Predicting communicative effectiveness in the international workplace: Support for TOEIC® Speaking test scores from linguistic laypersons

Jonathan Schmidgall and Donald E. Powers
Educational Testing Service, USA

Abstract
In this study we examined the extent to which TOEIC® Speaking test scores relate to evaluations by professionals in the international workplace, the target language use domain of TOEIC tests. Linguistic laypersons in 10 countries were invited to participate in an online research survey. The survey incorporated a stratified sample of test-taker (N = 99) responses to three representative tasks from the TOEIC Speaking test (reading a text aloud, responding to questions, expressing an opinion) that were cast as workplace role-play tasks. After completing each role-play task, participants used brief, descriptive six-point rating scales to rate the communicative effectiveness (comprehensibility, task fulfillment, elaboration, and coherence) of each of several speakers. Communicative effectiveness ratings from linguistic laypersons were strongly correlated with TOEIC Speaking test scaled scores (r = 0.84). In addition, regression analysis was used to plot the relationship between layperson and test-based evaluations of speaking proficiency. Results suggested that test takers’ performances can be expected to be perceived as effective at score ranges typically associated with important decisions. The results are discussed in terms of their implications for claims about the generalizability of TOEIC Speaking test score interpretations in relation to the evaluations of linguistic laypersons in the international workplace.

Keywords
Communicative effectiveness, comprehensibility, English as a lingua franca, speaking, validation, workplace

Corresponding author:
Jonathan Schmidgall, Educational Testing Service, 660 Rosedale Road, Princeton, NJ 08541, USA.
Email: jschmidgall@ets.org
One important aspect of validation research in language assessment is investigating the extent to which test-based interpretations about language ability relate to language proficiency in the real world (McNamara, 2004). In argument-based approaches to validation, evidence about the relationship between test-based and real-world interpretations about language ability is used to support (or rebut) claims about the meaningfulness and generalizability of score interpretations (Bachman & Palmer, 2010) or extrapolation to the target domain (Kane, 2013). One way in which this evidence may be obtained is by examining the relationship between test scores and relevant variables external to the test, or criterion-related validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Identifying an appropriate and compelling criterion variable for validation research can be challenging, particularly in the context of assessing English speaking proficiency for the international workplace. Legal and privacy concerns make it infeasible to obtain criterion measures in real-world settings, and even if available, outcome measures such as job performance are typically influenced by many factors unrelated to speaking proficiency. In addition, identifying an appropriate group of informants to evaluate real-world performance may be challenging. The characteristics of language users and settings in this target language use (TLU) domain are diverse: individuals use English to communicate with interlocutors who may come from a variety of first language (L1) backgrounds across many different industries, professional settings, and geographic locations. Furthermore, interactants may not share the same L1 background, and often English is not the dominant local language outside of international workplace settings. Some researchers have argued that this TLU domain is characterized by the use of English as a lingua franca, or ELF (e.g., Kankaanranta & Louhiala-Salminen, 2010), which leads to additional implications about the nature of language use and language users in the TLU domain. Regardless, a criterion variable should reflect how language users in this domain evaluate proficiency.

In Bachman and Palmer’s (2010) assessment use argument approach to validation, the claim that score interpretations are generalizable to the TLU domain is supported by two supporting claims, or warrants. The first warrant specifies that characteristics of test tasks correspond to important characteristics of tasks in the TLU domain, and may be supported by research that provides evidence of such correspondence. The second warrant states that the criteria by which proficiency is evaluated (e.g., by trained raters using scoring rubrics) is similar to how proficiency is evaluated in the TLU domain (e.g., intuitively by real-world interlocutors). Thus, evidence to support this second warrant prioritizes the judgment of competent language users in the TLU domain who may apply their own intuitive criteria for evaluating performance (Jacoby & McNamara, 1999). Sato and McNamara (2019) described this population as “linguistic laypersons,” that is, individuals who are not trained language professionals but who nevertheless are expected to make impressionistic judgments of speakers without reference to formal theories of communication. From this perspective, linguistic laypersons are the ultimate arbiters of communicative effectiveness in the TLU domain and evidence regarding the extent to which test scores predict laypersons’ judgments may either support or undermine claims about the generalizability of score interpretations in an assessment use argument.
In this study we investigated the extent to which an assessment of English speaking proficiency for the international workplace predicts evaluations of communicative effectiveness by proficient language users in the TLU domain. The criterion measure that is elicited from these proficient language users incorporates judgments of the following three aspects of communicative effectiveness: comprehensibility, task fulfillment, and coherence. This criterion information was obtained in the context of workplace role-play tasks from linguistic laypersons currently employed in settings where inner-circle varieties of English (e.g., British, American, Australian) were not expected to be the dominant local language outside workplace settings. We begin with a brief overview of the TOEIC® Speaking test and relevant claims in its validation argument using Bachman and Palmer’s (2010) assessment use argument approach, and a discussion of the proposed criterion measure of communicative effectiveness as perceived by linguistic laypersons.

Assessing English-speaking proficiency in everyday and workplace settings: The TOEIC Speaking test

The TOEIC Speaking test is designed to measure English speaking proficiency in everyday and workplace contexts. It is computer-delivered and consists of 11 speaking tasks that are scored by trained raters. The test was developed using an evidence-centered design methodology that resulted in three claims about speaking proficiency that form the construct definition of the test (see Hines, 2010). The first claim is that test takers can produce speech that is intelligible to native and proficient non-native English speakers. The first three test tasks (two read-aloud tasks, one picture-description task) are designed to elicit evidence of a test taker’s ability with respect to this first claim. Test-taker performance is evaluated by trained raters who use scoring rubrics designed with the claim and task type (i.e., read-aloud, picture description) in mind. The second claim is that test takers can produce multi-sentence-length text to convey straightforward information, questions, instructions, narratives, and so on. The next six test tasks (three respond-to-questions tasks, three respond-to-questions-with-information-provided tasks) elicit evidence of this ability, which is evaluated by raters who use a scoring rubric designed for this claim and task type. The third claim is that the test taker can produce multi-paragraph-length text to express complex ideas, using, as appropriate, reasons, evidence, and extended explanations. The final two test tasks (propose-a-solution task, express-an-opinion task) elicit evidence of this ability, which is evaluated using a different scoring rubric for each task type.

The three claims are conceptualized as hierarchical; for example, the ability to produce multi-sentence length text to convey information (claim 2) also includes the ability to produce intelligible speech (claim 1). Consequently, all the scoring rubrics incorporate the notion of intelligibility. As the test progresses, tasks become increasingly demanding in order to elicit evidence of the test taker’s ability with respect to the higher-order claims about speaking proficiency (e.g., to produce coherent, sustained discourse that is appropriate to the communicative goal of the task). Rater scores are aggregated in a scoring procedure that assigns weights to scores for each task, and total raw scores are transformed into a scale score that ranges from 0 to 200. The scale score is reported and interpreted as evidence of the test taker’s English speaking proficiency in everyday and workplace settings.
A number of research studies have been conducted in order to facilitate the interpretation of TOEIC Speaking test scale scores and provide evidence to support the claim in the assessment use argument for the TOEIC Speaking test that scaled scores are meaningful indicators of English speaking proficiency in everyday and workplace contexts. Proficiency level descriptors were developed using a scale-anchoring approach in order to provide more detailed information about the characteristics of test takers’ performances across each of eight levels, or ranges of scores (see Educational Testing Service, 2018). Scale scores have been mapped to CEFR levels A1 to C1 (Tannenbaum & Wylie, 2013) and STANAG 6001 proficiency levels 1 and 2 (Tannenbaum & Baron, 2013), and an accessible method for local standard-setting for score users has been published (Tannenbaum, 2013). In addition, test takers’ assessments of their speaking ability have exhibited a moderately strong relationship to scores on the TOEIC Speaking test (Powers et al., 2009; Powers & Powers, 2015).

The relevance of evaluations of speaking proficiency by linguistic laypersons

Studies of oral language proficiency have alternately described this population as naive listeners (Schmidgall, 2013), evaluators (Powers et al., 1999), and ordinary listeners (Bridgeman et al., 2012). Although researchers (e.g., Elder et al., 2017) have invoked the notion of linguistic laypersons more narrowly to focus on content-area specialists who are not language specialists, for the purpose of this study we adopt the broader definition of this population as non-language experts who are actively engaged in the TLU domain. Under this broader definition, the key characteristic of linguistic laypersons is that they use the target language to communicate in the TLU domain, and are thus potential interactants in the domain. In this sense, they may not necessarily be content-area specialists in the manner described by Elder et al. (2017), but context specialists: individuals who possess a substantial familiarity with the TLU domain. As language users and interactants in the TLU domain, they are experienced in using the target language to communicate and, as a result, they may be expected to form impressionistic judgments naturally about communicative success or failure. Some linguistic laypersons work in management or other positions of authority where their judgments of communication skills have an impact on hiring, placement, or promotion.

Linguistic laypersons have characteristics that are distinct from those of trained raters. Although linguistic laypersons and trained raters may be proficient users of the target language, linguistic laypersons usually lack the formal linguistic training and experience that is typically required for rating language proficiency (Elder et al., 2017; Sato, 2014, 2018). Thus, scoring rubrics designed for trained raters may be inappropriate for linguistic laypersons since they may presume formal linguistic expertise. In addition, linguistic laypersons lack rater training, which is designed to minimize the influence of construct-irrelevant factors and biases on evaluations of language proficiency. These factors may include familiarity with the speaker’s accent, familiarity with content, attitudes towards the speaker and L2-accented speech, language proficiency and linguistic awareness, and occupational background (Reed & Cohen, 2001; Yan & Ginther, 2018). The introduction
of rater training for linguistic laypersons may help reduce the influence of construct-irrelevant factors on their evaluations of language proficiency. However, such training may undermine the external validity of the unique perspective of linguistic laypersons: if formal training in linguistics or rating has the intended effect of extinguishing or homogenizing layperson perspectives, then the valuable distinction between laypersons and trained raters may be lost.

The distinction between the populations of linguistic layperson and trained raters is not simply a distinction between novice and expert raters, as laypersons have the additional characteristic of being active observers of language use in the TLU domain. This characteristic is generally not typical of the raters who are trained to evaluate test performance. More fundamentally, linguistic laypersons are qualified to evaluate language proficiency because they are potential interlocutors or interactants in the TLU domain, regardless of their level of linguistic training or rating experience. That is, although expert raters are qualified because of their linguistic training, laypersons may be the ultimate arbiters of real-world language proficiency (Elder et al., 2017), or at the least, alternative, credible arbiters of real-world language proficiency for the purpose of evaluating generalization or extrapolation to the TLU domain for a validity argument. We suggest that formal training may make laypersons less like most language users in the TLU domain, thereby undermining their natural status as alternative, credible arbiters of real-world language proficiency for our intended purpose.

Linguistic laypersons’ lack of training, together with the possibility that training may actually reduce the generalizability of their evaluations, underscores the appropriateness of utilizing intuitive, high-inference measures to capture layperson evaluations. In efforts to capture impressionistic, non-expert judgments of L2 communication ability, researchers in applied linguistics have proposed a number of related constructs, including communicative effectiveness (Björkman, 2011; Hu, 2017; Sato, 2012), communicative ability (Sato, 2014), communicative adequacy (Pallotti, 2009; Révész et al., 2016), communicative understanding (Bridgeman et al., 2012), communicative success (Powers et al., 1999), and functional adequacy (de Jong et al., 2012; Kuiken & Vedder, 2017, 2018).

These terms and their associated constructs are often used interchangeably but tend to share several features. First, they take inspiration from Hymes’ (1972) notion of communicative competence as ability for use, which emphasizes that, in addition to knowledge of language forms and pragmatic conventions, the ability to communicate involves non-linguistic factors such as task- and domain-related competencies (Harding & McNamara, 2018). Thus, many of these measures were conceptualized and developed to complement the kinds of linguistically oriented evaluations of complexity, accuracy, and fluency (CAF) that predominate in applied linguistics research (e.g., Kuiken & Vedder, 2018; Pallotti, 2009; Révész et al., 2016). Second, these constructs are typically conceptualized within a task-based approach to language assessment (Kuiken & Vedder, 2017; Littlemore, 2003; Sato, 2018). In a task-based approach, the context of language use (e.g., setting, role of speaker and listener, communication goal) needs to be clearly specified, and evaluation criteria may be dependent on the nature of the task (Kuiken & Vedder, 2017). Finally, the notion of L2 communicative ability is typically contrasted with native-like competence (Elder et al., 2017; Harding & McNamara, 2018; Hu, 2017).
This distinction reflects the intent to emphasize communicative impact over formal accuracy as well as the understanding that English is often used as a lingua franca in settings where native speakers of English may be absent (and where, therefore, native-speaker norms may not be observed).

**Communicative effectiveness as a criterion for validation research**

The relationship between language test scores and evaluations of L2 communicative ability (hereafter *communicative effectiveness*) by linguistic laypersons has implications for claims about the meaningfulness and generalizability of language test scores. Elder et al. (2017) reported the results of three studies of the extent to which speaking test scoring was aligned with the judgments of linguistic laypersons and found discrepancies between the evaluation criteria used for test scoring and the judgments of linguistic laypersons. To the extent that these discrepancies lead to relative differences between overall evaluations of language proficiency, they may undermine the claim that test-based interpretations about speaking proficiency generalize to the TLU domain (Bachman & Palmer, 2010; Knoch & Chapelle, 2018).

The ways in which communicative effectiveness has been conceptualized have several important considerations for designing measures of this construct for the purpose of validation. The context of language use needs to be clearly specified in order for the results to have adequate external validity (i.e., generalizability to real-world language use). Although speaking test tasks are typically designed with a communicative purpose in mind, features of the context (e.g., the intended audience, the role of the speaker and listener) may be underspecified in test tasks. Providing additional context to facilitate the judgments of linguistic laypersons (such as more detailed scenarios to contextualize test performance) could improve the external validity of the evaluation. A measure of communicative effectiveness should also take into consideration the nature of the task and the aspects of communicative effectiveness that are relevant, such as comprehensibility, content, task fulfillment, and coherence and cohesion (Kuiken & Vedder, 2017).

In order to maximize the reliability and external validity of the evaluation, the design of a measure of communicative effectiveness for the purpose of validation should also take into account the relevant population of linguistic laypersons. For example, if linguistic laypersons in the TLU domain are expected to lack both formal linguistic expertise and rating experience, then the use and interpretation of the measure should not require this expertise or experience. Arguably, this would enhance the external validity of the measure as it would allow linguistic laypersons to express their unique perspectives in an intuitive fashion and capture the variability inherent in the real world where interactants are likely to have different experiences and expectations. The potential cost of such an approach is decreased inter- or intra-rater reliability of the measure. In other words, all laypersons are unlikely ultimately to base their judgments of communicative effectiveness on the same criteria, when given minimalist descriptions of the construct, and these criteria may not be used consistently by individuals. One way to minimize the impact of lower interrater reliability on scoring or evaluation is to increase the number of raters who contribute to a score. In addition, the use of a scale with multiple items for a unidimensional construct can be expected to increase the internal consistency of the
measure, providing more precise within-rater measurement. For example, Bridgeman et al. (2012) investigated the extent to which TOEFL iBT Speaking test scores predicted linguistic laypersons’ (undergraduate students’) judgments of communicative understanding. The researchers used a scale consisting of four items, with each test performance rated by a virtual team of approximately five undergraduates. The use of multiple items whose scores were highly correlated improved the internal consistency of the measure, whereas the use of virtual teams of raters minimized the impact of rater-related variance.

With these considerations in mind, we designed a measure of communicative effectiveness relevant to interpretations about English speaking ability for the workplace as operationalized in the TOEIC Speaking test. Since it would be impractical to ask laypersons to evaluate the entirety of a test taker’s performance (responses to 11 TOEIC tasks), we selected three test tasks. Each of the three selected test tasks corresponded to a different ECD-based claim about speaking proficiency. We decided that this approach would provide the best representation of a test taker’s performance for a layperson’s evaluation without compromising the practicality of survey administration.

In conceptualizing our measure of communicative effectiveness, we also considered the aspects of communicative effectiveness that were, given the task associated with a claim, relevant to each claim. For example, the read-aloud task associated with Claim 1 requires test takers to read a text they did not themselves prepare. Consequently, any evaluation of task fulfillment, coherence, or elaboration (as defined below) should also reflect the content of the text, which is something that test takers did not produce. With this in mind, we believed that the only aspect of communicative effectiveness relevant to the read-aloud task was comprehensibility. Thus, our scale items for communicative effectiveness incorporate a task-based approach wherein evaluation criteria were aligned with the purpose of the task.

Based on these considerations, we developed a criterion measure that corresponds to the hierarchical structure of the TOEIC Speaking ECD-based claims while incorporating relevant aspects of communicative effectiveness. A diagram of the relationship between TOEIC Speaking test tasks, questions, and claims, and evaluations of communicative effectiveness is shown in Table 1.

Each aspect of communicative effectiveness described in the middle column in Table 1 corresponds to a TOEIC Speaking claim, and was designed to require little or no linguistic training for linguistic laypersons to interpret. For Claim 1, the corresponding aspect of communicative effectiveness is the perceived intelligibility or comprehensibility of the speaker. Although Claim 1 uses the term “intelligibility,” we incorporated the concept of comprehensibility (Derwing & Munro, 2009) for the communicative effectiveness scale. Definitions of the terms “intelligibility” and “comprehensibility” may vary, and the scoring rubrics associated with Claim 1 tasks include references to both intelligibility and comprehensibility. A measurement of comprehensibility typically includes impressionistic questions (e.g., “How easy was it to understand the speaker?”) followed by Likert-type scales (e.g., 1 = Not easy at all, . . ., 6 = Extremely easy). For Claim 2, relevant aspects of communicative effectiveness include comprehensibility and an impressionistic measure of how successfully the speaker carried out the prescribed communicative task. In order to create a parallel item type, this measure can include a
comparable question (e.g., “How successful was the speaker?”) and Likert-type scale (e.g., 1 = Not at all successful, . . ., 6 = Extremely successful). Finally, relevant aspects of communicative effectiveness for Claim 3 also include an impressionistic measure of the perceived competence of the speaker to produce clear, organized, fluent, and comprehensive discourse.

In addition, speaking test tasks are typically designed for the speaker or test taker as the audience of the task. TOEIC Speaking test task directions and stimulus material are oriented towards the test taker (e.g., the purpose of the task is to express an opinion). By modifying task directions and including a test taker’s response to the task, we developed a role-play scenario version of each task to orient the task towards the implied listener or interactant (e.g., the purpose of the task is to listen to a colleague express an opinion on a workplace topic). Thus, test tasks were recast as role-play scenarios in which speakers (test takers) could be evaluated by listeners (linguistic laypersons) as interlocutors in an imagined workplace context.

Based on this operationalization of communicative effectiveness for linguistic laypersons, we investigated the following research question:

1. What is the relationship between TOEIC Speaking test scaled scores and evaluations of communicative effectiveness by linguistic laypersons?

In addition, owing to the study’s emphasis on the judgment of linguistic laypersons for the purpose of validation research, sampling played an important role in the design of the study. With this in mind, we intended to use a principled approach to sampling speakers
(test takers) and linguistic laypersons, describe the demographic characteristics of the samples in detail, and ask linguistic laypersons about the nature of language use in the TLU domain. Consequently, a secondary purpose of the study was to investigate the following research question:

2. What are the characteristics of linguistic laypersons and their impressions of language use in the international workplace?

Method

In this study, we administered an online survey to obtain evaluations of TOEIC Speaking test takers by linguistic laypersons. The survey included role-play scenarios in which TOEIC Speaking test-taker responses were embedded, and questions about laypersons’ use of English in the international workplace. We conducted a scale analysis using laypersons’ evaluations, calculated communicative effectiveness scores, and estimated the relationship between these scores and TOEIC Speaking test scores, using correlation and linear regression. We used descriptive statistics to summarize laypersons’ perceptions of the use of English in the international workplace.

Participants

TOEIC Speaking test takers. Speaking responses from a sample of 99 TOEIC Speaking test takers were included in the study. These data were obtained from two TOEIC Speaking test forms that were representative of operational TOEIC Speaking test administrations in terms of test-taker demographics and psychometric characteristics of scores. Test takers that did not give permission to use their data for research purposes were excluded from the sample. Prior to selecting test takers for the stratified sample, test-taker responses (audio files) were screened to ensure their audio quality was adequate for the purpose of this study. Any test takers who had received a technical difficulty (TD) rating code during the scoring process were excluded. Two researchers reviewed the remaining audio files and responses with poor audio quality or substantial background noise were excluded from the sample. These responses were removed to ensure that layperson ratings would be minimally influenced by the technical audio quality of a test taker’s response, a factor unrelated to a test taker’s speaking proficiency.

The sample was stratified by proficiency (TOEIC Speaking score level 1–2, 3, 4, 5, 6, 7, 8) and region (Asia, Europe) in order to ensure proper coverage of the range of proficiency evaluated by the TOEIC Speaking test and characteristics of the largest subpopulations of test takers. In order to minimize any interaction between region and proficiency in the sample, test takers at each of the six score levels were sampled at a ratio of 2:1 (e.g., at score level 6, 11 test takers from Asian countries were randomly sampled from those available and five test takers from Europe were sampled from those available). In total, 70 test takers from Asian countries (Korea, Japan, Taiwan, Vietnam) and 29 test takers from European countries (France, Poland, Hungary) were included. The stratified sample was mostly consistent across score levels 1–2 (n = 11), 3 (n = 14), 4 (n = 16), 5
(n = 16), 6 (n = 16), 7 (n = 14), and 8 (n = 13). There was no apparent confounding of proficiency level and geographic region of test takers in the selected sample.

**Linguistic laypersons.** A total of 100 linguistic laypersons provided valid responses to the survey. Linguistic laypersons were recruited and compensated by a third-party data collection agency (Rakuten Insight), which maintains a large pool of online panels in Asia (and through partners, worldwide) who can meet the selection criteria for a wide variety of surveys. This type of data collection agency is often used by clients for market research and eliciting attitudes and opinions from carefully defined samples.

In order to be invited to participate in the survey, individuals needed to be currently employed in the international workplace, defined as “a workplace where English is frequently used as the lingua franca for communication between native and/or non-native speakers of English.” Linguistic laypersons who reported having either no work experience in this setting or less than “general working proficiency” in English were screened out of the survey. Quotas were implemented to ensure a sample of participants across five regions, including East Asia (China = 7, Japan = 13, South Korea = 12), Europe (France = 10, Germany = 10), the Middle East (Turkey = 10), South America (Brazil = 10, Columbia = 10), South Asia (India = 10), and Southeast Asia (Singapore = 8).

Among the 100 linguistic laypersons, 63 reported having more than 3 years’ experience in the international workplace, 22 had 1 to 3 years of experience, and 15 had less than 1 year of experience. The majority of linguistic laypersons described their ability to use English in everyday and workplace situations as “general working proficiency” (59), and the remaining described their ability as “full professional working proficiency” (41). Most linguistic laypersons self-identified as non-native speakers of English (88), and a slight majority were male (55). Linguistic laypersons belonged to a variety of age groups, including under 30 (20), 30 to 39 (39), 40 to 49 (26), 50 to 59 (13), and over 60 (2). A variety of occupations were represented in the sample, including linguistic laypersons who identified as management (21), professional specialists (19), scientific/technical professionals (16), clerical/administrative (13), services (10), teaching/training (6), marketing/sales (5), technicians (4), and other (6). Over 38 different industries were represented in the sample, with the largest number of linguistic laypersons coming from industries related to the service sector (26), manufacturing sector (19), and finance (12).

Thus, the sample primarily consisted of non-native but proficient users of English with at least one year of experience in the international workplace, was relatively well-balanced by gender, and included a variety of age groups, occupations, and industries. Although English proficiency was self-reported, the survey itself was in English and used a rigorous screening procedure (described later) to ensure data integrity. Consequently, it would have been difficult for survey participants to pass the screening procedure without having a threshold level of English proficiency.

**Instruments**

**TOEIC Speaking test.** As described in the introduction, five TOEIC Speaking test questions (one read aloud, three respond to questions, and one express an opinion) were identified as relevant for the purpose of this study. Test content and test-taker responses
to these five questions were incorporated into the research survey as described below in the Online survey section. Test-taker responses to the three respond to questions items, along with the question that prompted each test-taker response, were merged to a single overall response set for the respond to questions task. Test takers were randomly assigned (with replacement) to 100 participant IDs to ensure that each test taker’s response was rated by three different linguistic laypersons. Consequently, each test taker’s sampled performance (test responses to five items, condensed to three items for the survey) was randomly assigned to (and judged by) a total of nine different linguistic laypersons. Each test taker’s TOEIC Speaking test scaled score (0 to 200) was retained for the data analysis. Given the purpose of this study (i.e., the validation of TOEIC Speaking test scaled score interpretations) scaled scores were used rather than scores for the individual items used in the study. The reliability of TOEIC Speaking test scaled scores under operational conditions has been shown to be high enough to justify using scores for high-stakes decisions; for example, Schmidgall (2017) estimated the generalizability coefficient for the operational rating design, \( E\rho^2 = 0.89 \). In G theory, the generalizability coefficient is the ratio of universe-score variance to the expected observed score variance, which is similar to the reliability coefficient in classical test theory; a reliability coefficient around 0.90 is typically desirable when scores are used for high-stakes decisions (Webb et al., 2007).

**Online survey.** The online survey, designed to be completed within 30 minutes, included three sections: a preliminary section, a role-play evaluation of workplace communication, and a background questionnaire. The preliminary section included a set of qualifying questions (provided in supplemental materials) and an informed consent form. The informed consent form also specified the expected length of the survey, and warned participants that progressing through the survey unreasonably quickly could result in being screened out. The background section included demographic questions (e.g., gender, age, occupation) and questions about the linguistic layperson’s use of English in the international workplace (provided in supplemental materials). In order to orient the role-play task as a simulation of real-world workplace communication rather than TOEIC Speaking test performance, there was no mention of the TOEIC Speaking test or test takers in the online survey.

The main section of the survey, that is, role-play evaluation of the workplace, was further divided into three parts. Each part was based on a form of spoken communication valued in the international workplace as operationalized on the TOEIC Speaking test. The first part was depicted as an announcement, which corresponded to questions 1 and 2 (read aloud) in the TOEIC Speaking test. The second part focused on responses to phone surveys, which corresponded to questions 4, 5, and 6 (respond to questions) in the TOEIC Speaking test. The third part focused on coworker opinions on workplace topics, which corresponded to question 11 (express an opinion) in the TOEIC Speaking test.

Each part began with a brief explanation of the role-play. For example, the announcement part began with the following instructions:

> In this situation, you will listen to brief announcements or messages made by local businesses or media outlets. In this case, the local context is one in which English is spoken as a second or foreign language or lingua franca.
After you listen to each announcement or message, you will be asked to give your impression of it.

When providing your impressions, please consider and evaluate each speaker independently. In other words, please try to avoid comparing one speaker to another or using comparison as the basis for evaluation.

After reading the explanation of the role-play within each section, participants then read a specific scenario for that role-play task. The three role-play scenarios included in the announcement part were as follows:

Imagine that you have a young child and would like to enroll the child in a local day-care center that has been recommended by colleagues. You call the day-care center and hear a recorded message. (Scenario A)

Imagine that you are listening to a radio program that is broadcasting local news in English. During a commercial break, you hear the following advertisement. (Scenario B)

Imagine that you are traveling and listening to a radio program that is broadcasting local news in English. The newscaster provides an update about a town in the area. (Scenario C)

All of the role-play scenarios were written by the researchers, who based them on the content and context of the related TOEIC Speaking test tasks. For the TOEIC Speaking test read aloud task, no interactional context beyond the stimulus text was provided, so plausible contexts for the announcement role-play scenarios were based on the content of the stimulus text. For the TOEIC Speaking test respond to questions task, the “market survey” context is implied but not elaborated, so we further elaborated the scenario for phone survey role-plays to make it salient for laypersons. For the TOEIC Speaking test express an opinion task, the stimulus consists of a question on a workplace topic but does not elaborate the context in which that question is asked; consequently, we created a description of a general workplace context for the opinion survey role-play part (i.e., an informal discussion with a small group of colleagues). For each scenario within the opinion survey part, we slightly modified the corresponding test stimulus material to ensure it connected to the general workplace context already described.

After reading each role-play scenario, laypersons listened to an imagined interlocutor’s response within the scenario. This response was actually a TOEIC Speaking test taker’s response to the corresponding test task. A sample script for each role-play task is provided in supplemental materials, but here is an example scenario and test taker response for the opinion survey role-play:

**Scenario:**

Your colleagues are discussing what they think is most important to consider when deciding whether to accept a job offer: the company’s location, the amount of vacation time, or the salary.

Please listen to one colleague’s opinion.
Audio of test taker’s full response to TOEIC speaking test “express an opinion” task

Sample transcript: I think the salary is the most important thing when considering whether to accept the job or not . . . and . . . the reason I think this is because . . . well, I work in order for . . . uh . . . support my family, and myself as well, and even if company’s location is . . . uh . . . close to my home, or I have a lot of vacation time, then there’s no point in . . . doing that job, unless I learn the right amount of salary, that . . . I’ll be able to support myself . . . uh . . . and to make a lot out of my job, and not just salary, but . . . I want confidence in that job, so . . . I think salary is the most important thing.

After listening to the imagined interlocutor’s response within the roleplay scenario, linguistic laypersons completed a set of evaluation questions. The announcement role-plays were evaluated based on comprehensibility (three items). The phone survey role-plays were evaluated based on comprehensibility (three items) and task fulfillment (one item). The coworker’s opinion on a workplace topic role-plays were evaluated based on comprehensibility (three items), task fulfillment (one item), coherence (one item), and elaboration (one item). Each evaluation question was rated on a six-point Likert-type scale from which linguistic laypersons selected the descriptor that best reflected their evaluation of the speaker with respect to the question. Consequently, a rating of “3” on any item always indicated “Somewhat (comprehensible, effective, coherent, etc.).” The evaluation questions are reproduced in the Appendix.

In order to familiarize participants with the role-play evaluation, all linguistic laypersons listened to and evaluated the speech of the same test taker for an example item in the first part, the announcement role-play. Laypersons then listened to and evaluated the speech of three different test takers during each part of the main section. The order in which each of the three scenarios were presented within each part was also randomized. In total, each linguistic layperson listened to and evaluated the speech of 10 test takers: the same test taker for the example, and nine different test takers for the main section. Timing checks were programmed into the first two role-play scenarios, and participants who completed them unreasonably quickly were given feedback and prompted to slow down via an onscreen pop-up window.

After completing the main section of the survey, linguistic laypersons completed background questions that included demographic questions (gender, age, occupation, industry, nationality) and questions about their language experience. The latter questions focused on linguistic laypersons’ experience with different varieties of English (described as “accents” in the survey, a more familiar concept to linguistic laypersons), languages (besides English) used with “at least limited working proficiency,” and questions about the use of English in their workplace.

Analysis

In order to ensure the integrity of data obtained from linguistic laypersons via the online survey, a rigorous screening procedure was implemented. The purpose of the screening procedure was to ensure that participants responded to survey questions in good faith. Screening was conducted independently of and prior to any data analysis
related to the research questions. Survey participants who read instructions too quickly, did not listen to the entirety of workplace role-play tasks, or completed evaluation questions too quickly were given feedback via on-screen warnings that they were moving through the survey too quickly to provide valid responses. Participants who completed the entire survey in an unreasonably short amount of time (as determined by initial usability testing) were screened out of the analysis, as were those who provided responses that were unreasonably invariant across the three parts of the survey (i.e., did not vary in their evaluations of the nine role-play scenarios they were randomly assigned from a stratified sample, an implausible outcome). In total, 202 participants responded to the survey, but 102 were screened out for failing to meet the criteria for providing valid responses.

Responses to background questions were summarized. Using R, classical item and scale analyses were conducted for evaluations of communicative effectiveness in order to ensure communicative effectiveness scores had adequate reliability. As described in the Results section, the communicative effectiveness score for each TOEIC Speaking test taker was calculated by averaging responses to individual evaluation questions across tasks. Descriptive statistics for the communicative effectiveness scale were examined to investigate the assumption of normality.

In order to address the research question, we computed the correlation between TOEIC Speaking test scaled scores and evaluations of communicative effectiveness. There were no missing data in the final validated dataset. The assumption of linearity and homoscedasticity that underlie correlational analysis and linear regression were investigated by examining the bivariate scatterplot and Q–Q plot for TOEIC Speaking test scaled scores and communicative effectiveness scores (see Agresti & Finlay, 2009). Ordinary least squares (OLS) regression was used to plot the linear relationship using TOEIC Speaking test scores to predict communicative effectiveness scores. The OLS regression line was inserted into the bivariate scatterplot obtained from the correlation analysis in order to determine the TOEIC Speaking test scaled scores corresponding to “4” on the communicative effectiveness scale, which is interpreted as being a “comprehensible, effective, coherent, and comprehensive speaker.”

Results

Characteristics of language use in the international workplace

In this section, we examine the characteristics of language use, and describe the context of communication, for linguistic laypersons in the TLU domain of the international workplace. The results of the background survey of participants who met qualification criteria for participation and provided valid responses are summarized here.

In an open-ended survey question, participants were asked to list the types of “English accents” with which they were familiar and were provided illustrative examples that included American English, British English, and Indian English. The purpose of this question was to prompt descriptions of the range of English varieties regularly encountered in the TLU domain. On average, participants listed 1.2 different English accents in response to this question (median = 1, range = 0 to 5). The most widely listed accent
was American English (69%), followed by British English (31%) and Indian English (10%). Other accents listed by participants included Korean and French (2% each), and “Asian,” Brazilian, Chinese, “European,” German, Japanese, Russian, Singlish, and Spanish (1% each).

Participants also listed the languages (besides English) that they could use “with at least limited working proficiency.” On average, participants listed 1.3 languages (median = 1, range = 0 to 5). The languages listed by participants largely corresponded to their nationalities; for example, Indian participants primarily listed Hindi and potentially another regional Indian or South Asian language (e.g., Marathi, Bengali, Konkani, Telegu, Urdu).

A set of three questions asked participants to describe the use and importance of English in their workplace. The first two questions in this set asked participants to indicate the approximate frequency of English communication in general, and the frequency of communication with non-native speakers of English in particular, on a typical day in their workplace. The results of these questions are summarized in Table 2.

As shown in Table 2, a majority of participants (55) indicated that 40% or less of their communication in the workplace on a typical day is in English, with the median response indicating that 21% to 40% of communication is in English. A majority of participants (52) also indicated that 40% or less of their communication in the workplace on a typical day involves non-native speakers of English, with the median response indicating that 21% to 40% of communication involves non-native speakers of English.

Participants were also asked to indicate how important they believed it is to have “some working proficiency” for reading, listening, speaking, and writing in English for their workplace. In general, all four skills were perceived as “Important” or “Very important” (91% for reading, 85% for listening, 84% for speaking, 78% for writing).

In summary, participants generally agreed that working proficiency in English language skills is important in their workplace, but that they varied in the extent to which they reported using English to communicate during a typical working day. Almost all participants indicated that they were multilingual (i.e., used with at least limited working proficiency a language besides English). Finally, the only English accents that a sizable number of participants indicated they were familiar with were American English (69%) and British English (31%).
Communicative effectiveness scale analysis

A scale analysis was conducted for the evaluation questions used to elicit participants’ evaluations of communicative effectiveness for each of the three role-play tasks. The purpose of the analysis was to justify combining evaluation questions into an overall communicative effectiveness score and to analyze the psychometric characteristics of the resulting measurement. There were no missing data. Across tasks, participants used the full range of responses to scale items (1 to 6), and the distribution of linguistic laypersons’ responses exhibited minimal skew and kurtosis across evaluations. Since communicative effectiveness was conceptualized as a task-based construct, scale analysis was first conducted at the task level (i.e., for each of the three workplace role-play tasks). Each evaluation question for each TOEIC Speaking test taker’s response was evaluated by three different linguistic laypersons, and these evaluations were averaged in order to reflect the aggregated evaluation for each set of linguistic laypersons by task (e.g., role-play task 1 score = [layperson 1 average score + layperson 2 average score + layperson 3 average score] / 3). This approach was also consistent with the intended interpretation of communicative effectiveness scores for the purpose of this study: perceived effectiveness to linguistic laypersons representative of the TLU domain.

For the first role-play task (announcement), all three evaluation items had high discrimination, the mean interitem correlation was 0.89, and the internal consistency of the communicative effectiveness scale was $\alpha = 0.96$. The intraclass correlation between ratings was ICC(1) = 0.46. For the second role-play task (phone survey), all four evaluation items had high discrimination, the mean interitem correlation was 0.83, and the internal consistency of the communicative effectiveness scale was $\alpha = 0.95$. The intraclass correlation between ratings was ICC(1) = 0.36. For the third role-play task (opinion on a workplace topic), all six evaluation items had high discrimination, the mean interitem correlation was 0.85, and the internal consistency of the communicative effectiveness scale was $\alpha = 0.97$. The intraclass correlation between ratings was ICC(1) = 0.45. For the overall communicative effectiveness scale, which consisted of average scores for each task, the mean interitem correlation (i.e., correlation between average task scores) was 0.57 and the internal consistency of the scale was $\alpha = 0.80$.

The relationship between TOEIC Speaking test scaled scores and communicative effectiveness

The correlation between TOEIC Speaking test scaled scores and overall communicative effectiveness scores was $r = 0.84$. This can be considered a large effect size based on Plonsky and Oswald’s (2014) recommendations for interpreting the correlation coefficient as an indicator of effect size in second language research. A bivariate scatterplot of this relationship is shown in Figure 1.

Although the primary interest of this study was the relationship between TOEIC Speaking scaled scores and overall communicative effectiveness scores, correlations between TOEIC Speaking scaled scores and communicative effectiveness scores for each of the three workplace role-play tasks were also computed. For the announcement role-play, the correlation was $r = 0.62$. For the phone survey role-play, the
correlation was $r = 0.78$. Finally, the correlation for the opinion on a workplace role-
play was $r = 0.72$.

Regression analysis was used to plot the relationship between TOEIC Speaking test
scaled scores and overall communicative effectiveness ratings. TOEIC Speaking test
scaled scores were significant predictors of communicative effectiveness ratings, $t(98) =
23.42$, $p < .001$. The OLS regression line conveying the relationship between TOEIC
Speaking test scaled scores and communicative effectiveness ratings is shown in the
bivariate scatterplot in Figure 1.

As shown in Figure 1, as TOEIC Speaking test scaled scores increased, layperson
evaluations of communicative effectiveness increased. The regression line shows how
TOEIC Speaking test scaled scores related to evaluations of communicative effective-
ness. For example, TOEIC Speaking test takers who were evaluated as comprehensible,
effective, coherent, and producing sustained and fluent discourse (i.e., an overall com-
municative effectiveness score of 4) received a TOEIC Speaking test scaled score of
approximately 160, on average. Conversely, test takers who were evaluated as “some-
what” comprehensible, effective, coherent, and producing sustained and fluent discourse
(i.e., an overall communicative effectiveness score of 3) received a TOEIC Speaking test
scaled score of approximately 100, on average. Typically, more important decisions
based on TOEIC Speaking scores are made within this range; for example, the cut score
for CEFR level B2 for the TOEIC Speaking test is 160. For the range of communicative
effectiveness scores 3.6 to 4.4 (i.e., closer to “effective” (4) than “somewhat effective”
(3) or “very effective”), the corresponding range of TOEIC Speaking scaled scores is
approximately 130 to 190.
Discussion

This study focused on the possible role of linguistic laypersons in the development of a validity argument to support the interpretation and use of scores from the TOEIC Speaking test. Validity arguments are multi-faceted, but at their heart they are accumulations of various kinds of evidence from a variety of sources to support the interpretation and use of test scores (Bachman & Palmer, 2010). Some evidence, like laypersons’ judgments, may be useful for supporting inferences about the extent to which interpretations about proficiency based on test scores generalize (Bachman & Palmer, 2010) or extrapolate (Kane, 2013) to the TLU domain. Other kinds of evidence, such as the evaluations provided by trained raters, are more central to claims about the consistency and meaning of score interpretations. The focus of our paper is on the former, and not on the critical role of the judgment provided by trained raters as the basis for score reporting. For the TOEIC Speaking test’s validity argument, evidence from the latter has been discussed in considerable detail elsewhere (e.g., Hines, 2010; Schmidgall, 2017). Evidence obtained from linguistic laypersons and from trained raters relates to different claims in the validity argument, and serves functionally different purposes. One does not supplant the other.

In this study we found that TOEIC Speaking test scores were strong predictors of evaluations of communicative effectiveness by linguistic laypersons in the international workplace ($r = 0.84$). These results provide evidence that the manner in which speaking proficiency is evaluated based on TOEIC Speaking test performances generalizes well to how communicative effectiveness is evaluated by linguistic laypersons based on analogous TLU domain tasks. In addition, the results of regression analysis provide an additional way in which to interpret TOEIC Speaking test scaled scores, that is, as predictors of evaluations of communicative effectiveness by linguistic laypersons. We believe that the way in which we have designed (and analyzed results from) a communicative effectiveness scale can facilitate interpretation of TOEIC Speaking test scores in relation to communicative effectiveness; for example, a score of 4 implies that, on average, linguistic laypersons perceive the speaker to be comprehensible, coherent, and to fulfill task demands. As a task-based construct, this perception should be interpreted with respect to the types of communicative tasks included in the evaluation: giving announcements, responding to questions, and expressing opinions.

Although the relationship between language test scores and evaluations by linguistic laypersons has been investigated in several previous studies (e.g., Bridgeman et al., 2012; Sato, 2018), this study is distinguished, we believe, by a more purposefully implemented task-based approach to evaluation by presenting actual test performances to linguistic laypersons in real-world role-play scenarios. In doing so, we hoped to enhance the external validity of such evaluations.

In this study, we also documented the characteristics of study participants (a sample of linguistic laypersons) in order to investigate some common assumptions about this population. One assumption is that the TLU domain of the international workplace is characterized by the use of English as a lingua franca (ELF), which implies that a substantial portion of English communication involves L2 English speakers (e.g., Jenkins & Leung, 2014; Kankaanranta & Louhiala-Salminen, 2010; Park & Wee, 2012, chapter 4). The qualifying criteria for participating in this study included (a) being a user of English in the international workplace and (b) being based in a geographical location where
English has not been the historically dominant local language (from a World Englishes perspective, the outer and expanding circle). Most (88%) of participants self-identified as non-native speakers of English. Surprisingly, a majority of participants reported that, on a typical workday, less than 40% of their communication involves non-native speakers of English (in English or another language). Another potential implication of the ELF view is the expectation of encountering a range of varieties of English in the TLU domain. However, when asked to list the English accents that were familiar to them, only two varieties were frequently mentioned (i.e., American and British English). These findings may be the result of certain inadequacies in our questioning: the former question about the frequency of communication with non-native speakers of English may have been unclear, and responses to the latter question may have been influenced by examples of accents (American, British, Indian) included in the question itself. Alternatively, it is possible that participants understood the questions as intended, but that, despite their occurrence, these phenomena are simply not widely or easily noticed by many language users in the workplace. Nonetheless, we believe that these assumptions (particularly when making generalizations about aspects of language use and interaction in the international workplace) may deserve further study.

This study focused on aspects of Bachman and Palmer’s (2010) notions of the meaningfulness and generalizability of score interpretations, and specifically on the following: (a) the extent to which scores on a speaking test can be interpreted as indicators of speaking proficiency (meaningfulness); and (b) whether test-based evaluations of speaking proficiency based on scoring criteria and rubrics correspond to the evaluations of language users in the TLU domain (generalizability). Some researchers have questioned the extent to which the pool of trained raters for large-scale assessments such as the TOEIC (which are often dominated by L1 English speakers) are equipped to evaluate “the diverse ways in which English is used internationally” (Jenkins & Leung, 2014, p. 1609). In gathering the evaluations of a diverse international sample of laypersons (mostly L2 English users) from the TLU domain, we found that TOEIC Speaking test scores were strong predictors of the aggregated judgment of laypersons.

Another important aspect of generalizability is the extent to which the characteristics of assessment tasks correspond to TLU domain tasks, which this study does not address. In their review of the TOEIC tests, Im and Cheng (2019) argued that some aspects of TLU domain tasks (e.g., interaction, which is critical to the notion of ELF) may not be adequately represented in TOEIC test tasks. Thus, additional research is needed in order to gain a better understanding of the extent to which features of TLU domain tasks are underrepresented in test tasks. Such research would have, we think, implications for claims about the generalizability (and ultimately, the meaningfulness) of TOEIC Speaking test score interpretations.

The use of layperson perspectives in language testing research has promising applications, we believe, but it involves practical challenges that should be carefully considered. A key feature (and major strength) of our study, we believe, was the consideration of perspectives of laypersons as a validation criterion. This approach was not intended to diminish the importance of trained raters, or to suggest that the judgments of laypersons can substitute for those of trained raters. We postulated that a lay perspective might be different from, more diverse than, and possibly more authentic than the more homogeneous perspectives
of trained raters. We acknowledge, however, that our relatively small sample allowed only an analysis of mean ratings, and not individual differences among linguistic laypersons. Thus, our analysis may have masked interesting and relevant differences among laypersons having presumably heterogeneous perspectives, thereby failing to fully understand the unique value laypersons’ judgments may hold as a validation criterion. In a sense, our method of analysis may have homogenized laypersons’ evaluations in much the same way as training does for professional raters.

Another challenging aspect of using layperson ratings is designing scales that are sufficiently defined yet easy to use without being overly prescriptive. Although the types of items used in our scales are commonly used in research investigating the comprehensibility of speech (e.g., Bridgeman et al., 2012; Derwing & Munro, 2009; Schmidgall, 2013), as one reviewer noted, the scale item for one relevant aspect of communicative effectiveness, “elaboration,” seemed incongruent with its intended meaning as defined by the relevant ECD claim. Specifically, the item in the communicative effectiveness scale appeared to emphasize fluency more than elaboration, which was the intended meaning. We note this as a potential weakness of the scale, as layperson perceptions of fluency are arguably more relevant to ECD Claim 1 than Claim 3, to which the scale item was intended to relate. In attempting to ensure that rating criteria were easily understood by layperson raters, in some instances we have deviated slightly from the precise wording of the ECD claims made by the test.

The use of online panels may facilitate efficient targeting of large samples of laypersons for survey research, but it involves a number of challenges as well. Companies such as Rakuten Insight (used in our study) and Qualtrics develop and maintain large panels that are intended to facilitate the selection of representative samples, and they are often used in market research. By utilizing these services, researchers may be able to more efficiently target and survey language users in particular TLU domains. These companies have their own quality standards, but they are primarily focused on verifying panelists’ identities and the quality of the background information they provide. Thus, researchers will need to impose data quality checks of their own, through either pre-analysis data validation or screening. In our study, more than 50% of the initial completed responses were screened out during the pre-analysis data validation phase. This is not necessarily unusual for this type of online survey research. For example, Im (2019, p. 77) used Qualtrics to conduct a survey of employers in South Korea and screened out 272 of 553 responses (49%) during the pre-analysis data validation phase. In a similarly designed validation study for the TOEIC Writing test, we also used Rakuten Insight to survey laypersons and 52 of 152 (34%) participants’ data were screened out prior to analysis for not meeting validation criteria (Schmidgall & Powers, 2020).

Research has established that impressionistic judgments of communicative effectiveness by linguistic laypersons may be influenced by various factors including familiarity with the speaker’s accent and attitudes towards the speaker (e.g., Schmidgall, 2013). As one reviewer noted, by gathering additional information about the attitudes of our study participants, we could have investigated rater bias. Although this might have made an interesting and important contribution, we were unable to gather these data owing to practical limitations (i.e., survey length and a tradeoff that would require reducing the number of participants evaluating each test-taker response) and concern about burdening our study participants and thereby decreasing their willingness to furnish other
information that we deemed more critical to our efforts. As previously noted, we attempted to control for these factors by averaging the ratings of a set of randomly assigned linguistic laypersons, so that each speaker’s evaluation of communicative effectiveness was based on the collective judgment of a diverse group of real-world interlocutors. Given this design, we believe that it is unlikely that individual or subgroup biases had a substantial influence on the results unless they exist systematically at the population level. Regardless, our goal was to obtain impressionistic evaluations by laypersons as they may exist in the TLU domain, which may include bias. We do not argue that layperson judgment should necessarily be the gold standard for this approach to criterion-related validation, but offer it as another approach that some may find informative.

Finally, the results of this study showed that TOEIC Speaking test scores were highly correlated with pooled evaluations of communicative effectiveness. However, the correlation with individual judgments would be expected to be much lower owing to variation between individuals. As the communicative effective scale analysis showed, interrater agreement was generally low, whereas internal consistency was high (although possibly inflated by the halo effect). Individual variation in communicative effectiveness judgments might be reduced through more training and the use of less impressionistic rating scales. Although training may help reduce the impact of biases and personal experiences, the purpose of this study was to capture the intuitive judgments of linguistic laypersons as they are made in the TLU domain. Ultimately, the more training that linguistic laypersons receive to minimize the impact of biases, the less unique their perspectives may become, and consequently, the less representative they may be of language users in the TLU domain.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: The research reported in this paper was funded by the TOEIC program at Educational Testing Service. Any opinions expressed in this paper are those of the authors and not necessarily of Educational Testing Service.

ORCID iD
Jonathan Schmidgall https://orcid.org/0000-0003-2560-0646

Supplemental material
Supplemental material for this article is available online.

References
Agresti, A., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Pearson.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. American Educational Research Association.
Bachman, L. F., & Palmer, A. S. (2010). Language assessment in practice. Oxford University Press.

Björkman, B. (2011). Pragmatic strategies in English as an academic lingua franca: Ways of achieving communicative effectiveness? Journal of Pragmatics, 43, 950–964. https://doi.org/10.1016/j.pragma.2010.07.033

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2012). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. Language Testing, 29(1), 91–108.

de Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. Studies in Second Language Acquisition, 34, 5–34. https://doi.org/10.1017/S0272263111000489

Derwing, T. M., & Munro, M. J. (2009). Comprehensibility as a factor in listener interaction preferences: Implications for the workplace. The Canadian Modern Language Review, 66(2), 181–202.

Educational Testing Service (2018). User guide: Speaking and writing. Author.

Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. Language and Communication, 57, 14–21. https://doi.org/10.1016/j.langcom.2016.12.005

Harding, L., & McNamara, T. (2018). Language assessment: The challenge of ELF. In J. Jenkins, W. Baker & M. Dewey (Eds.), The Routledge handbook of English as a lingua franca (pp. 570–582). Routledge.

Hines, S. (2010). Evidence-centered design: The TOEIC Speaking and Writing tests. In D. Powers (Ed.), TOEIC compendium (1st ed., pp. 7.1–7.31). Educational Testing Service.

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.

Hymes, D. (1972). On communicative competence. In J. Pride & J. Holmes (Eds.), Sociolinguistics. Penguin.

Im, G.-H. (2019). Stakeholder voices: Validity argument for score meaning and use of the Test of English for International Communication (TOEIC) in international business workplaces in South Korea [Unpublished doctoral dissertation]. Queen’s University, Kingston, Ontario.

Im, G.-H., & Cheng, L. (2019). The Test of English for International Communication (TOEIC). Language Testing, 36, 315–324. https://doi.org/10.1177/0265532219828252

Jacoby, S., & McNamara, T. (1999). Locating competence. English for Specific Purposes, 18(3), 213–241. https://doi.org/10.1016/S0889-4906(97)00053-7

Jenkins, J., & Leung, C. (2014). English as a lingua franca. In A. Kunnan (Ed.), The companion to language assessment (pp. 1607–1616). Wiley.

Kane, M. T. (2013). Validating the interpretation and use of test scores. Journal of Educational Measurement, 50, 1–73. https://doi.org/10.1111/jedm.12000

Kankaanranta, A., & Louhiala-Salminen, L. (2010). “English? – Oh, it’s just work!”: A study of ELF users’ perceptions. English for Specific Purposes, 29, 204–209. https://doi.org/10.1016/j.esp.2009.06.004

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. Language Testing, 35, 477–499. https://doi.org/10.1177/0265532217710049

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. Language Testing, 34, 321–336. https://doi.org/10.1177/0265532216663991

Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi & Y. Kim (Eds.), Task-based approaches to teaching and assessing pragmatics (pp. 265–285). Benjamins.

Littlemore, J. (2003). The communicative effectiveness of different types of communication strategy. System, 31, 331–347. http://dx.doi.org/10.1016/S0346-251X(03)00046-0
McNamara, T. (2004). Language testing. In A. Davies & C. Elder (Eds.), The handbook of applied linguistics (pp. 763–783). Blackwell.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. Applied Linguistics, 30, 590–601. https://doi.org/10.1093/applin/amp045

Park, J. S.-Y., & Wee, L. (2012). Markets of English: Linguistic capital and language policy in a globalizing world. Routledge.

Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. Language Learning, 64(4), 878–912.

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

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, 129–149. https://doi.org/10.1177/0265532214551855

Powers, D. E., Schedl, M. A., Leung, S. W., & Butler, F. A. (1999). Validating the revised Test of Spoken English against a criterion of communicative success. Language Testing, 16, 399–425. https://doi.org/10.1177/026553229901600401

Reed, D. J., & Cohen, A. D. (2001). Revisiting raters and ratings in oral language assessment. In C. Elder, A. Brown, E. Grove, N. Iwashita, T. Lumley, T. McNamara & K. O’Loughlin (Eds.), Experimenting with uncertainty: Essays in honour of Alan Davies (pp. 82–96). Cambridge University Press.

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. Applied Linguistics, 37, 828–848. https://doi.org/10.1093/applin/amu069

Sato, T. (2012). The contribution of test-takers’ speech content to scores on an English oral proficiency test. Language Testing, 29, 223–241. https://doi.org/10.1177/0265532211421162

Sato, T. (2014). Linguistic laypersons’ perspective on second language oral communication ability [Unpublished doctoral dissertation]. University of Melbourne, Australia.

Sato, T. (2018). The gap between communicative ability measurements: General-purpose English speaking tests and linguistic laypersons’ judgments. Papers in Language Testing and Assessment, 7(1), 1–31.

Sato, T., & McNamara, T. (2019). What counts in second language oral communication ability? The perspective of linguistic laypersons. Applied Linguistics, 40(6), 814–916. https://doi.org/10.1093/applin/amy032

Schmidgall, J. (2013). Modeling speaker proficiency, comprehensibility, and perceived competence in a language use domain [Unpublished doctoral dissertation]. University of California, Los Angeles, CA.

Schmidgall, J. (2017). The consistency of TOEIC Speaking scores across ratings and tasks (Research Report No. RR-17-46). Educational Testing Service. https://doi.org/10.1002/ets2.12178

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, Article 100492. https://doi.org/10.1016/j.asw.2020.100492

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.), TOEIC compendium (2nd ed., pp. 8.0–8.12). Educational Testing Service.

Tannenbaum, R. J., & Baron, P. A. (2013). Mapping TOEIC test scores to the STANAG 6001 language proficiency levels. In D. Powers (Ed.), TOEIC compendium (2nd ed., pp. 7.0–7.17). Educational Testing Service.
Tannenbaum, R. J., & Wylie, E. C. (2013). Mapping TOEIC and TOEIC Bridge® test scores to the Common European Framework of Reference. In D. Powers (Ed.), TOEIC compendium (2nd ed., pp. 6.1–6.10). Educational Testing Service.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability coefficients and generalizability theory. Handbook of Statistics, 26, 81–124.

Yan, X., & Ginther, A. (2018). Listeners and raters: Similarities and differences in evaluation of accented speech. In O. Kang & A. Ginther (Eds.), Assessment in second language pronunciation (pp. 67–88). Routledge.

Appendix

Evaluation questions for communicative effectiveness

Comprehensibility (Parts 1–3)

- How certain are you that you understood the speaker?
  - Not certain at all / Not very certain / Somewhat certain / Certain / Very certain / Extremely certain
- How easy was it to understand the speaker?
  - Not easy at all / Not very easy / Somewhat easy / Easy / Very easy / Extremely easy
- How comprehensible was the speaker?
  - Not comprehensible at all / Not very comprehensible / Somewhat comprehensible / Comprehensible / Very comprehensible / Extremely comprehensible

Task fulfillment (Parts 2–3)

- (Part 2) Overall, how effectively did the speaker answer the questions?
- (Part 3) Overall, how effectively did the speaker present and support his or her opinion on the topic?
  - Not effectively at all / Not very effectively / Somewhat effectively / Effectively / Very effectively / Extremely effectively

Coherence (Part 3)

- How coherent was the speaker’s response?
  - Not coherent at all / Scarcely coherent / Somewhat coherent / Coherent / Very coherent / Extremely coherent

Elaboration (Part 3)

- How sustained (fluent, not halting) was the speaker’s response?
  - Not sustained at all / Scarcely sustained / Somewhat sustained / Sustained / Very sustained / Extremely sustained