]	40[b]	52[c]	76[d]	2.86	0.97	.36	CEFR B1; ACTFL Int-H
W23	write a paragraph to describe the sequence of an everyday routine	24[a]	38[b]	48[b]	75[d]	2.84	0.97	.38	CLB5
W24	write 1–2 paragraphs about a familiar topic, expressing a main idea and supporting it with some detail	10[a]	33[b]	43[b]	68[c]	2.69	0.99	.37	CLB6
W14	write a short essay or report, passing on information or giving reasons in support of, or against, a particular point of view	14[a]	31[b]	35[b]	64[c]	2.58	1.00	.35	CEFR B2
	Sample size for score interval	21	94	272	597				

Note. ACTFL = American Council on the Teaching Foreign Languages; CEFR = Common European Framework of Reference; CLB = Canadian Language Benchmarks. The pattern of color shading is a rough indicator of the percentages of test takers who can perform each task at each proficiency level.

[a][0–29] [b][30–49] [c][50–69] [d][70–79] [e][80–89] [f][90–100]

The percentage of participants who indicated they could perform each of the tasks in Tables 3 and 4 increased across each TOEIC Bridge test proficiency level. For example, the first task in Table 4 is "write basic personal identification information, words, simple phrases, and a few sentences about highly familiar information related to everyday life" (ID# W20). As the TOEIC Bridge Writing test proficiency level increased from 1 to 4, the percentage of participants who indicated they could perform the task increased from 62% to 74% (Level 1 to 2), from 74% to 88% (Level 2 to 3), and from 88% to 97% (Level 3 to 4). If TOEIC Bridge test proficiency levels were poor indicators of test takers' speaking and writing proficiency at beginning to low intermediate levels, we would expect to observe a less consistent pattern of results. Across both surveys, all 48 tasks conformed to this pattern.

In addition, Tables 3 and 4 show the language proficiency standards and levels that correspond to each task in the survey. Overall, the percentages of test takers at different TOEIC Bridge Speaking and Writing proficiency levels who reported being able to perform different speaking tasks corresponds to what one might expect based on language proficiency standards. In the case of CEFR descriptors, tasks corresponding to CEFR Levels Pre-A1, A1, A2, A2+, B1, and B2 have been included in the survey. As tasks are rated increasingly difficult to perform by participants, CEFR levels associated with tasks generally increase. For example, in Table 3, the speaking task "read aloud a very short, rehearsed statement" (ID# S08) is associated with CEFR Level A1. TOEIC Bridge test proficiency Level 2 is associated with this CEFR level (see Schmidgall, 2019), and 82% of participants at this level reported being able to perform this task with some degree of confidence. Also in Table 3, the task "narrate a story or relate the plot of a book or film and describe my reactions" (ID# S12) is associated with CEFR Level B1. Only 22% of participants at TOEIC Bridge Speaking proficiency Level 2 reported being able to perform this task, whereas 65% of participants at proficiency Level 4, associated with CEFR Level B1, reported being able to perform the task. In general, a similar pattern can be observed for tasks associated with CLB proficiency levels (from 1 to 6) and ACTFL proficiency levels (from Novice High to Intermediate High).

Summary of the Results of Study 1

TOEIC Bridge Speaking and Writing test scores had medium correlations with self-assessments of speaking (r = .51) and writing (r = .46). Although these are not large correlations, they compare favorably to the results of similar validity studies that have used self-assessments as a criterion measure of speaking and writing skills. In a study of the relationship between TOEIC Speaking and Writing test scores and self-assessments of speaking and writing ability, Powers et al. (2009) estimated similar correlations for speaking (r = .54) and writing (r = .52). Li (2015) examined the relationship between the Michigan English Placement Test (MEPT; **www.michiganassessment.org/blog/category/mept**) Writing scores and self-assessments of writing ability (r = .37), and between *TOEFL iBT*® Speaking and Writing scores and self-assessments of speaking (r = .37) and writing (r = .22) ability. In an earlier study, Powers et al. (2003) investigated the relationship between LanguEdge Speaking and Writing scores and self-assessments of speaking (r = .43) and writing (r = .26) ability. With this context in mind, the results of this study provide support for the claim that TOEIC Bridge Speaking and Writing test scores are meaningful indicators of speaking and writing ability. In addition, the pattern of results is largely consistent with expectations based on the design of the test and its consideration of relevant language proficiency standards.

Study 2: Listening and Reading

Test takers who participated in pretesting of redesigned TOEIC Bridge Listening and Reading test forms in Japan (n = 2,063) and Taiwan (n = 3,109) also completed a paper-based self-assessment survey. As shown in Table 5, the mean TOEIC Bridge Reading and Listening test scores were higher for the Taiwanese sample compared to the Japanese sample of participants. Among the Japanese participants who reported demographic information, approximately 33% were female, and the average age was 16 (ages ranged from 10 to 20). Among the Taiwanese participants who reported demographic information, approximately 58% were female, and the average age was also 16 (ages ranged from 11 to 24). The majority of Japanese participants were enrolled in high school (55%), and most Taiwanese participants were enrolled in vocational high schools (81%).

TABLE 5

Demographic Characteristics of Participants in the Listening and Reading Can-Do Survey

TOEIC Bridge Listening and Reading can-do survey sample	*n*	TOEIC Bridge Listening *M* (*SD*)	TOEIC Bridge Reading *M* (*SD*)	% Female	Age in years *M* (*SD*)
Japan	2,063	27.21 (7.25)	31.38 (7.56)	33	16.9 (0.8)
Taiwan	3,109	34.73 (9.00)	38.15 (8.32)	58	16.7 (2.8)

The development of the listening and reading survey mirrored the approach used for the speaking and writing survey in Study 1. The survey was developed in English and then translated for administration to participants in local languages. Unlike the speaking and writing survey, however, the reading and listening survey largely emphasized tasks from one set of language proficiency standards, the CEFR. This was done in order to make the listening and reading survey more comparable to self-assessments administered for the legacy version of the TOEIC Bridge Listening and Reading tests, which only utilized descriptors from the CEFR (e.g., Powers, Bravo, et al., 2008; Powers & Simpson, 2008; Powers et al., 2013; Powers & Yan, 2013; Powers, Kim, & Weng, 2008).

The paper-based survey consisted of can-do statements that described communicative tasks that involved listening or reading skills. The listening section included 20 statements, seven based on their relevance to the TOEIC Bridge Listening test construct definition, and the remaining 13 based on their relevance to the CEFR. The reading section included 19 statements, six based on their relevance to the TOEIC Bridge Reading test construct definition and the remaining 13 based on their relevance to the CEFR. Similar to Study 1, the statements were selected from the CEFR in order to represent a range of tasks across proficiency levels (stratified by proficiency level) and distinct activities (to avoid too much overlap between descriptors within each section).

Analysis

Because the listening and reading survey was paper-based, the initial validity check was only able to include an analysis of invariant responses; it was unable incorporate an analysis of response times. Using the same procedure to identify invariant responders as described for the speaking and writing survey in Study 1,587 participants were screened out in the validity check, reducing the overall sample of participants in the listening and reading survey to 4,585 for the analysis (Japan = 1,918; Taiwan = 2,667).

Using the same approach as Study 1, scale analysis was conducted for the can-do scale (listening and reading) and included estimates of item difficulty, item–total correlations, and estimates of reliability using Cronbach's alpha. Self-assessment scores for each skill were estimated by calculating the average of responses to individual self-assessment items. The relationship between TOEIC Bridge scaled scores and self-assessment scores for each subpopulation and the overall sample was calculated via Pearson correlations. Finally, tables were prepared for each self-assessment scale that estimated the percentage of participants at each TOEIC Bridge test score level (1 to 4) that indicated they were likely to be able to perform each communicative task. As in Study 1, two different standards were considered for rescaling the results for the purpose of these tables. Ultimately, the more stringent standard was used after considering the interpretability of results and the expectation that learners in this population have been historically more likely to focus (and be assessed) on their listening and reading skills. This approach is also consistent with similar research that has been conducted with this learner population (e.g., Powers et al., 2008; Powers & Simpson, 2008; Powers & Yan, 2013). After the tables were prepared, the pattern of results was analyzed to determine the extent to which it conformed to expectations in order to answer the third research question.

Results of Study 2

Table 6 shows the correlations between TOEIC Listening and Reading scores and test takers' self-assessments of their ability to perform reading and listening tasks. Again, all measures had adequate reliability: The reliability of TOEIC Bridge Listening and Reading scores using coefficient alpha has ranged from alpha = .88 to .93 (see Lin et al., 2019), and the reliability of the can-do listening and reading scales using coefficient alpha ranged from alpha = .96 to .98.

TABLE 6

Correlations Among Listening and Reading Can-Do Self-Assessments and TOEIC Bridge Test Scores for the Overall Sample (N = 4,585)

Measure	*M* (*SD*)	TOEIC Bridge Listening score	TOEIC Bridge Reading score	Can-do listening score
TOEIC Bridge test score				
Listening	31.56 (9.01)			
Reading	35.24 (8.54)	.79		
Can-do score				
Listening	3.78 (0.74)	.55	.52	
Reading	3.65 (0.77)	.54	.54	.87

Note. All correlations are significant at the $p < .001$ level.

As shown in Table 6, TOEIC Bridge Listening test scores had a medium correlation with self-assessed listening skills ($r = .55$), and TOEIC Bridge Reading test scores had a medium correlation with self-assessed reading skills ($r = .54$). Again, this relationship differed slightly by subgroups. TOEIC Bridge Listening and Reading test scores had a large correlation with self-assessed listening and reading skills for the Taiwanese ($r = .61$ and .59, respectively) participants as compared to small correlations for the Japanese ($r = .28$ and .28) participants. This difference does not appear to be attributable to a difference in the reliability (internal consistency) of can-do scores across subpopulations, as the measures of internal consistency of the listening and reading can-do scales for Japanese participants (alpha = .96, .97) and Taiwanese participants (alpha = .98, .98) were high.

Tables 7 and 8 indicate the percentages of test takers at each TOEIC Bridge Listening and Reading score level that we defined (according to their reports) as likely to be able to perform each of the tasks in the survey. Again, the list of tasks is arranged by easiest to most difficult based on the mean response on the original rating scale (1 to 5) for each task. The correlations between TOEIC Bridge test scores and ratings for each task ranged from $r = .38$ to .51 (median $r = .46$) for listening tasks and from $r = .36$ to .49 (median $r = .43$) for reading tasks. The tables use the same highlighting convention introduced earlier, and results are summarized by proficiency levels in Appendices C (listening) and D (reading) using the same method described in Study 1.

TABLE 7

Percentages of TOEIC Bridge Test Takers, by Listening Proficiency Level, Who Indicated They Could Perform Various English Listening Tasks Easily or With Little Difficulty

		TOEIC Bridge Listening proficiency level							
		1	2	3	4				
		TOEIC Bridge Listening scaled score							
ID#	Descriptor ("I can…")	15	16–25	26–38	39–50	*M*	*SD*	Correlation with TOEIC Bridge Listening	Corresponding standard(s)
L08	understand simple questions in social situations such as "How are you?" and "Where do you live?"	56[c]	76[d]	88[e]	97[f]	4.30	0.77	.38	CEFR Pre-A1
L09	identify a few common key words and expressions (for example, "Help!" and "Watch out!")	49[b]	71[d]	83[e]	96[f]	4.21	0.83	.38	CEFR Pre-A1
L10	recognize familiar words and simple phrases when people speak slowly and clearly	42[b]	65[c]	80[e]	94[f]	4.10	0.83	.41	CEFR A1
L11	understand short, simple instructions addressed carefully and slowly	39[b]	60[c]	77[d]	94[f]	4.04	0.85	.43	CEFR A1
L05	understand simple greetings and introductions	46[b]	63[c]	77[d]	94[f]	4.03	0.83	.43	CEFR Pre-A1 to A2
L03	understand short announcements when they are spoken slowly and clearly	37[b]	59[c]	76[d]	94[f]	4.03	0.85	.44	CEFR A1 to A2
L12	understand questions addressed carefully and slowly	38[b]	55[c]	73[d]	93[f]	3.97	0.87	.44	CEFR A1
L13	understand simple, everyday conversations if conducted slowly and clearly	38[b]	52[c]	72[d]	93[f]	3.95	0.87	.45	CEFR A2
L01	understand simple descriptions of people, places, objects, and actions	35[b]	48[b]	69[c]	92[f]	3.92	0.89	.46	CEFR A1 to A2
L02	understand short conversations related to everyday life (for example, making a purchase)	31[b]	42[b]	65[c]	91[f]	3.83	0.90	.49	CEFR A1 to A2
L04	understand words and phrases that are commonly used in everyday life, relating to people, places, things, and basic activities	33[b]	44[b]	63[c]	88[e]	3.79	0.88	.45	CEFR Pre-A1 to A2
L18	understand someone who is speaking slowly and deliberately about his or her hobbies and interests	27[a]	39[b]	62[c]	88[e]	3.75	0.91	.47	CEFR B1
L14	understand when speakers agree and disagree in a conversation conducted slowly and clearly	31[b]	38[b]	59[c]	87[e]	3.72	0.95	.47	CEFR A2+

ID#	Descriptor ("I can…")	TOEIC Bridge Listening proficiency level: 1	2	3	4	*M*	*SD*	Correlation with TOEIC Bridge Listening	Corresponding standard(s)
		TOEIC Bridge Listening scaled score: 15	16–25	26–38	39–50				
L15	understand the main point of simple messages and short, clear announcements	29[a]	36[b]	56[c]	85[e]	3.66	0.93	.46	CEFR A2
L16	generally identify the topic of a conversation around me if the speakers are talking slowly and clearly	27[a]	34[b]	57[c]	85[e]	3.65	0.96	.46	CEFR A2+
L17	understand the main points and important details in stories (for example, a description of a vacation), provided the speaker talks slowly and clearly	19[a]	32[b]	51[c]	82[e]	3.58	0.95	.46	CEFR B1
L07	understand the main idea in short announcements or talks	19[a]	29[a]	50[c]	84[e]	3.56	0.93	.51	CEFR A2 to B1
L19	understand a person in social situations talking about his or her background, family, or interests	17[a]	25[a]	43[b]	76[d]	3.42	0.98	.47	CEFR B1+
L06	understand a request that is indirect or implied	16[a]	16[a]	30[b]	70[d]	3.19	1.01	.48	CEFR B1
L20	understand extended speech and lectures and follow complex arguments on familiar topics	11[a]	13[a]	25[a]	61[c]	2.99	1.106739	.46	CEFR B2
	Sample size for score interval	167	1112	2180	1126				

Note. CEFR = Common European Framework of Reference. The pattern of color shading is a rough indicator of the percentages of test takers who can perform each task at each proficiency level.

[a][0–29] [b][30–49] [c][50–69] [d][70–79] [e][80–89] [f][90–100]

TABLE 8

Percentages of TOEIC Bridge Test Takers, by Reading Proficiency Level, Who Indicated They Could Perform Various English Listening Tasks Easily or With Little Difficulty

ID#	Descriptor ("I can…")	TOEIC Bridge Reading proficiency level: 1 / TOEIC Bridge Reading scaled score: 15–18	2 / 19–33	3 / 34–44	4 / 45–50	*M*	*SD*	Correlation with TOEIC Bridge Reading	Corresponding standard(s)
R07	understand simple everyday signs such as "Parking," "Station," "Stop"	46[b]	67[c]	84[e]	94[f]	4.10	0.83	.36	CEFR Pre-A1
R08	recognize familiar words if they are accompanied by pictures, such as in a menu	47[b]	63[c]	81[e]	94[f]	4.03	0.83	.40	CEFR Pre-A1
R10	understand familiar words and very simple sentences	38[b]	56[c]	80[e]	94[f]	3.97	0.86	.44	CEFR A1
R02	understand short informational and descriptive texts about people, places, objects, and actions	40[b]	55[c]	78[d]	92[f]	3.92	0.86	.43	CEFR A2
R01	understand short, simple correspondence	41[b]	52[c]	74[d]	90[f]	3.88	0.89	.41	CEFR A2
R09	find and understand simple, important information such as costs, dates, and locations in reading material that has visuals such as brochures or advertisements	35[b]	50[c]	72[d]	91[f]	3.83	0.89	.43	CEFR A1
R04	understand nonlinear written texts (for example, signs, schedules)	34[b]	47[b]	70[d]	91[f]	3.79	0.89	.42	CEFR A2
R13	understand a train or bus schedule	36[b]	44[b]	69[c]	90[f]	3.74	0.94	.42	CEFR A2
R11	understand short, simple messages in texts, emails, or on social networks	32[b]	42[b]	68[c]	89[e]	3.71	0.95	.44	CEFR A2
R03	understand written instructions and directions (for example, a basic recipe, simple travel directions)	29[a]	40[b]	66[c]	89[e]	3.68	0.95	.46	CEFR A2
R15	identify specific information in short text or articles that are written in simple language	31[b]	37[b]	65[c]	89[e]	3.66	0.94	.47	CEFR A2+
R14	understand simple, step-by-step instructions	30[b]	35[b]	64[c]	88[e]	3.63	0.98	.47	CEFR A2
R12	understand a simple email from a friend	30[b]	34[b]	63[c]	87[e]	3.61	0.97	.47	CEFR A2
R06	understand the main idea and stated details in short, written texts	26[a]	33[b]	61[c]	88[e]	3.60	0.92	.49	CEFR B1 to B1+
R17	understand the main points of an article on a familiar topic	23[a]	30[b]	58[c]	84[e]	3.51	1.01	.48	CEFR B1

ID#	Descriptor ("I can…")	TOEIC Bridge Reading proficiency level 1	2	3	4	*M*	*SD*	Correlation with TOEIC Bridge Reading	Corresponding standard(s)
		TOEIC Bridge Reading scaled score 15–18	19–33	34–44	45–50				
R16	read information about products (for example, advertisements)	27[a]	29[a]	52[c]	79[d]	3.45	0.97	.43	CEFR B1
R05	infer the meaning of unknown written words through context clues	27[a]	25[a]	46[b]	77[d]	3.34	1.00	.43	CEFR A2+ to B1+
R18	understand the viewpoints expressed in articles and reports about contemporary issues or problems	15[a]	17[a]	38[b]	68[c]	3.08	1.09	.42	CEFR B2
R19	understand a popular novel	15[a]	11[a]	30[b]	62[c]	2.88	1.12	.43	CEFR B2
	Sample size for score interval	186	1643	2005	751				

Note. CEFR = Common European Framework of Reference. The pattern of color shading is a rough indicator of the percentages of test takers who can perform each task at each proficiency level.

[a][0–29] [b][30–49] [c][50–69] [d][70–79] [e][80–89] [f][90–100]

For almost all of the tasks in Tables 7 and 8, the percentage of participants who indicated they could perform the task increased across each redesigned TOEIC Bridge test proficiency level. For example, the first task in Table 8 is "understand simple everyday signs such as 'Parking,' 'Station,' 'Stop'" (ID# R07). As the TOEIC Bridge Reading test proficiency level increased from 1 to 4, the percentage of participants who indicated they could perform the task increased from 47% to 67% (Level 1 to 2), from 67% to 84% (Level 2 to 3), and from 84% to 94% (Level 3 to 4). If TOEIC Bridge test proficiency levels are poor indicators of test takers' listening and reading proficiency at beginning to low intermediate levels, we would expect to observe a less consistent pattern of results. Across both surveys, only three of 39 tasks violated this pattern (ID# L06, R05, R19), and for these tasks, the discrepancy was between estimates at the lowest levels of proficiency with respect to their ability to perform more difficult tasks.

Tables 7 and 8 also show the CEFR levels that correspond to relevant tasks in the survey. Some of the tasks are directly related to the construct definition of the TOEIC Bridge Listening or Reading test and may be relevant to multiple CEFR levels; consequently, these tasks are not directly relevant to a specific CEFR level. Overall, the pattern of results conforms to the expectations that (a) participants indicated they were increasingly less able to perform tasks as associated CEFR proficiency levels increased from Pre-A1 to B2, and (b) the percentage of participants at each TOEIC Bridge test proficiency level who indicated they were likely to perform each task was largely consistent with the task's classification in terms of its CEFR proficiency level. For example, in Table 7, the listening task "understand simple questions in social situations" (ID# L08) is associated with CEFR Level Pre-A1. TOEIC Bridge test proficiency Level 1 is associated with this CEFR level, and 56% of participants at this level reported being able to perform this task. In comparison, the listening task "understand the main points of simple messages and short, clear

announcements" (ID# L15) is associated with CEFR Level A2. Only 29% of participants at TOEIC Bridge test proficiency Level 1 indicated they could perform this task, whereas 56% of participants at TOEIC Bridge test proficiency Level 3, associated with CEFR Level A2, indicated they could perform the task. Although the degree of correspondence varied somewhat across items, the overall pattern was consistent with expectations.

Summary of the Results of Study 2

Redesigned TOEIC Bridge Listening and Reading test scores had medium correlations with self-assessments of listening ($r = .55$) and reading ($r = .54$). Again, these results compare favorably with previous research that examined the relationship between reading and listening test scores and self-assessments. Validity studies for the legacy version of the TOEIC Bridge test found correlations assessments ranging from $r = .35$ to .51 between listening test scores and self-assessments, and ranging from $r = .22$ to .49 between reading test scores and self-assessments (Powers, Bravo, et al., 2008; Powers & Simpson, 2008; Powers & Yan, 2013). Thus, the results of this study provide empirical support for the claim that TOEIC Bridge Listening and Reading test scores are meaningful indicators of listening and reading ability. In addition, the pattern of results is generally consistent with expectations based on the design of the test and its consideration of relevant language proficiency standards.

DISCUSSION

The results of Studies 1 and 2 provide evidence to support the claim that redesigned TOEIC Bridge test scores are meaningful indicators of test takers' beginning to intermediate English listening, reading, speaking, and writing proficiency in the context of everyday life. The studies found medium correlations between TOEIC Bridge test scores and self-assessments of test takers' ability to perform everyday listening ($r = .55$), reading ($r = .54$), speaking ($r = .51$), and writing ($r = .46$) tasks relevant to beginning to intermediate levels of English proficiency. The strength of these correlations compares favorably with the results of similar validity studies, as discussed in the summary of each study. In addition, the pattern of results across TOEIC Bridge proficiency levels for each task suggests that TOEIC Bridge tests are able to clearly differentiate test takers at beginning to intermediate levels of English proficiency. Put more simply, higher performing TOEIC Bridge test takers were much more likely to report that they could perform each task. Finally, the pattern of results across tasks for each language skill suggests that TOEIC Bridge proficiency levels are reasonably well aligned with expectations regarding the kinds of tasks that test takers at each level should be able to perform based on how each proficiency level has been theorized.

The results of this study also provide information that may be referenced by score users to clarify the meaning of TOEIC Bridge test scores as they pertain to proficiency levels. The tables produced by the study (i.e., Tables 3, 4, 7, and 8) provide some indication of the extent to which test takers at different proficiency levels may be able to complete tasks of varying complexity, and the accompanying Appendices A, B, C, and D summarize these tasks by language skill and proficiency level. This information can be used to get a broader sense of what learners at different proficiency levels can be expected to accomplish and provides additional evidence to support claims about TOEIC Bridge test score mapping with language proficiency standards such as the CEFR.

Several important limitations should be kept in mind when interpreting the results of this study, including the estimates provided for individual tasks in the can-do surveys (Tables 3, 4, 7, and 8, and the appendices). First, the results are based on samples of test takers in Japan and Taiwan, and estimates may be expected to vary across different subpopulations of test takers. Second, our study included relatively small samples of test takers at some proficiency levels (e.g., proficiency Level 1 for reading and listening), and the overall sample used for Study 1 (speaking and writing) is relatively small; larger samples may be expected to produce more robust estimates. Third, self-assessments may be expected to be more accurate for tasks that learners have previously experienced (Ross, 1998). For example, it is unlikely that test takers at low English proficiency levels have attempted to read a popular novel in English (reading task ID# R19), so self-assessments at these levels involve a higher degree of inference on the part of learners. In comparing the results of Study 1 and Study 2, it is important to keep in mind that the studies involved slightly different populations of test takers. Although both studies involved samples of test takers in Japan and Taiwan who would be included in the target population of TOEIC Bridge test takers, participants in Study 1 were generally much older than participants in Study 2 (the average age in Japan was 34.1 for Study 1 and 16.9 for Study 2). In addition, self-assessments were collected at the same time as TOEIC Bridge test scores for Study 2, but self-assessments were collected approximately 2 months after TOEIC Bridge test scores were obtained in Study 1. Due to the potential interaction between learner characteristics (e.g., experience) and self-assessments, direct comparisons between the results of the studies should be made with caution. Finally, test takers classified at the highest proficiency level on the redesigned TOEIC Bridge test (Level 4) may vary in terms of their proficiency level (from low intermediate to advanced) because the test is not designed to discriminate levels of more advanced proficiency. Consequently, inferences about what test takers at TOEIC Bridge test proficiency Level 4 are able to do should be made more cautiously.

The method used in this study builds on previous validity studies using self-assessments by including can-do descriptors that were more purposefully linked to expectations about what test takers should be able to do at different proficiency levels based on the design of the test and its relation to language proficiency standards. Language proficiency standards, such as the CEFR, typically use can-do descriptors to exemplify performance at different levels of proficiency. This design is a natural fit for self-assessment and establishes expectations that provide a basis for interpreting self-assessment ratings by test takers. It is important to note that descriptors in language proficiency standards are often conceptualized and ordered based on expert judgment and may evolve over time and that individual learner profiles with respect to descriptors may vary. Consequently, it is probably unreasonable to expect perfect alignment between proficiency levels and self-assessment ratings, even if proficiency levels were derived from an assessment built with a specific set of language proficiency standards in mind (see Summers et al., 2019). With this important caveat, this study shows how the use of standards-based descriptors may enhance the use of self-assessments in validity research by establishing clearer expectations regarding how test takers' responses to specific tasks may be evaluated.

REFERENCES

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What do we really know about our abilities and knowledge. *Personality and Individual Differences, 33*(4), 587–605. **https://doi.org/10.1016/S0191-8869(01)00174-X**

American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines 2012*. **http://www.actfl.org/sites/default/files/pdfs/ACTFLProficiencyGuidelines2012_FINAL.pdf**

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

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing, 6*(1), 14–29. **https://doi.org/10.1177/026553228900600104**

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

Centre for Canadian Language Benchmarks. (2012). *Canadian language benchmarks: English as a second language for adults.* **http://www.cic.gc.ca/english/pdf/pub/language-benchmarks.pdf**

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Erlbaum.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment (CEFR)*. **https://www.coe.int/lang-cefr**

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.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59*(4), 395–430. **https://doi.org/10.3102/00346543059004395**

Ito, T., Kawaguchi, K., & Ohta, R. (2005). *A study of the relationship between TOEIC scores and functional job performance: Self-assessment of foreign language proficiency* [White paper]. The Institute for International Business Communications. **http://www.iibc-global.org/library/redirect_only/library/toeic_data/toeic_en/pdf/newsletter/1_E.pdf**

Iwamoto, N. (2015). *Effects of L2 affective factors on self-assessment of speaking* [Unpublished doctoral dissertation]. Temple University.

Johansson, S. (2013). The relationship between students' self-assessed reading skills and other measures of achievement. *Large-Scale Assessments in Education, 1*(3), 1–17. **https://doi.org/10.1186/2196-0739-1-3**

Li, Z. (2015). Using an English self-assessment tool to validate an English placement test. *Papers in Language Testing and Assessment, 4*(1), 59–96.

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

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67*(3), 280–296. **https://doi.org/10.1037/0021-9010.67.3.280**

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100. **https://doi.org/10.1177/0265532208097337**

Mislevy, R. J. (2013). Modeling language for assessment. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics.* Blackwell. **https://doi.org/10.1002/9781405198431.wbeal0770**

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**

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. **https://doi.org/10.1111/lang.12079**

Powers, D. E., Bravo, G. M., Sinharay, S., Saldivia, L. E., Simpson, A. G., & Weng, V. Z. (2008). *Relating scores on the TOEIC Bridge to student perceptions of proficiency in English* (Research Memorandum No. RM-08-02). ETS.

Powers, D. E., Kim, H.-J., & Weng, V. (2008). *The redesigned TOEIC (Listening and Reading) test: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-08-56). ETS. **https://doi.org/10.1002/j.2333-8504.2008.tb02142.x**

Powers, D. E., Kim, H.-J., Yu, F., Weng, V. Z., & VanWinkle, W. (2009). *The TOEIC Speaking and Writing tests: Relations to test-taker perceptions of proficiency in English* (Research Report No. RR-09-18). ETS. **https://doi.org/10.1002/j.2333-8504.2009.tb02175.x**

Powers, D. E., Mercadante, R., & Yan, F. (2013). Validating TOEIC Bridge scores against teacher ratings for vocational students in China. In D. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies: Volume II* (pp. 4.0–4.11). ETS.

Powers, D. E., & Powers, A. (2015). The incremental contribution of *TOEIC*® Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing, 32*(2), 151–167. **https://doi.org/10.1177/0265532214551855**

Powers, D. E., Roever, C., Huff, K. L., & Trapani, C. S. (2003). *Validating LanguEdge courseware scores against faculty ratings and student self-assessments* (Research Report No. RR-03-11). ETS. **https://doi.org/10.1002/j.2333-8504.2003.tb01903.x**

Powers, D. E., & Simpson, A. (2008). *Validating TOEIC Bridge scores against teacher and student ratings: A small-scale study* (Research Memorandum No. RM-08-03). ETS.

Powers, D. E., & Yan, F. (2013). TOEIC Bridge scores: Validity evidence from Korea and Japan. In D. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies: Volume II* (pp. 5.0–5.10). ETS.

Ross, J. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation, 11*(10), 1–13. **http://hdl.handle.net/1807/30005**

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1–20. **https://doi.org/10.1177/026553229801500101**

Schmidgall, J. (2019). *Mapping the redesigned TOEIC Bridge tests to the CEFR* (Research Memorandum No. RM-19-09). ETS.

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**

Schmidgall, J., & Xi, X. (2020). Validation of language assessments. In C. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 1123–1158). Wiley.

Shrauger, J. S., & Osberg, T.M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. *Psychological Bulletin, 90*(2), 322–351. **https://doi.org/10.1037/0033-2909.90.2.322**

Summers, M. M., Cox, T. L., McMurry, B. L., & Dewey, D. P. (2019). Investigating the use of the ACTFL can-do statements in a self-assessment for student placement in an intensive English program. *System, 80*, 269–287. **https://doi.org/10.1016/j.system.2018.12.012**

Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967–1974* (pp. 53–65). TESOL.

APPENDIX A. CAN-DO TABLE FOR TOEIC BRIDGE SPEAKING

	Task Speaking Scaled Score 15 to 22 (Proficiency Level 1)
Probably can do	None
Probably can do with difficulty	Ask a few simple, formulaic questions in social situations (for example: "How are you?," "Where are you from?," "What do you do for fun?") Give basic personal information in response to a direct question from a supportive listener (for example: your name, where you are from) Read aloud a very short, rehearsed statement
Probably cannot do	Give simple directions Give a short, rehearsed, basic presentation on a familiar subject Open a short conversation with someone who is familiar and supportive Use simple phrases and sentences to describe where I live and people I know Ask a variety of questions to obtain simple information about everyday things (for example: directions, prices, and services) Make simple requests, offers, and suggestions Ask for and provide basic information about everyday topics Can explain what I like or dislike about something Give simple, common, routine instructions and directions to a familiar person Describe people, objects, places, and activities Participate in a very short, simple phone call with a familiar person Narrate and sequence simple events Handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself Express an opinion or plan and give a reason for it Give detailed accounts of experiences, describing feelings and reactions Use simple words and phrases fluently and accurately in social situations Narrate a story or relate the plot of a book or film and describe my reactions Converse with ease and confidence when dealing with everyday tasks and social situations Agree, disagree, and give opinions in small group discussions or meetings Explain a viewpoint on a topical issue giving the advantages and disadvantages of various options Give a detailed presentation (~7 minutes long) about a familiar topic

Speaking Scaled Score 23 to 36 (Proficiency Level 2)	
Probably can do	None
Probably can do with difficulty	Ask a few simple, formulaic questions in social situations (for example: "How are you?," "Where are you from?," "What do you do for fun?") Give basic personal information in response to a direct question from a supportive listener (for example: your name, where you are from) Read aloud a very short, rehearsed statement Give simple directions Give a short, rehearsed, basic presentation on a familiar subject Open a short conversation with someone who is familiar and supportive Use simple phrases and sentences to describe where I live and people I know Ask a variety of questions to obtain simple information about everyday things (for example: directions, prices, and services) Make simple requests, offers, and suggestions Ask for and provide basic information about everyday topics Can explain what I like or dislike about something Describe people, objects, places, and activities
Probably cannot do	Give simple, common, routine instructions and directions to a familiar person Participate in a very short, simple phone call with a familiar person Narrate and sequence simple events Handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself Express an opinion or plan and give a reason for it Give detailed accounts of experiences, describing feelings and reactions Use simple words and phrases fluently and accurately in social situations Narrate a story or relate the plot of a book or film and describe my reactions Converse with ease and confidence when dealing with everyday tasks and social situations Agree, disagree, and give opinions in small group discussions or meetings Explain a viewpoint on a topical issue giving the advantages and disadvantages of various options Give a detailed presentation (~7 minutes long) about a familiar topic

Speaking Scaled Score 37 to 42 (Proficiency Level 3)	
Probably can do	Ask a few simple, formulaic questions in social situations (for example: "How are you?," "Where are you from?," "What do you do for fun?") Give basic personal information in response to a direct question from a supportive listener (for example: your name, where you are from) Read aloud a very short, rehearsed statement
Probably can do with difficulty	Give simple directions Give a short, rehearsed, basic presentation on a familiar subject Open a short conversation with someone who is familiar and supportive Use simple phrases and sentences to describe where I live and people I know Ask a variety of questions to obtain simple information about everyday things (for example: directions, prices, and services) Make simple requests, offers, and suggestions Ask for and provide basic information about everyday topics Can explain what I like or dislike about something Give simple, common, routine instructions and directions to a familiar person Describe people, objects, places, and activities Participate in a very short, simple phone call with a familiar person Narrate and sequence simple events Handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself Express an opinion or plan and give a reason for it
Probably cannot do	Give detailed accounts of experiences, describing feelings and reactions Use simple words and phrases fluently and accurately in social situations Narrate a story or relate the plot of a book or film and describe my reactions Converse with ease and confidence when dealing with everyday tasks and social situations Agree, disagree, and give opinions in small group discussions or meetings Explain a viewpoint on a topical issue giving the advantages and disadvantages of various options Give a detailed presentation (~7 minutes long) about a familiar topic

Speaking Scaled Score 43 to 50 (Proficiency Level 4)	
Probably can do	Ask a few simple, formulaic questions in social situations (for example: "How are you?," "Where are you from?," "What do you do for fun?") Give basic personal information in response to a direct question from a supportive listener (for example: your name, where you are from) Read aloud a very short, rehearsed statement Give simple directions Give a short, rehearsed, basic presentation on a familiar subject Open a short conversation with someone who is familiar and supportive Use simple phrases and sentences to describe where I live and people I know Ask a variety of questions to obtain simple information about everyday things (for example: directions, prices, and services)
Probably can do with difficulty	Make simple requests, offers, and suggestions Ask for and provide basic information about everyday topics Can explain what I like or dislike about something Give simple, common, routine instructions and directions to a familiar person Describe people, objects, places, and activities Participate in a very short, simple phone call with a familiar person Narrate and sequence simple events Handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself Express an opinion or plan and give a reason for it Give detailed accounts of experiences, describing feelings and reactions Use simple words and phrases fluently and accurately in social situations Narrate a story or relate the plot of a book or film and describe my reactions Converse with ease and confidence when dealing with everyday tasks and social situations Agree, disagree, and give opinions in small group discussions or meetings Explain a viewpoint on a topical issue giving the advantages and disadvantages of various options Give a detailed presentation (~7 minutes long) about a familiar topic
Probably cannot do	None

APPENDIX B. CAN-DO TABLE FOR TOEIC BRIDGE WRITING

	Task Writing Scaled Score 15 to 19 (Proficiency Level 1)
Probably can do	None
Probably can do with difficulty	Write basic personal identification information, words, simple phrases, and a few sentences about highly familiar information related to everyday life Write very simple messages and personal online postings as a series of very short sentences about hobbies, likes/dislikes, etc., relying on the aid of a translation tool
Probably cannot do	Write a series of simple phrases and sentences linked with simple connectors like "and," "but," and "because" Write simple isolated phrases and sentences Post simple online greetings, using basic formulaic expressions and emoticons Write 3–5 sentences describing a familiar person Ask for and provide basic information about everyday topics Copy numbers, letters, words, short phrases, or sentences from simple lists or very short passages, for personal use or to complete short tasks Make simple requests, offers, and suggestions Write simple sentences on very familiar topics Complete simple forms that require basic personal information or familiar information and some responses to 15–20 simple questions Write basic emails or letters to request information Narrate and sequence simple events Write very short, basic descriptions of events, past activities, and personal experiences Write short, simple communications, compositions, and requests for information about personal preferences, daily routines, common events, and other personal topics Describe people, objects, places, and activities Engage in basic social communication online (e.g., writing a simple message on a virtual card for special occasions, sharing news, and making/confirming arrangements to meet) Write statements and formulate questions based on familiar topics Make personal online postings about experiences, feelings, and events and respond individually to the comments of others in some detail, though my vocabulary may be limited Express a simple opinion and give a reason for it Write compositions and simple summaries related to work or school experiences Write a paragraph to describe the sequence of an everyday routine Write 1–2 paragraphs about a familiar topic, expressing a main idea and supporting it with some detail Write a short essay or report, passing on information or giving reasons in support of or against a particular point of view

Writing Scaled Score 20 to 31 (Proficiency Level 2)	
Probably can do	None
Probably can do with difficulty	Write basic personal identification information, words, simple phrases, and a few sentences about highly familiar information related to everyday life Write a series of simple phrases and sentences linked with simple connectors like "and," "but," and "because" Write very simple messages and personal online postings as a series of very short sentences about hobbies, likes/dislikes, etc., relying on the aid of a translation tool Write simple isolated phrases and sentences Post simple online greetings, using basic formulaic expressions and emoticons Write 3–5 sentences describing a familiar person Copy numbers, letters, words, short phrases, or sentences from simple lists or very short passages, for personal use or to complete short tasks Make simple requests, offers, and suggestions Write simple sentences on very familiar topics Complete simple forms that require basic personal information or familiar information and some responses to 15–20 simple questions Write very short, basic descriptions of events, past activities, and personal experiences Write short, simple communications, compositions, and requests for information about personal preferences, daily routines, common events, and other personal topics Describe people, objects, places, and activities Engage in basic social communication online (e.g., writing a simple message on a virtual card for special occasions, sharing news, and making/confirming arrangements to meet)
Probably cannot do	Ask for and provide basic information about everyday topics Write basic emails or letters to request information Narrate and sequence simple events Write statements and formulate questions based on familiar topics Make personal online postings about experiences, feelings, and events and respond individually to the comments of others in some detail, though my vocabulary may be limited Express a simple opinion and give a reason for it Write compositions and simple summaries related to work or school experiences Write a paragraph to describe the sequence of an everyday routine Write 1–2 paragraphs about a familiar topic, expressing a main idea and supporting it with some detail Write a short essay or report, passing on information or giving reasons in support of or against a particular point of view

Writing Scaled Score 32 to 42 (Proficiency Level 3)	
Probably can do	None
Probably can do with difficulty	Write basic personal identification information, words, simple phrases, and a few sentences about highly familiar information related to everyday life Write a series of simple phrases and sentences linked with simple connectors like "and," "but," and "because" Write very simple messages and personal online postings as a series of very short sentences about hobbies, likes/dislikes, etc., relying on the aid of a translation tool Write simple isolated phrases and sentences Post simple online greetings, using basic formulaic expressions and emoticons Write 3–5 sentences describing a familiar person Ask for and provide basic information about everyday topics Copy numbers, letters, words, short phrases or sentences from simple lists or very short passages, for personal use or to complete short tasks Make simple requests, offers, and suggestions Write simple sentences on very familiar topics Complete simple forms that require basic personal information or familiar information and some responses to 15–20 simple questions Write basic emails or letters to request information Narrate and sequence simple events Write very short, basic descriptions of events, past activities, and personal experiences Write short, simple communications, compositions, and requests for information about personal preferences, daily routines, common events, and other personal topics Describe people, objects, places, and activities Engage in basic social communication online (e.g., writing a simple message on a virtual card for special occasions, sharing news, and making/confirming arrangements to meet) Write statements and formulate questions based on familiar topics Make personal online postings about experiences, feelings, and events and respond individually to the comments of others in some detail, though my vocabulary may be limited Express a simple opinion and give a reason for it Write compositions and simple summaries related to work or school experiences
Probably cannot do	Write a paragraph to describe the sequence of an everyday routine Write 1–2 paragraphs about a familiar topic, expressing a main idea and supporting it with some detail Write a short essay or report, passing on information or giving reasons in support of or against a particular point of view

Writing Scaled Score 43 to 50 (Proficiency Level 4)	
Probably can do	Write basic personal identification information, words, simple phrases, and a few sentences about highly familiar information related to everyday life Write a series of simple phrases and sentences linked with simple connectors like "and," "but," and "because" Write very simple messages and personal online postings as a series of very short sentences about hobbies, likes/dislikes, etc., relying on the aid of a translation tool Write simple isolated phrases and sentences
Probably can do with difficulty	Post simple online greetings, using basic formulaic expressions and emoticons Write 3–5 sentences describing a familiar person Ask for and provide basic information about everyday topics Copy numbers, letters, words, short phrases or sentences from simple lists or very short passages, for personal use or to complete short tasks Make simple requests, offers, and suggestions Write simple sentences on very familiar topics Complete simple forms that require basic personal information or familiar information and some responses to 15–20 simple questions Write basic emails or letters to request information Narrate and sequence simple events Write very short, basic descriptions of events, past activities, and personal experiences Write short, simple communications, compositions, and requests for information about personal preferences, daily routines, common events, and other personal topics Describe people, objects, places, and activities Engage in basic social communication online (e.g., writing a simple message on a virtual card for special occasions, sharing news, and making/confirming arrangements to meet) Write statements and formulate questions based on familiar topics Make personal online postings about experiences, feelings, and events and respond individually to the comments of others in some detail, though my vocabulary may be limited Express a simple opinion and give a reason for it Write compositions and simple summaries related to work or school experiences Write a paragraph to describe the sequence of an everyday routine Write 1–2 paragraphs about a familiar topic, expressing a main idea and supporting it with some detail Write a short essay or report, passing on information or giving reasons in support of or against a particular point of view
Probably cannot do	None

APPENDIX C. CAN-DO TABLE FOR REDESIGNED TOEIC BRIDGE LISTENING

	Task Listening Scaled Score 15 (Proficiency Level 1)
Probably can do	Understand simple questions in social situations such as "How are you?" and "Where do you live?"
Probably can do with difficulty	Recognize familiar words and simple phrases when people speak slowly and clearly Understand short, simple instructions addressed carefully and slowly Understand simple greetings and introductions Understand short announcements when they are spoken slowly and clearly Understand questions addressed carefully and slowly Understand simple, everyday conversations if conducted slowly and clearly Understand simple descriptions of people, places, objects, and actions Understand short conversations related to everyday life (for example, making a purchase) Understand words and phrases that are commonly used in everyday life, relating to people, places, things, and basic activities Understand someone who is speaking slowly and deliberately about his or her hobbies and interests Understand when speakers agree and disagree in a conversation conducted slowly and clearly Understand the main point of simple messages and short, clear announcements Generally identify the topic of a conversation around me if the speakers are talking slowly and clearly Understand the main points and important details in stories (for example, a description of a vacation), provided the speaker talks slowly and clearly Understand the main idea in short announcements or talks Understand a person in social situations talking about his or her background, family, or interests
Probably cannot do	Understand a request that is indirect or implied Understand extended speech and lectures, and follow complex arguments on familiar topics

Listening Scaled Score 16 to 25 (Proficiency Level 2)	
Probably can do	Understand simple questions in social situations such as "How are you?" and "Where do you live?" Identify a few common key words and expressions (for example, "Help!" "Watch out!") Recognize familiar words and simple phrases when people speak slowly and clearly Understand short, simple instructions addressed carefully and slowly Understand simple greetings and introductions Understand short announcements when they are spoken slowly and clearly Understand questions addressed carefully and slowly Understand simple, everyday conversations if conducted slowly and clearly
Probably can do with difficulty	Understand simple descriptions of people, places, objects, and actions Understand short conversations related to everyday life (for example, making a purchase) Understand words and phrases that are commonly used in everyday life, relating to people, places, things, and basic activities Understand someone who is speaking slowly and deliberately about his or her hobbies and interests Understand when speakers agree and disagree in a conversation conducted slowly and clearly Understand the main point of simple messages and short, clear announcements Generally identify the topic of a conversation around me if the speakers are talking slowly and clearly Understand the main points and important details in stories (for example, a description of a vacation), provided the speaker talks slowly and clearly Understand the main idea in short announcements or talks Understand a person in social situations talking about his or her background, family, or interests Understand a request that is indirect or implied
Probably cannot do	Understand extended speech and lectures, and follow complex arguments on familiar topics

Listening Scaled Score 26 to 38 (Proficiency Level 3)	
Probably can do	Understand simple questions in social situations such as "How are you?" and "Where do you live?" Identify a few common key words and expressions (for example, "Help!" "Watch out!") Recognize familiar words and simple phrases when people speak slowly and clearly Understand short, simple instructions addressed carefully and slowly Understand simple greetings and introductions Understand short announcements when they are spoken slowly and clearly Understand questions addressed carefully and slowly Understand simple, everyday conversations if conducted slowly and clearly Understand simple descriptions of people, places, objects, and actions Understand short conversations related to everyday life (for example, making a purchase) Understand words and phrases that are commonly used in everyday life, relating to people, places, things, and basic activities Understand someone who is speaking slowly and deliberately about his or her hobbies and interests Understand when speakers agree and disagree in a conversation conducted slowly and clearly Understand the main point of simple messages and short, clear announcements Generally identify the topic of a conversation around me if the speakers are talking slowly and clearly Understand the main points and important details in stories (for example, a description of a vacation), provided the speaker talks slowly and clearly Understand the main idea in short announcements or talks
Probably can do with difficulty	Understand a person in social situations talking about his or her background, family, or interests Understand a request that is indirect or implied Understand extended speech and lectures, and follow complex arguments on familiar topics
Probably cannot do	None

Listening Scaled Score 39 to 50 (Proficiency Level 4)	
Probably can do	Understand simple questions in social situations such as "How are you?" and "Where do you live?" Identify a few common key words and expressions (for example, "Help!" "Watch out!") Recognize familiar words and simple phrases when people speak slowly and clearly Understand short, simple instructions addressed carefully and slowly Understand simple greetings and introductions Understand short announcements when they are spoken slowly and clearly Understand questions addressed carefully and slowly Understand simple, everyday conversations if conducted slowly and clearly Understand simple descriptions of people, places, objects, and actions Understand short conversations related to everyday life (for example, making a purchase) Understand words and phrases that are commonly used in everyday life, relating to people, places, things, and basic activities Understand someone who is speaking slowly and deliberately about his or her hobbies and interests Understand when speakers agree and disagree in a conversation conducted slowly and clearly Understand the main point of simple messages and short, clear announcements Generally identify the topic of a conversation around me if the speakers are talking slowly and clearly Understand the main points and important details in stories (for example, a description of a vacation), provided the speaker talks slowly and clearly Understand the main idea in short announcements or talks Understand a person in social situations talking about his or her background, family, or interests Understand a request that is indirect or implied Understand extended speech and lectures, and follow complex arguments on familiar topics
Probably can do with difficulty	None
Probably cannot do	None

APPENDIX D. CAN-DO TABLE FOR REDESIGNED TOEIC BRIDGE READING

	Task Reading Scaled Score 15 to 18 (Proficiency Level 1)
Probably can do	None
Probably can do with difficulty	Understand simple everyday signs such as "Parking," "Station," "Stop" Recognize familiar words if they are accompanied by pictures, such as in a menu Understand familiar words and very simple sentences Understand short informational and descriptive texts about people, places, objects, and actions Understand short, simple correspondence Find and understand simple, important information such as costs, dates, and locations in reading material that has visuals such as brochures or advertisements Understand nonlinear written texts (for example, signs, schedules) Understand a train or bus schedule Understand short, simple messages in texts, emails, or on social networks Understand written instructions and directions (for example: a basic recipe, simple travel directions) Identify specific information in short text or articles that are written in simple language Understand simple, step-by-step instructions Understand a simple email from a friend Understand the main idea and stated details in short, written texts Understand the main points of an article on a familiar topic Read information about products (for example, advertisements) Infer the meaning of unknown written words through context clues Understand the viewpoints expressed in articles and reports about contemporary issues or problems
Probably cannot do	Understand a popular novel

Reading Scaled Score 19 to 33 (Proficiency Level 2)	
Probably can do	Understand simple everyday signs such as "Parking," "Station," "Stop" Recognize familiar words if they are accompanied by pictures, such as in a menu Understand familiar words and very simple sentences Understand short informational and descriptive texts about people, places, objects, and actions Understand short, simple correspondence Find and understand simple, important information such as costs, dates, and locations in reading material that has visuals such as brochures or advertisements
Probably can do with difficulty	Understand nonlinear written texts (for example, signs, schedules) Understand a train or bus schedule Understand short, simple messages in texts, emails, or on social networks Understand written instructions and directions (for example: a basic recipe, simple travel directions) Identify specific information in short text or articles that are written in simple language Understand simple, step-by-step instructions Understand a simple email from a friend Understand the main idea and stated details in short, written texts Understand the main points of an article on a familiar topic Read information about products (for example, advertisements) Infer the meaning of unknown written words through context clues Understand the viewpoints expressed in articles and reports about contemporary issues or problems
Probably cannot do	Understand a popular novel

Reading Scaled Score 34 to 44 (Proficiency Level 3)	
Probably can do	Understand simple everyday signs such as "Parking," "Station," "Stop" Recognize familiar words if they are accompanied by pictures, such as in a menu Understand familiar words and very simple sentences Understand short informational and descriptive texts about people, places, objects, and actions Understand short, simple correspondence Find and understand simple, important information such as costs, dates, and locations in reading material that has visuals such as brochures or advertisements Understand nonlinear written texts (for example, signs, schedules) Understand a train or bus schedule Understand short, simple messages in texts, emails, or on social networks Understand written instructions and directions (for example: a basic recipe, simple travel directions) Identify specific information in short text or articles that are written in simple language Understand simple, step-by-step instructions Understand a simple email from a friend Understand the main idea and stated details in short, written texts Understand the main points of an article on a familiar topic Read information about products (for example, advertisements)
Probably can do with difficulty	Infer the meaning of unknown written words through context clues Understand the viewpoints expressed in articles and reports about contemporary issues or problems Understand a popular novel
Probably cannot do	None

Reading Scaled Score 45 to 50 (Proficiency Level 4)	
Probably can do	Understand simple everyday signs such as "Parking," "Station," "Stop" Recognize familiar words if they are accompanied by pictures, such as in a menu Understand familiar words and very simple sentences Understand short informational and descriptive texts about people, places, objects, and actions Understand short, simple correspondence Find and understand simple, important information such as costs, dates, and locations in reading material that has visuals such as brochures or advertisements Understand nonlinear written texts (for example, signs, schedules) Understand a train or bus schedule Understand short, simple messages in texts, emails, or on social networks Understand written instructions and directions (for example: a basic recipe, simple travel directions) Identify specific information in short text or articles that are written in simple language Understand simple, step-by-step instructions Understand a simple email from a friend Understand the main idea and stated details in short, written texts Understand the main points of an article on a familiar topic Read information about products (for example, advertisements) Infer the meaning of unknown written words through context clues Understand the viewpoints expressed in articles and reports about contemporary issues or problems Understand a popular novel
Probably can do with difficulty	None
Probably cannot do	None

MAKING THE CASE FOR THE QUALITY AND USE OF A NEW LANGUAGE PROFICIENCY ASSESSMENT: VALIDITY ARGUMENT FOR THE REDESIGNED *TOEIC BRIDGE*® TESTS

Jonathan Schmidgall, Jaime Cid, Elizabeth Carter Grissom, and Lucy Li

An assessment should be designed to measure knowledge, skills, and abilities for a purpose, and stakeholders—test takers, score users, and others affected by an assessment—should approach the enterprise with some healthy skepticism. In a world where a variety of seemingly comparable assessments may appear to meet a specific need—such as an evaluation of language skills for an admissions or placement decision—an understanding of the basic principles of test quality and appropriate test use can help stakeholders more critically assess marketing claims and their own preconceptions about assessment. Research on stakeholders' conceptions of assessment has shown that personal beliefs and attitudes toward assessment, as well as understanding of the principles of effective test use and the purposes of assessment, can vary substantially among individuals and groups (Brown, 2008). For score users, having adequate assessment literacy, or being able to know the difference between sound and unsound assessment (Stiggins, 1995), can help maximize the beneficial outcomes and minimize the negative consequences of using an assessment to make decisions or maximize the overall usefulness of the assessment.

One fundamental aspect of assessment literacy is understanding the basic principles behind the proper use of language tests, including essential concepts such as reliability, validity, and fairness (Davies, 2008). Reliability is fundamentally about consistency, typically the consistency of test scores. As traditionally conceived, validity pertains to the meaning of scores and whether they mean what they are intended to mean. Fairness is about the absence of bias or whether the assessment disadvantages one group versus another. If reliability is low, test scores are inconsistent and a test taker's score may primarily depend on the rater, the specific form of the test taken, or any other number of factors irrelevant to the ability being tested. If validity and fairness are shown to be minimal or limited, scores will not be meaningful or may provide information that is not impartial or relevant enough for their intended use.

These principles are interdependent because weakness or strength in one aspect of measurement quality may influence or have implications for another. For example, an assessment that produces inconsistent scores (low reliability) is unlikely to produce very meaningful scores (weak validity). But even when an assessment produces highly consistent scores—such as when automated scoring is used—it may result in interpretations about ability that are extremely narrow, limited, or inadequate for their intended use (potentially weak validity). Given the complexity of these principles and their interdependence in practice, how can stakeholders—even those with sufficient assessment literacy—evaluate whether an assessment is designed to meet their needs?

The argument-based approach to validation has proven to be a promising framework for articulating and evaluating claims about measurement quality and assessment use and has been widely adopted in language testing (Schmidgall & Xi, 2020). In the argument-based approach, test developers systematically specify a series of claims about qualities of the assessment and its attended use and provide evidence to support those claims (Kane, 2006). This framework provides several benefits. It is flexible, comprehensive, and "helps us make sense of disparate lines of evidence and argument" (Mislevy, 2012, p. 94). By making the claims and evidence for test use explicit, it promotes transparency and identifies weaknesses in the argument for test use (Bachman, 2005). This approach also focuses less on the philosophical foundations of validity—which can be difficult for nonexperts to navigate—by placing the focus on specific claims made by a test developer (Kane & Bridgeman, 2017).

One approach to constructing a validity argument is the assessment use argument (AUA; Bachman & Palmer, 2010). The AUA was originally developed in the context of language assessment and has been utilized for the *TOEIC*® tests (Schmidgall, 2017). The AUA consists of four major claims, typically about the qualities of test scores, interpretations about test takers' abilities based on scores, decisions based on score interpretations, and consequences of decisions and of the use of the test. In a sense, the AUA presents a simplified narrative about the complex process of assessment from test administration and scoring to the appropriate and effective use of test scores. These four major claims encompass traditional qualities of measurement such as reliability, validity, and fairness; they also specify the role of important stakeholders (e.g., test takers, score users) in the use of an assessment.

The AUA is structured as a hierarchical set of statements (claims) made by the test developer regarding how test scores should be interpreted and used to make decisions. Each claim represents an inference made based on data and is elaborated by more specific statements (warrants). Warrants are supported by evidence (backing) and subject to criticism (rebuttals). The AUA is evaluated by examining the plausibility of the claims, particularly in light of backing and rebuttals for its warrants.

Figure 1 summarizes the AUA for the *TOEIC Bridge*® tests. The TOEIC Bridge tests were designed to measure beginning to low-intermediate English language proficiency in the context of everyday adult life (Schmidgall et al., 2019). The TOEIC Bridge tests include modules for listening and reading, speaking, and writing. For each of the skills tested—up to four, depending on a score user's needs—a score is reported that is intended to be interpreted as a measure of listening, reading, speaking, or writing proficiency. If an evaluation of overall language proficiency is needed, all four skills should be tested. The TOEIC Bridge tests were designed to support three primary intended uses: selection, placement, and evaluation of readiness for more advanced study or evaluation.

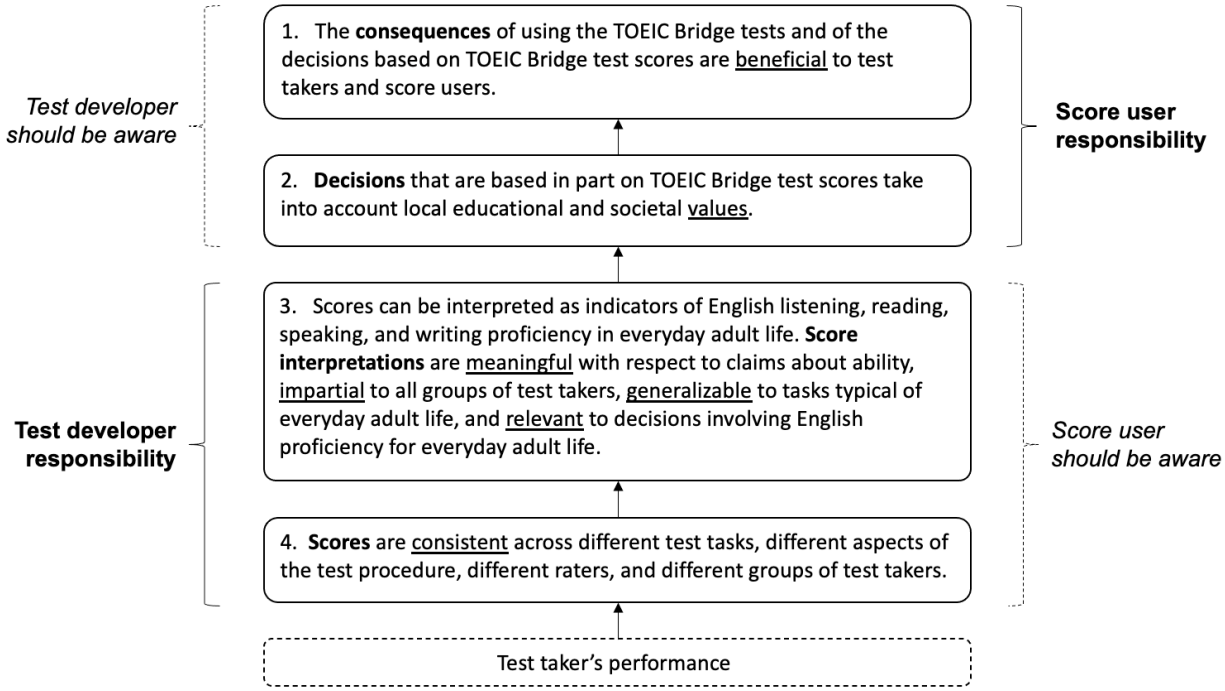

Figure 1. Assessment use argument for the redesigned TOEIC Bridge Tests.

The AUA starts with the assumption that the purpose of a test is to bring about beneficial consequences (see Figure 1, Claim 1): for example, to select a student whose English-speaking proficiency is appropriately suited to a training program. The consequences of the decision should be beneficial to the organization, the student selected, and ultimately students who did not get selected as they presumably would have been placed in a training program that was much too difficult or inappropriate to their current level of development. Test developers should collaborate with organizations to produce evidence—or identify problems—related to this fundamental goal of the use of a test.

To bring about these consequences, decisions must be made based on test scores, which are indicators of ability (see Figure 1, Claim 2). The decision-making procedure should be equitable and consider existing social and organizational values as well as legal requirements. Test developers should provide guidance to organizations and collect evidence to support claims about how a test may be used to make decisions.

When considering claims about the measurement quality of a test, it may be easier to start from the test taker's performance itself (see Figure 1, lowest box). A test taker's performance is based on the types of items or tasks, the number of items, and the content involved in the items. These features of test design are critical in that they determine the sample of language ability that is evaluated, influence score consistency (reliability) and how scores should be interpreted (validity), and ultimately frame how scores are used to make decisions.

Test scores are derived from test takers' performances, and their most important quality is consistency (see Figure 1, Claim 4). Scores may be assigned by a single rater or multiple raters using a variety of rubrics, scoring rules, and transformations; whatever the method used, a score assigned to a test taker

based on a test performance should be consistent. In other words, variation in scores should be related to language ability and should not depend on the particular rater, task, or form of the test a test taker receives.

Scores are used to make interpretations about ability (see Figure 1, Claim 3). Although it is vital for scores to be consistent, the interpretation of scores need to be valid and fair: meaningful, impartial, and generalizable beyond the test. In addition, interpretations about ability based on scores need to be relevant and sufficient for the decision to be made. Test developers must make explicit claims about score interpretations and back them with multifaceted evidence from test design and research.

The design and research basis of the test should provide evidence to justify claims about scores and their interpretations, and collaborative research with test users can provide evidence to support the use of test scores to make decisions. As shown in Figure 1, the test developer is responsible for specifying and supporting claims about scores and their interpretations (Claims 3 and 4), whereas the test developer should work with score users to support claims about test use (Claims 1 and 2; Bachman & Palmer, 2010). This is the structure of an argument for test use, a research-based approach to promoting higher quality decision-making and test use.

The warrants and evidence supporting each claim are elaborated in the rest of this document, beginning with claims about the consistency or reliability of scores (Figure 1, Claim 4) and the interpretation of scores (Figure 1, Claim 3). Because the test developer is principally responsibility for backing Claims 3 and 4, these claims will be given the most attention in this paper. Taken together, Claims 3 and 4 essentially reflect the argument and evidence for the measurement quality of the TOEIC Bridge tests, encompassing the traditional qualities of reliability, validity, and fairness.

Claims about test use—the decisions based on test scores, and their consequences—are strongly influenced by the particular decision-making context, over which the test developer has more limited control. Nevertheless, the TOEIC Bridge tests were initially designed for three primary uses, and these uses and their intended outcomes are elaborated in claims about decisions (Figure 1, Claim 2) and consequences (Figure 1, Claim 1) of the use of the TOEIC Bridge tests.

The AUA presented in Figure 1 (and elaborated in the rest of this paper) is somewhat generalized and simplified, as a single AUA is used for all four TOEIC Bridge tests and their intended uses. A narrower approach would involve constructing an AUA for each test (e.g., the TOEIC Bridge Speaking test) and intended use (e.g., its use for selection for a specific training program). A more applied approach would involve constructing an AUA for a specific use and TOEIC Bridge test(s) involved in supporting that use. We have adopted a more generalized and simplified approach to present important claims and synthesize key evidence that would be expected to apply to different implementations of AUAs for TOEIC Bridge tests. Because the major claims (e.g., consistency of scores, meaningfulness of score interpretations) are essentially identical across the four tests, claims and warrants (elaborating statements that support claims) apply to all four tests unless specified. The evidence supporting each claim and warrant for each test is discussed separately.

TEST PERFORMANCE

Depending on a score user's needs, test takers may complete the TOEIC Bridge Listening, Reading, Speaking, and/or Writing tests. Each test requires test takers to use their English language knowledge and skills; in other words, test takers' performance is their demonstration of their knowledge and skills.

The TOEIC Bridge Listening test includes four parts, with a total of 50 multiple-choice questions. In the first part, Four Pictures, test takers hear one short phrase or sentence spoken aloud and must choose the picture that the phrase or sentence describes. 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. 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. 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. The total testing time is approximately 25 minutes.

The TOEIC Bridge Reading test has three parts, with a total of 50 multiple-choice questions. In Sentence Completion, test takers are presented with a sentence that has a missing word or phrase. Test takers must then select the word or phrase, from among four options, that best completes the sentence. 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 among four options. In Reading Comprehension Passages, test takers must read everyday texts (e.g., notices, letters, forms, advertisements) and answer two or three questions about each text. The total time allowed for the test is 35 minutes.

The TOEIC Bridge Speaking test consists of eight questions and takes approximately 15 minutes to complete. In the first two questions, test takers read aloud a short presentational text that is displayed on their screen. In the third and fourth questions, 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. In the fifth question, 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). In the sixth question, 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). In the seventh question, test takers look at four pictures that illustrate a story and narrate the story in their own words, describing places, people, actions, and feelings. In the eighth question, 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.

The TOEIC Bridge Writing test includes nine questions and lasts approximately 37 minutes. In the first three questions, test takers must drag and drop words (or phrases) to form a grammatically correct sentence. In the next three questions, test takers view a picture on their screen and use two supplied words (or phrases) to write one sentence. In the seventh question, 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 and to avoid the use of texting language. In the eighth question, test takers write a short narrative about an everyday topic (e.g., a time when you helped a friend). In the ninth question, test takers read and respond to questions in an e-mail.

In order to translate TOEIC Bridge test performance into a useful evaluation of ability, the test is scored. The test score, in effect, is a transformation of the test performance: from a demonstration of language knowledge and skills to a number. Based on measurement theory, an essentially important quality of scores is their consistency.

CLAIM 4: SCORES ARE CONSISTENT

Score consistency is important, because if it is inadequate, test scores may not be meaningful—and the test performance would only be valuable as practice. Consistency (or reliability), as a concept, suggests that scores should be minimally influenced by aspects of the test and testing procedure unrelated to language ability. There are many of these potential factors, and they should not be overlooked. With this in mind, we make the following claim about TOEIC Bridge test score consistency, supported by nine warrants:

For each TOEIC Bridge test, **scores** are *consistent* across different test tasks, different aspects of the test procedure, different raters (for Speaking and Writing), and different groups of test takers.

Consistency, Warrant 1

TOEIC Bridge tests are administered in a standard way every time they are offered.

Backing

The test is administered globally by local ETS Preferred Network (EPN) members who are required to comply with the *TOEIC*® program guidelines set forth in a policies and procedures document and test administration supplement manual. The policies and procedures document provides a mandate for test administration processes, including preadmin, test day, and postadmin activities; irregularities; emergencies; and more. It also includes a thorough overview of testing environment requirements, including lighting, noise, appropriate writing surfaces, seating arrangement, comfort, and accommodation considerations. The test administration supplement manual provides step-by-step instructions for test administrators and proctors to follow during the test. Both documents provide detailed guidance concerning test security procedures and meet ETS test integrity standards. All test administrations are subject to unannounced audits by ETS's Office of Testing Integrity. Any TOEIC test EPN that violates TOEIC test operation policies and procedures would be terminated.

Consistency, Warrant 2

Procedures for producing test scores are well-specified and are adhered to.

Backing

For the TOEIC Bridge Listening and Reading tests, all statistical analyses are conducted by ETS Psychometric Analysis and Research staff. The procedures for scoring test items and producing scaled section and total scores are elaborated in a statistical procedures document. This document specifies procedures for score key management, data file management, item analysis, differential item functioning analysis, equating, scoring, and scaling (i.e., converting raw scores to scaled scores). Tests are scored using software that undergoes a series of quality control checks to ensure that the system accurately scores all tests. The software has multiple security features programmed into it to prevent unauthorized access to any of the scoring keys or conversion tables. Hand scoring may be used as a backup verification to electronic scoring; if the hand score differs from the electronic score the cause of the discrepancy is identified by reviewing each response and comparing the response against the official answer key to confirm the correct score.

For the TOEIC Bridge Speaking and Writing tests, the procedures for scoring test items and producing scaled scores are elaborated in a scoring rules document. The TOEIC program also provides training for raters and monitors the accuracy and reliability of scoring to help ensure that raters apply the scoring rubric accurately and consistently. Speaking and writing test responses are scored centrally through the ETS Online Network for Evaluation (ONE), and each rating session is overseen by a scoring leader (see Everson & Hines, 2010).

Consistency, Warrant 3

TOEIC Bridge Speaking and Writing test raters are trained, certified, and monitored.

Backing

As described in detail by Everson and Hines (2010), raters must be college graduates with experience teaching English to learners at the high school, university, or adult levels. Prior to operational rating, raters complete a training program and pass a certification test using the ETS Online Network for Evaluation. Before each operational rating session, raters must pass a calibration test that assesses their readiness to score for that specific scoring day. ETS professional staff monitor the accuracy and quality of scoring by overseeing operational rater scoring in the ETS Online Network for Evaluation. To aid raters in scoring, ETS staff create topic notes to help raters approach the responses of certain items within a form. This extra measure is taken to help ensure that raters assign scores accurately and fairly to test takers' responses across all administrations that use a particular form.

Consistency, Warrant 4

TOEIC Bridge Speaking and Writing test raters are trained to avoid bias for or against different groups of test takers.

Backing

Raters are instructed to leave institutional or personal biases aside—such as standards for native English speaker speech or writing—while rating and interpreting the scoring guide. Raters are provided benchmarks and training samples, and scoring leaders periodically provide feedback that may include comments on the acceptability of linguistic features (e.g., pronunciation, grammatical structures, vocabulary) to reduce any impact of institutional or personal biases. In addition, the operational scoring system is designed to reduce the impact of individual raters' biases by randomly assigning raters to score test-taker responses and by having multiple raters assigned to score an individual test taker's responses (see Everson & Hines, 2010).

Consistency, Warrant 5

Raw test scores are internally consistent (internal consistency reliability).

Backing

Reliability estimates of internal consistency have been adequately high. Coefficient alpha, a measure of internal consistency, provides an indication of the consistency of test takers' responses to all items in each test. Reliability estimates of internal consistency based on the TOEIC Bridge field test were .88 and .89 for two listening test forms, .93 for each of two reading test forms, .83 and .86 for two speaking test forms, and .73 and .75 for two writing test forms (Lin et al., 2019). Overall, data from operational administrations have produced reliabilities comparable to those of the redesigned TOEIC Bridge field test.

Consistency, Warrant 6

Ratings of different TOEIC Bridge Speaking and Writing test raters are consistent (interrater reliability).

Backing

Based on the redesigned TOEIC Bridge field test, interrater agreement for the speaking and writing tests was adequately high (Lin et al., 2019). For the speaking test, the percentage of exact agreement for individual items ranged from 57% to 81% on both forms, and the percentage of exact plus adjacent ratings was greater than 99% for most speaking test items. Weighted kappa ranged from 56% to 89% on both forms. For the writing test, the percentage of exact agreement for individual items ranged from 56% to 89%, and the percentage of exact plus adjacent ratings was greater than 99% for most writing test items. Weighted kappa ranged from 76% to 92% on both forms.

Consistency, Warrant 7

Scaled test scores from different forms of the test are consistent (equivalence, or equivalent forms reliability).

Backing

For the TOEIC Bridge Listening and Reading tests, equivalence is achieved through the use of equating and a robust test development process. As the strongest form of linking between the scores on different test forms, equating compensates for small differences across forms and allows the scores from each test form to be used interchangeably, as if they had come from the same test. Several requirements need to be met for equating: The test forms must measure the same construct at the same general level of difficulty and with the same accuracy (Holland & Dorans, 2006). For the equal construct requirement, all test forms are built to the same content specifications. Item writers (both outside and inside ETS) receive detailed guidelines that supplement the test specifications. Item writers receive training in the guidelines and frequent detailed feedback. Tests are assembled using procedures that ensure a balance of points tested and difficulty in order to ensure form equivalency and reduce unintended variance. The detailed content of each test form is closely monitored and documented. For the equal difficulty requirement, test developers try their best to make the forms as similar as possible in difficulty, although it is impossible to create absolutely equivalent forms in operational work. Therefore, in order to be fair to all TOEIC test takers, equating is used to adjust test results based on the difficulty level of each new test form and derive the scaled scores from test takers' raw scores. As a result, the reported scaled scores obtained from different alternate test forms are comparable, regardless of any potential differences in form difficulty.

For the TOEIC Bridge Speaking and Writing tests, score equivalence or comparability is controlled through consistent item/test development and constant scoring rubrics. Test performance and statistics across forms are carefully monitored in every administration and over time.

Consistency, Warrant 8

Scaled scores are consistent across test administrations (stability, or test-retest reliability).

Backing

Although there is currently no direct evidence available pertaining to test-retest reliability, test performance and statistics across administrations (e.g., score means and reliability) are carefully monitored over time. Overall, scale scores across operational administrations are reasonably close with variations consistent with the ability of groups of test takers.

Consistency, Warrant 9

Scaled test scores are of comparable consistency across different groups of test takers.

Backing

The reliability and standard error of measurement are evaluated for gender groups within countries that participate in operational administrations. The standard error of measurement—as another indicator of score consistency—estimates the average variation expected in a test taker's score from one test form to another. The results have indicated that test scores are equally reliable between male and female test takers within countries. The results have also found that SEMs for the countries that participate in operational administrations are comparable.

When scores are consistent—and the evidence for this demonstrated—they can be interpreted as an indicator of knowledge, skills, or abilities. But of what ability, and is it useful to help make decisions? Scores may be consistent but be an indication of a different ability (or abilities) than expected; for example, a math test that requires strong language skills. Scores may be consistent, but consistently biased against different groups of test takers. Thus, when taking the next step to make an inference about ability based on test scores, the qualities of score interpretations are important to understand.

CLAIM 3: SCORE INTERPRETATIONS ARE MEANINGFUL, IMPARTIAL, GENERALIZABLE, AND RELEVANT

Test scores themselves are just numbers. To be useful, the numbers need to be interpreted. These interpretations about ability are essentially a transformation of the test score: a number becomes a reflection of knowledge or abilities. Fundamentally, score interpretations should be meaningful and impartial, generalize to a real-world setting, and be relevant to how they will be used. If test users have inadequate knowledge of (and confidence in) these qualities, they may end up using inaccurate or inadequate information to make decisions. Based on these principles, we state this claim about interpreting TOEIC Bridge test scores, with supporting warrants for each quality of score interpretations:

TOEIC Bridge test **scores** can be **interpreted** as indicators of English listening, reading, speaking, and writing proficiency for beginning to low-intermediate learners of English for everyday adult life. These interpretations are *meaningful* with respect to theoretically-based definitions of ability; *impartial* to all groups of test takers; *generalizable* to language use tasks typical of everyday adult life; and *relevant* to selection, placement, and proficiency-level-verification decisions for beginning to low-intermediate English proficiency. Because the TOEIC program encourages the use of multiple measures for decision-making, sufficiency is not claimed for this assessment.

Meaningfulness

Score interpretations will be more meaningful if they are based on a well-articulated definition of the construct (i.e., the targeted knowledge, skills, or abilities). The construct definition specifies the intended meaning of scores and should clearly influence the design of the test. This process of "operationalizing the construct" in test design and administration also influences the meaningfulness of scores; for example, assessments designed with similar construct definitions in mind may end up looking quite different. Consequently, warrants pertaining to the meaningfulness of score interpretations should clearly define what scores are intended to mean and state how the operationalization of the construct in test design and administration supports the intended interpretation about ability.

Meaningfulness, Warrant 1

The construct definitions reflect an interactionalist approach in which a construct is defined based on the interaction between ability and context (Bachman, 2007).

Scores are interpreted in terms of claims about test takers' knowledge, skills, and abilities. Specific claims about listening, reading, speaking, and writing proficiency are summarized below.

- TOEIC Bridge Listening comprehension: In English, test takers can understand commonly occurring spoken texts, demonstrating the ability to
  - understand simple descriptions of people, places, objects, and actions;
  - understand short dialogues or conversations on topics related to everyday life; and
  - understand short spoken monologues as they occur in everyday life when they are spoken slowly and clearly.
- TOEIC Bridge Reading comprehension: In English, test takers can understand commonly occurring written texts, demonstrating the ability to
  - understand nonlinear written texts (e.g., signs, schedules);
  - understand written instructions and directions;
  - understand short, simple correspondence; and
  - understand short informational, descriptive, and expository written texts about people, places, objects, and actions.
- TOEIC Bridge Speaking: In spoken English, test takers can perform simple communication tasks, demonstrating the ability to
  - ask for and provide basic information;
  - describe people, objects, places, and activities;
  - express a simple opinion or plan and give a reason for it;
  - give simple directions;
  - make simple requests, offers, and suggestions; and
  - narrate and sequence simple events.

- TOEIC Bridge Writing: In written English, test takers can perform simple communication tasks, demonstrating the ability to
  - ask for and provide basic information;
  - make simple requests, offers, and suggestions;
  - express thanks;
  - express a simple opinion and give a reason for it;
  - describe people, objects, places, and activities; and
  - narrate and sequence simple events.

In addition to ability, context forms the other part of the construct definition. Context is determined by the target language use domain, which is broadly defined as English for everyday adult life. The domain definition included three subdomains, including personal, public, and the workplace. The construct definition may be distinguished from related constructs (e.g., general English proficiency, communicative competence) and domains (e.g., English for academic purposes).

Backing. These construct definitions, which also represent claims about ability within an evidence-centered design (ECD) framework (Mislevy et al., 2003), were the outcome of a review of theory, research literature, and language proficiency standards with the mandate for test design in mind, as detailed in the test framework paper published by Schmidgall et al. (2019). This framework paper details the process used to produce the construct definitions, including the construction of a theory of action (logic model) to articulate the intended use of the tests and a domain analysis. The domain analysis provided a justification for defining the target language use domain of everyday adult life and the English listening, reading, speaking, and writing proficiency relevant to beginning to low-intermediate learners. The full construct definition for each of the four tests includes a broad statement about what the test intends to measure, a list of the communication goals relevant to the use of English at beginning to low-intermediate levels in the context of everyday adult life (also listed above), and an outline of the linguistic knowledge and competencies needed to achieve the communication goals (e.g., grammatical knowledge). The target language use domain of "everyday adult life" is a more general-purpose domain that emphasizes tasks and contexts that are expected to be familiar to adults and young adults. This domain definition is elaborated in the test framework paper (Schmidgall et al., 2019, pp. 5–9) and was influenced by the language use contexts defined (e.g., personal, public, and occupational contexts) and the general approach advocated by the authors of the Common European Framework of Reference (Council of Europe, 2001, 2018). As described in the framework paper, this work involved considering the types and characteristics of tasks and topics relevant to the domain.

The construct definitions were operationalized into claims about listening, reading, speaking, and writing proficiency during the test development process. As described by Everson and his colleagues, the construct definitions include communication goals that are essentially definitions of task paradigms, or the types of situations that will allow test takers to show evidence of their proficiencies (Everson et al., 2019). For each test, the test development team created a range of prototype tasks (task models) that

were explicitly connected to relevant communication goals and aspects of linguistic knowledge that the tasks were expected to engage. Task and test specifications were refined after a pilot test and finalized after a field test. The use of ECD, a systematic approach to test development, produced documentation of the process and involved data collections (e.g., cognitive labs, surveys, item performance) that helped to establish a stronger and more transparent link between the theoretically based construct definitions and claims about ability based on test scores.

Meaningfulness, Warrant 2

Item specifications clearly describe the characteristics of the tasks that test takers will perform during the test, which will elicit evidence of relevant language skills for beginning to low-intermediate English learners.

Backing. Item and test specifications were developed using an ECD approach, as described by Everson et al. (2019). The test specifications were reviewed by a team of ETS assessment specialists and by external clients.

Item writers (both outside and inside ETS) receive detailed guidelines that supplement the item specifications. External item writers receive training in the guidelines and frequent detailed feedback. For the listening and reading tests, each item is classified as to the specific ECD-based claim it supports. These claims (classifications) are reviewed as part of the regular item development process. The item review process includes a review of content for suitability for the intended population as well as fairness and editorial reviews.

Tests are assembled using procedures that help to ensure a balance of aspects of the construct tested and difficulty to maintain form equivalency and reduce unintended variance. The test forms receive a summative review, a fresh eyes review (i.e., a review by a test developer who has not previously worked on the form), and a coordinator review of each test form section. The test forms are also reviewed before administrations by ETS's partners. An ETS assessment specialist responds to all comments made by test reviewers.

Meaningfulness, Warrant 3

The procedures for administrating the TOEIC Bridge enable test takers to perform at their highest level of ability.

Backing. In order to provide the best representation of their ability, test takers need to understand test and item directions and have adequate time to engage in test activities. ETS researchers conducted cognitive interviews with test takers before the pilot and field tests and surveyed test takers after the pilot and field tests (see Everson et al., 2019). One purpose of this research was to identify whether any of the test or item directions were difficult for test takers to understand, and another purpose was to gather test-taker perceptions of the adequacy of preparation and response time for speaking and writing test tasks. Based on this research, slight modifications were made to item directions and task timing prior to the

field test. A final survey of test takers conducted after the field test found that a large majority indicated that directions were not difficult to understand, and a majority of test takers found that preparation and/or response times for speaking and writing tasks were good, with a few exceptions. Based on test-taker feedback, final adjustments were made to item directions and task timing to ensure test takers would be able to provide the best representation of their ability.

Test takers are also encouraged to provide feedback on their test-taking experience to their local test administrator or the TOEIC program directly. Test takers are provided with Candidate Comment Forms to express concerns, complaints, or questions following a test administration.

Meaningfulness, Warrant 4

The scoring procedures focus on aspects of reading, listening, speaking, and writing skills relevant to everyday adult life.

Backing. The scoring rubrics for the speaking and writing tests were developed within the context of an ECD approach to test development where scoring criteria were explicitly linked to claims about ability (Everson et al., 2019). These claims were specified based on the construct definition for each test, which derived from a review of the domain of everyday adult life, research on language proficiency, and relevant language proficiency standards (Schmidgall et al., 2019).

The listening and reading tests are scored using keys generated during the test development process. The performance of keys and distractors is monitored to minimize construct-irrelevant variance. If a problem is identified during scoring, it is corrected by ETS staff before scores are released.

Meaningfulness, Warrant 5

TOEIC Bridge Listening, Reading, Speaking, and Writing test tasks engage test takers' reading, listening, speaking, and writing skills, respectively.

Backing. For the speaking and writing tests, the findings of cognitive interviews and surveys conducted with test takers during the test development process provide support for this warrant (Everson et al., 2019). Cognitive interviews were conducted before the pilot and field tests to investigate test takers' responses, processes, and perceptions of items. Test-taker feedback was generally positive, and potential usability issues noted by test takers informed minor adjustments to item design. Test takers who completed the speaking and writing field test completed a follow-up survey in their local language, which allowed them to provide feedback on the usability of the test, item-specific perceptions, and general impressions. The results indicated that item features were functioning as test developers intended, allowing test takers to provide a demonstration of their speaking and writing skills.

This warrant may be supported indirectly through documentation of the test development process (Everson et al., 2019). For all TOEIC Bridge tests, highly qualified and trained item writers develop items that must pass multiple reviews to help ensure that items target the knowledge, skills, and abilities articulated in the construct definition and operationalized in test specifications.

Meaningfulness, Warrant 6

TOEIC Bridge Listening, Reading, Speaking, and Writing test scores can be interpreted as indicators of English language listening, reading, speaking, and writing proficiency, respectively, for beginning to low-intermediate learners of English for everyday adult life.

Backing. This warrant is supported by the documentation of the test development process described in Schmidgall et al. (2019) and Everson et al. (2019) and also through the findings of statistical analysis and research studies.

In their statistical analysis of the TOEIC Bridge field test data, Lin et al. (2019) estimated the correlations between listening, reading, speaking, and writing test scores. The four sets of test scores were moderately correlated, which suggested they were measuring something different. Sets of test scores that have a stronger theoretical relationship under a four-skills model of language proficiency had slightly higher correlations than those with a weaker theoretical relationship. For example, reading and listening test scores (receptive skills) and reading and writing test scores (which share the written channel) were more highly correlated than reading and speaking test scores (r = .78, .74, and .66, respectively).

Another strand of research evidence comes from test takers themselves in the form of self-assessments of their own language skills. Self-assessments have been shown to be useful in a variety of contexts, especially in the assessment of language skills (Powers & Powers, 2015). Language learners often have more complete access to the full spectrum of their successes and failures than do external evaluators, who have much more limited access to their behavior and thus may hold a much narrower view of their language skills (Upshur, 1975). For language skills that are not directly observable, such as listening and reading comprehension, language learners may be in a unique position to have insight into their competencies. Schmidgall (2020) conducted two research studies in which TOEIC Bridge test scores were compared to test takers' self-evaluations of their ability to complete everyday listening, reading, speaking, or writing tasks in English and found that TOEIC Bridge Listening, Reading, Speaking, and Writing test scores were moderately correlated with self-assessments (r = .55, .54, .51, and .46, respectively). These results compare favorably with the results of similar studies of the relationship between self-assessments and criterion measures; in a meta-analysis that included 67 studies, Li and Zhang (2021) found that the overall correlation between self-assessment and language performance was .466. The pattern of results in the TOEIC Bridge test self-assessment study aligned with the findings of the meta-analysis, where listening had the strongest average correlation between self-assessment and criterion measure (r = .49), followed by reading (r = .45), speaking (r = .44), and writing (r = .38). In the TOEIC Bridge test study, the trustworthiness of the self-reports as a validity criterion was supported by their high degree of internal consistency reliability (coefficient alpha = .96 to .99) and their correspondence with language tasks representing selected levels of relevant language proficiency standards.

Meaningfulness, Warrant 7

The TOEIC program communicates the meaning of test scores in terms that are clearly understandable to stakeholders.

Backing. The meaning of test scores is communicated to test takers and score users in test preparation materials, examinee handbooks, score user guides, and on ETS's and partners' websites. Examinee handbooks for the TOEIC Bridge Listening and Reading tests and for the Speaking and Writing tests are oriented toward test takers and include a summary of the construct targeted by each test, the meaning of test scores, a sample score report, sample test items, and proficiency level descriptors (ETS, 2019a, 2019c). Score user guides are available for the TOEIC Bridge Listening and Reading tests and Speaking and Writing tests, are oriented towards score users, and contain the same essential information about the meaning of test scores (ETS, 2019b, 2019d). This material is always reviewed by multiple groups (e.g., research, marketing, and business staff) prior to publication to ensure general understandability and relevance to the intended audience.

Impartiality

The design, administration, and scoring of the TOEIC Bridge tests adhere to the *ETS Standards for Quality and Fairness* (ETS, 2014), which includes the requirement that testing programs treat test takers "comparably and fairly regardless of differences in characteristics that are not relevant to the intended use" of the test (p. 19).

Impartiality, Warrant 1

The TOEIC Bridge tasks do not include response formats or content that may inappropriately favor or disfavor some test takers.

Impartiality, Warrant 2

The TOEIC Bridge tasks do not include content that may be offensive to test takers.

Backing. ETS recruits test development staff from a broad range of backgrounds to have a wider range of perspectives incorporated into the test development process. Test development staff have taught English in different countries or have had experience with English learners from various cultures. During the test development process, items are screened to help ensure they are not culturally specific and that a range of international names are represented in the content. All items receive a fairness review from assessment staff trained in the ETS Standards for Quality and Fairness (2014) to minimize the possibility of sexist, racist, or otherwise offensive test content. Every effort is made to avoid language, language usage, and cultural contexts specific to Australia, Britain, or the United States. Each test form is constructed and reviewed so that the accents of spoken material do not have marked variation across forms; both men and women are included performing a variety of roles; and stimulus material is balanced in terms of speaker gender, gender depicted in visual content, nationality, and race.

After large public administrations using new test forms, statistical analysis is conducted to confirm that items and overall test forms are functioning properly. For the listening and reading tests, this routinely includes differential item functioning analysis for gender. In addition, test takers are provided with information about how to contact ETS directly if there are concerns about the test.

Impartiality, Warrant 3

The procedures for producing a score report are clearly described in a manner understandable to all test takers.

Backing. The scoring process is briefly described in examinee handbooks in the TOEIC Bridge Listening and Reading (or Speaking and Writing) Scores section (ETS, 2019a, 2019c). The process is described in slightly more detail in the user guides in the TOEIC Bridge Listening and Reading (or Speaking and Writing) Tests Results section (ETS, 2019b, 2019d). The TOEIC Bridge Speaking and Writing examinee handbook includes the scoring rubric for each question to be completely transparent about how each question is evaluated by trained raters (ETS, 2019c).

Impartiality, Warrant 4

Test takers are treated impartially during all aspects of test administration:

- Test takers have equal access to information about TOEIC Bridge test content and procedures and have an equal opportunity to prepare.
- Test takers have equal access to the TOEIC Bridge test, in terms of cost, location, and familiarity with conditions and equipment.
- Test takers with disabilities have equal opportunity to demonstrate their language proficiency (reading, listening, speaking, writing).

Backing. Descriptions of test content are provided online and in the score user guides and examinee handbooks (ETS, 2019a, 2019b, 2019c, 2019d). All this material is easily accessible on ETS's website and through ETS's local partners. Local partners are required to provide test takers with a copy of the examinee handbook, which contains information about the characteristics of the test and item types, how personal information is protected, how to request accommodations, the intended uses of test scores, conditions under which results will be reported and to whom, how long scores will be available and usable, permitted and prohibited items, scoring information, score cancellation and hold policies, rescore policies, and sample questions.

Test administration is managed by EPN members, who are required to post registration information online. Test administrators are required to follow testing procedures contained in the TOEIC test administration policies and procedures manual and the administration supplement manual. These documents specifically address how to answer test-taker questions.

Local EPN members set test fees locally and are required to provide test takers and score users with appropriate test fee information. Local EPN members that require assistance setting test fees or fee-waiver programs may consult with TOEIC program management for guidance. EPN members are expected to make test administration information available in the local language and provide support in the local language.

The TOEIC program also offers appropriate and reasonable accommodations for test takers with disabilities. Available accommodations are elaborated in the *Guide for Test Takers With Disabilities* (ETS, 2013). ETS makes available Braille, reading only, spoken (audio), and large print versions of the TOEIC Bridge test. The scores achieved on an accommodated test are equivalent to the scores from a standard test administration and are not flagged. The test forms that are used for accommodated testing are re-equated by ETS psychometricians, when necessary, to help ensure the scores are comparable to a standard administration. When a representative receives a request for accommodated testing, the request can be reviewed by ETS to determine how to best meet the needs of the test taker; additional accommodations may be approved by the ETS Office of Disability Policy.

Generalizability

The quality of score interpretations is strengthened by evidence that language use (and its evaluation) on the test corresponds to language use (and its evaluation) in the real world.

Generalizability, Warrant 1

The characteristics of the TOEIC Bridge test tasks correspond to reading, listening, speaking, and writing language use tasks performed in everyday adult life.

Backing. Documentation from the test design process provides support for this warrant. As described in two research papers describing the conceptualization and design of the TOEIC Bridge tests (Everson et al., 2019; Schmidgall et al., 2019), test tasks were identified based on a review of theoretical and empirical literature and relevant language proficiency standards.

Relevance

Ultimately, score interpretations need to provide information about knowledge, skills, or abilities that is relevant to what a score user ideally needs to know about English skills in order to make decisions. Even if score interpretations are shown to be meaningful, impartial, and generalizable, they may provide information that is not relevant to a particular score user's needs. A warrant pertaining to relevance should elaborate the decision-making contexts to which score interpretations are suited.

Relevance, Warrant 1

TOEIC Bridge score-based interpretations provide information that is useful to make selection decisions, make placement decisions for instructional or training purposes, and verify current level of proficiency to determine readiness for more advanced study based on beginning to low-intermediate level English (reading, listening, speaking, and writing) proficiency for everyday adult life.

Backing. The test design process for the TOEIC Bridge tests began with a consideration of the types of decisions that the tests needed to support and the intended outcomes of those decisions (Schmidgall et al., 2019). This consideration was formalized in a logic model that informed the construct definition and, consequently, intended meaning of test scores. The TOEIC Bridge test user guides describe the intended uses of scores, provide guidance on appropriate test use, and advise score users to explicitly examine the suitability of the TOEIC Bridge test for their specific intended use (ETS, 2019b, 2019d). In addition, the TOEIC program's score retention policy is intended to promote appropriate use of TOEIC Bridge test scores. Specifically, scores are valid for decision-making purposes up to 2 years from the date of the test administration (for a rationale for this 2-year score retention policy, see Powers & Lall, 2013).

The extent to which test scores are consistent and score interpretations are meaningful, impartial, and generalizable essentially characterizes the "measurement quality" of the test (Schmidgall et al., 2018). If score interpretations are also relevant to the type of decisions (e.g., placement) that score users would like to make, they can have more confidence that the test will be useful. But ultimately, the usefulness of the test is still contingent on qualities of the decisions themselves.

CLAIM 2: DECISIONS TAKE LOCAL VALUES INTO ACCOUNT

Test scores are used to make decisions: This use involves another transformation of data and requires a claim about the quality of the transformed data. Typically, decision rules translate test scores into decision categories by determining the minimum test score required for each decision category. A test developer should provide guidance on the types of decisions that a test was designed to support as well as the intended qualities of those decisions (Bachman & Palmer, 2010). We state the following claim about decisions based on TOEIC Bridge test scores:

Selection decisions, placement decisions, and proficiency-level verifications that are based in part on *TOEIC Bridge* test scores take into account local educational, organizational, and/or societal *values*.

Table 1 summarizes these decision categories, the stakeholders expected to be affected by decisions, and the individuals expected to be responsible for making the decisions.

TABLE 1

Intended Uses of TOEIC Bridge Tests

Decision category	Stakeholders who will be affected by the decision	Individual(s) responsible for making the decision
Selection	Test takers, score users	Score users
Placement	Test takers, score users (including teachers)	Score users
Proficiency-level verification	Test takers, score users (including teachers)	Test takers and/or score users

Values Sensitivity

Decisions should reflect a score user's values as driven by the needs of their local decision-making or policy context. For any particular score user, the process of making decisions based on test scores should involve a purposeful consideration of the type and level of language needs associated with a decision-making category. It should also involve a consideration of the relative seriousness of false positive and false negative decision errors. For example, if a score user is very concerned about the possibility of individuals with insufficient language skills being selected, decision rules (e.g., the selection of a cut score) should reflect this concern.

Values Sensitivity, Warrant 1

Relevant regulations and proficiency requirements are considered in the decisions made by score users.

Backing. Guidance for appropriate score use is provided in the TOEIC Bridge test user guides (ETS, 2019b, 2019d). This guidance includes a list of appropriate uses and recommendations such as considering the relevance of TOEIC Bridge score interpretations to decisions and the use of multiple criteria for decision-making. If score users have any questions about appropriate test use in their context, they are encouraged to work with EPN members.

ETS provides several resources for score users who may need support in setting minimum proficiency standards for their decision-making purpose. TOEIC Bridge test scores have been mapped to CEFR levels A1, A2, and B1 to support CEFR-level classification based on test scores (see Schmidgall, 2021). The TOEIC program has also produced a guide for how to conduct a local standard-setting study to determine the minimum cut score needed for a decision-making purpose (see Tannenbaum, 2013).

CLAIM 1: CONSEQUENCES ARE BENEFICIAL

Ultimately, the purpose of assessment is to facilitate beneficial consequences. There consequences are associated with the decisions made based on test scores and consequences of merely using the test. These consequences represent another transformation of data that has desirable qualities. With this in mind, we state the following claim:

The **consequences** of using the *TOEIC Bridge* tests and of the decisions that are made based on TOEIC Bridge test scores are *beneficial* to test takers and score users.

There are two groups of primary stakeholders, or those most directly affected by the consequences of decisions based on the use of the test: test takers and score users (including teachers). The consequences of test use and of decisions are considered with these primary stakeholders in mind.

Consequences of Using the TOEIC Bridge Tests

Consequences of Use, Warrant 1

The consequences of using the TOEIC Bridge test and of the decisions that are made will be beneficial to test takers and score users.

Backing. The rationale behind the expectation that consequences of use (and of decisions) will be beneficial is elaborated in the test framework paper (see Schmidgall et al., 2019). First, the expectation that the use of the test will be beneficial for test takers and teachers is based on the use of appropriate models of language proficiency to guide test design. Second, this expectation is based on the logic model used to guide test design (Schmidgall et al., 2019, p. 3). In this logic model, decisions based on TOEIC Bridge tests are expected to produce beneficial intermediate effects and ultimate effects. Intermediate effects may include, for example, enabling score users to select individuals who have desired levels of English proficiency for vocational training institutions. Ultimate effects may include, for example, students benefitting from training that is aligned with their needs. In the future, more research is needed to fully elaborate and support the logic model as a theory of action.

Consequences of Use, Warrant 2

Score reports are treated confidentially.

Backing. TOEIC Bridge test score reports (individual or institutional) are confidential and can be released only by authorization of the individual or institution or by compulsion of legal processes. The TOEIC program recognizes test takers' rights to privacy with regard to information that is stored in data or research files held by ETS and its local EPN members and recognizes the responsibility to protect test takers from unauthorized disclosure of the information. This commitment is also stated in the examinee handbooks (ETS, 2019a, 2019c).

Consequences of Use, Warrant 3

Score reports are presented in ways that are clear and understandable to test takers and score users.

Backing. The information included on score reports is described in test preparation material, examinee handbooks, user guides, and on ETS's and its local EPN members' websites. An examinee handbook for the TOEIC Bridge Listening and Reading tests and Speaking and Writing tests is oriented toward test takers and includes a sample score report and description of its content (ETS, 2019a, 2019c). A score user guide for the TOEIC Listening and Reading tests and Speaking and Writing tests is oriented toward score users and contains the same essential information about score reports (ETS, 2019b, 2019d). This material is always reviewed by multiple groups (e.g., research, marketing, and business staff) prior to publication to help ensure general understandability and relevance to the intended audience.

Consequences of Use, Warrant 4

In language instructional settings, the TOEIC Bridge tests help promote good instructional practice and effective learning, and the use of the assessment is thus beneficial to students, instructors, and the program.

Backing. The TOEIC Bridge tests are intended to promote good instructional practice and effective learning through the use of appropriate models of language proficiency to guide test design. The definitions of listening, reading, speaking, and writing proficiency that guided test development were informed by theoretical and empirical research in second language learning and assessment and reflect

the competencies that well-established language proficiency standards associate with beginning to low-intermediate levels of proficiency (see Schmidgall et al., 2019). TOEIC Bridge test tasks are intended to reflect real-life communication and therefore preparation for the test should not be distinct from the development language skills for communication. The tests are also expected to mirror (or encourage) the types of communicative activities that effective language programs use to develop language proficiency.

The intent to have a positive impact on language teaching and learning—in particular, for language programs targeting beginning to low-intermediate learners who are learning English for everyday adult life—is explicitly stated in the logic model that guided the development of the TOEIC Bridge tests (Schmidgall et al., 2019, p. 3). The ultimate outcomes (or effects) that the tests intend to promote includes the improvement of English teaching and learning practices.

Consequences of the Decisions Made Based on TOEIC Bridge Test Scores

Consequences of Decisions, Warrant 1

The consequences of the decisions that are made will be beneficial to test takers and score users.

Backing. The TOEIC Bridge tests' user guides were designed to assist users of TOEIC Bridge test scores and include information regarding the appropriate and inappropriate use of the tests (ETS, 2019b, 2019d). In addition, the logic model articulated during test design describes the intended consequences of decisions in terms of intermediate and ultimate impacts (Schmidgall et al., 2019, p. 3). For making selection decisions, the intended outcome is that score users are able to select (recruit, admit) individuals who have the desired levels of English ability (e.g., for vocational training institutions). When making placement decisions, score users should be able to place students or employees into appropriate language training courses. When using the tests to verify current levels of English proficiency to determine readiness for more advanced study, test takers or score users should be able to use that information to target appropriate study material effectively. These intermediate effects should promote the ultimate outcomes of allowing organizations to fulfill their mission, students or employees to benefit from training aligned with their needs, and English teaching and learning practices to improve.

DISCUSSION

In this paper, we elaborated the claims and supporting evidence for the argument for redesigned TOEIC Bridge test use, including the consistency of scores; meaningfulness, impartiality, and generalizability of score interpretations; values sensitivity of decisions; and beneficence of decisions and test use. Most of the supporting evidence comes from documentation and research associated with the test design process, so additional documentation and research should be generated to help evaluate the extent to which claims are supported by operational test use. Ideally, this evidence will include working with score users to help evaluate the extent to which claims about the decisions and consequences of using TOEIC Bridge test scores are supported in practice.

An important aspect of constructing and evaluating a validity argument is to identify potential weaknesses in the evidentiary basis and to consider potential rebuttals to claims (Kane, 2006). Taking a critical view of claims can help test developers and other stakeholders avoid turning a validity argument into a checklist for best practices, presented with a confirmationist bias (Haertel, 1999). Although it is certainly desirable to be thorough and to follow best professional practices, all stakeholders—including, and especially, test developers—benefit from a clear understanding of the relative strengths and weaknesses of the validity argument for a specific assessment. With this point in mind, we highlight aspects of the evidentiary basis for existing warrants that could be strengthened and several warrants that could be introduced to further strengthen claims when evidence is produced.

There is a reasonably elaborate set of warrants supporting the claim that TOEIC Bridge test scores are consistent across different test tasks, different aspects of the test procedure, different raters, and different groups of test takers. As more operational data are produced, substantial research can be conducted to expand the evidence available pertaining to the interrater reliability of the TOEIC Bridge Speaking and Writing test scores, which is currently based on estimates of rater agreement from the field test, the equivalence of the TOEIC Bridge Speaking and Writing test scores, and stability of test scores (see the Claim 4 section: Consistency, Warrant 6; Consistency, Warrant 7; and Consistency, Warrant 8; respectively).

Interpretations about a test taker's English proficiency for everyday adult life based on TOEIC Bridge test scores are claimed to be meaningful, impartial, generalizable, and relevant to specific types of decisions about language proficiency. Warrants about the meaningfulness of scores are primarily supported by documentation and research completed as part of the test development process. The warrant that test tasks engage relevant language skills (see the Claim 3 section, Meaningfulness Warrant 5) could be further supplemented by investigating the response processes of test takers during the assessment. The evidentiary basis for the warrant that test scores are meaningful indicators of English proficiency for everyday adult life (see the Claim 3 section, Meaningfulness Warrant 6) could be expanded by examining how scores predict or relate to independent observations of test takers' proficiency. But perhaps warrants pertaining to the generalizability of score interpretations are in most need of additional evidentiary support. Future research studies could examine the extent to which test tasks approximate relevant, real-world language tasks (see the Claim 3 section, Generalizability Warrant 1), the extent to which test-based evaluations of proficiency correspond to those made in real-world settings, and other predictive validity research (for examples, see Schmidgall & Powers, 2020, 2021).

Finally, collaboration with score users is needed to expand the evidentiary basis supporting claims about decisions based on TOEIC Bridge test scores and the consequences of those decisions and of the use of the tests. For decisions, this may involve support for standard-setting activities to help score users determine the minimal level of proficiency that is needed in their specific decision-making context and gathering documentation on the types of decisions that are based on test scores and how decision-making procedures are formulated (see the Decisions section, Values Sensitivity Warrant 1). Research in this area would not only expand the evidentiary basis for warrants but potentially add new warrants

supporting the claim about decisions based on TOEIC Bridge test scores; for example, that cut scores are set to minimize the most serious classification errors. For consequences, research investigating the impact of the test on teachers and learners would help expand the evidential basis (see the Claim 1 section, Consequences of Use, Warrants 1 and 4) as well as investigations of the efficacy of decisions based on test scores (Consequences of Decisions, Warrant 1).

Another way to further support claims about decisions and consequences would be to expand the initial logic model that informed test design into a theory of action (see Schmidgall et al., 2019, pp. 2–3). The logic model specified the decisions (or "hypothesized actions") the tests were intended to support as well as their intended consequences (or "intermediate" and "ultimate effects"). A theory of action includes documentation—typically, research—that summarizes the evidence backing the causal claims made in a logic model; for example, that when using TOEIC Bridge tests for selection purposes, score users select individuals who have the desired levels of English ability.

In presenting the validity argument for the redesigned TOEIC Bridge tests, we have attempted to transparently and explicitly state claims about the tests' measurement quality and intended uses, coherently synthesize the available evidence, and identify areas for future research. As Chapelle (2012) has noted, validity arguments can have many different audiences, and our intent was to adequately elaborate the validity argument for measurement professionals while minimizing profession-specific jargon to make it accessible to all score users. The assessment use argument framework enables a simplified but coherent overview of the main claims about measurement quality and test use—the qualities of scores, score interpretations, decisions, and consequences—while supporting a more nuanced evaluation through the elaboration of specific warrants and their evidential basis.

REFERENCES

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34. **https://doi.org/10.1207/s15434311laq0201_1**

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 41–72). University of Ottawa Press. **https://doi.org/10.2307/j.ctt1ckpccf.9**

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

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students.* Nova Science.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. . . *Language Testing, 29*(1), 19–27. **https://doi.org/10.1177/0265532211417211**

Council of Europe. (2001). *The Common European Framework of Reference for Languages: Learning, teaching, assessment.* **https://rm.coe.int/16802fc1bf**

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

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327–347. **https://doi.org/10.1177/0265532208090156**

ETS. (2013). *Guide for test takers with disabilities: TOEIC test, TOEIC Bridge test, TFI test.* **https://www.ets.org/s/toeic/pdf/guide-for-test-takers-with-disabilities.pdf**

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

ETS. (2019a). *TOEIC Bridge Listening and Reading tests: Examinee handbook.* **https://www.ets.org/s/toeic/pdf/examinee_handbook_redesigned_toeic_bridge_listening_and_reading_tests.pdf**

ETS. (2019b). *TOEIC Bridge Listening and Reading tests: Score user guide.* **https://www.ets.org/s/toeic/pdf/user_guide_redesigned_toeic_bridge_listening_and_reading_tests.pdf**

ETS. (2019c). *TOEIC Bridge Speaking and Writing tests: Examinee handbook.* **https://www.ets.org/s/toeic/pdf/examinee_handbook_redesigned_toeic_bridge_speaking_and_writing_tests.pdf**

ETS. (2019d). *TOEIC Bridge Speaking and Writing tests: Score user guide.* **https://www.ets.org/s/toeic/pdf/user_guide_redesigned_toeic_bridge_speaking_and_writing_tests.pdf**

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.

Everson, P., & Hines, S. (2010). How ETS scores the TOEIC Speaking and Writing tests responses. In D. Powers (Ed.), *TOEIC compendium* (1st ed., pp. 8.1–8.9). ETS.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9. **https://doi.org/10.1111/j.1745-3992.1999.tb00276.x**

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Praeger.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Praeger.

Kane, M. T., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. Bennett & M. von Davier (Eds.), *Advancing human assessment* (pp. 489–552). Springer. **https://doi.org/10.1007/978-3-319-58689-2_16**

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing, 38*(2), 189–218. **https://doi.org/10.1177/0265532220932481**

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

Mislevy, R. J. (2012). The case for informal argument. *Measurement, 10*(1–2), 93–96. **https://doi.org/10.1080/15366367.2012.682525**

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). Routledge.

Powers, D. E., & Lall, V. (2013). *Supporting an expiration policy for English language proficiency test scores* (Research Memorandum No. RM-13-09). ETS.

Powers, D. E., & Powers, A. (2015). The incremental contribution of TOEIC Listening, Reading, Speaking, and Writing tests to predicting performance on real-life English language tasks. *Language Testing, 32*(2), 151–167. **https://doi.org/10.1177/0265532214551855**

Schmidgall, J. (2017). *Articulating and evaluating validity arguments for the TOEIC tests* (Research Report No. RR-17-51). ETS. **https://doi.org/10.1002/ets2.12182**

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. (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.

Schmidgall, J., Getman, E., & Zu, J. (2018). Screener tests need validation too: Weighing an argument for test use against practical concerns. *Language Testing, 35*(4), 583–607. **https://doi.org/10.1177/0265532217718600**

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**

Schmidgall, J., & Powers, D. E. (2020). TOEIC Writing test scores as indicators of the functional adequacy of writing in the international workplace: Evaluation by linguistic laypersons. *Assessing Writing, 46*, 1–13. **https://doi.org/10.1016/j.asw.2020.100492**

Schmidgall, J., & Powers, D. E. (2021). Predicting communicative effectiveness in the international workplace: Support for TOEIC Speaking test scores from linguistic laypersons. *Language Testing, 38*(2), 302–325. **https://doi.org/10.1177/0265532220941803**

Schmidgall, J., & Xi, X. (2020). Validation of language assessments. In C. Chapelle (Ed.), *The concise encyclopedia of applied linguistics* (pp. 1123–1158). Wiley.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238–245.

Tannenbaum, R. J. (2013). Setting standards on the TOEIC Listening and Reading test and the TOEIC Speaking and Writing tests: A recommended procedure. In D. Powers (Ed.), *The research foundation for the TOEIC tests: A compendium of studies* (2nd ed., pp. 8.0–8.12). ETS. **https://www.ets.org/Media/Research/pdf/TC2-08.pdf**

Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing 1967—1974* (pp. 53–65). TESOL.

TOEIC® Research Studies

148590-148590 • UNLWEB1221

828041

www.ets.org