Research Report

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

ETS RR–21-20

Jonathan Schmidgall
Jaime Cid
Elizabeth Carter Grissom
Lucy Li

December 2021

ETS Research Report Series

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

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

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

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, & Lucy Li

ETS, Princeton, NJ

The redesigned *TOEIC Bridge*® tests were designed to evaluate test takers' English listening, reading, speaking, and writing skills in the context of everyday adult life. In this paper, we summarize the initial validity argument that supports the use of test scores for the purpose of selection, placement, and evaluation of a test taker's English skills. The validity argument consists of four major claims that provide a coherent narrative about the measurement quality and intended uses of test scores. Each major claim in the validity argument is supported by more specific claims and a summary of supporting evidence. By considering the claims and supporting evidence presented in the validity argument, readers should be able to better evaluate whether the TOEIC Bridge tests are appropriate for their situation.

Keywords English proficiency; fairness; reliability; score users; test takers; validity

doi:10.1002/ets2.12335

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?

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

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.

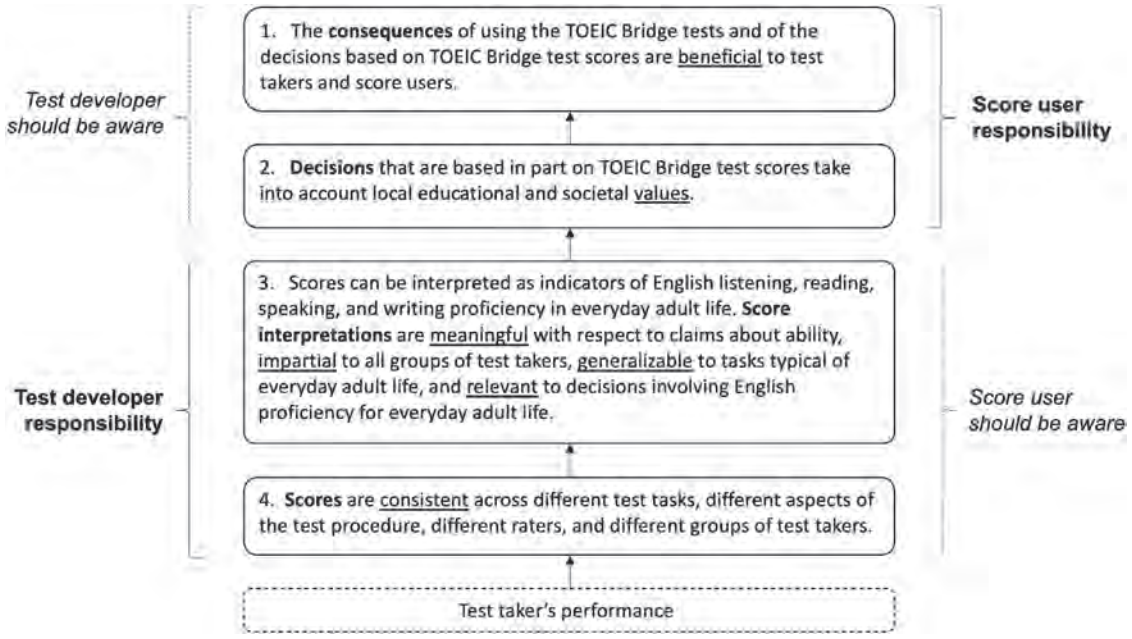

Figure 1 Assessment use argument for the redesigned TOEIC Bridge tests.

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., dialogs) 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., monologs) 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 EPN that violates TOEIC 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 dialogs or conversations on topics related to everyday life; and
    - understand short spoken monologs 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-skill 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 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 study, the trustworthiness of the self-reports as a validity criterion was supported by their high degree of internal consistency reliability (coefficient alpha = .96 – .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 toward 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 cancelation 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 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.

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.

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, 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 include 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 (e.g., 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 AUA 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). American Council on Education; 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, *2020*, 1, 27

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-Blackwell.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, *77*(3), 238–45.

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.

Suggested citation:

Schmidgall, J., Cid, J., Carter Grissom, E., & Li, L. (2021). *Making the case for the quality and use of a new language proficiency assessment: Validity argument for the redesigned* TOEIC Bridge® *Tests* (Research Report No. RR-21-20). ETS. https://doi.org/10.1002/ets2.12335

Action Editor: John Norris

Reviewers: Brent Bridgeman and Larry Davis