ETS

Listening. Learning. Leading.®

Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments

John W. Young,
Youngsoon So & Gary J. Ockey
Educational Testing Service

Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments

John W. Young, Youngsoon So, & Gary J. Ockey

Educational Testing Service

Table of Contents

Introduction 3
Definitions of Key Terms 4
Planning and Developing an Assessment 6
Using Selected-Response Questions 7
Scoring Constructed-Response Test Items 9
Statistical Analyses of Test Results 13
Validity Research 16
Providing Guidance to Stakeholders 17
Giving a Voice to Stakeholders in the Testing Process 19
Summary 19
Bibliography 20

Introduction

Educational Testing Service (ETS) is committed to ensuring that our assessments and other products are of the highest technical quality and as free from bias as possible. To meet this commitment, all ETS assessments and products undergo rigorous formal reviews to ensure that they adhere to the ETS fairness guidelines, which are set forth in a series of six publications to date (ETS, 2002, 2003, 2005, 2007, 2009a, 2009b). These publications document the standards and best practices for quality and fairness that ETS strives to adhere to in the development of all of our assessments and products.

This publication, *Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments*, adds to the ETS series on fairness, and focuses on the recommended best practices for the development of English language proficiency assessments taken by international test-taker populations. Assessing English learners requires attention to certain challenges not encountered in most other assessment contexts. For instance, the language of the assessment items and instructions—English—is also the ability that the test aims to measure. The diversity of the global English learner population in terms of language learning backgrounds, purposes and motivations for learning, and cultural background, among other factors, represents an additional challenge to test developers. This publication recognizes these and other issues related to assessing international English learners and proposes guidelines for test development to ensure validity and fairness in the assessment process. *Guidelines for Best Test Development Practices to Ensure Validity and Fairness for International English Language Proficiency Assessments* highlights issues relevant to the assessment of English in an international setting. This publication complements two existing ETS publications, *ETS International Principles for Fairness Review of Assessments* (ETS, 2007), which focuses primarily on general fairness concerns and the importance of considering local religious, cultural, and political values in the development of assessments used with international test-takers, and *Guidelines for the Assessment of English Language Learners* (ETS, 2009b), which spotlights assessments for K–12 English learners in the United States. The *ETS International Principles for Fairness Review of Assessments* (ETS, 2009a) focuses on general principles of fairness in an international context and how these can be balanced with assessment principles. Readers interested in assessing English learners in international settings may find all three of these complementary publications to be valuable sources of information.

In developing these guidelines, the authors reviewed a number of existing professional standards documents in educational assessment and language testing, including the AERA/APA/NCME Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999); the International Test Commission's International Guidelines for Test Use (ITC, 2000); the European Association for Language Testing and Assessment's Guidelines for Good Practice in Language Testing and Assessment (EALTA, 2006); the Association of Language Testers in Europe's the ALTE Code of Practice (ALTE, 2001); the Japan Language Testing Association's Code of Good Testing Practice (JLTA, n.d.); and the International Language Testing Association's Guidelines for Practice (ILTA, 2007). In addition, the authors consulted with internal and external experts in language assessment while developing the guidelines contained in this publication. This publication is intended to be widely distributed and accessible to all stakeholders and interested parties. It can be found on the ETS website: www.ets.org.

The use of an assessment affects different groups of stakeholders in different ways. For issues of validity and fairness, it is likely that different groups of stakeholders have different concerns, and consequently different expectations. This publication is primarily intended to serve the needs of educational agencies and organizations involved in the development, administration, and scoring of international English language proficiency assessments. Others, such as individuals and groups using international English language proficiency assessments for admissions and selection, or for diagnostic feedback in instructional programs, and international English teachers and students may also find the publication useful.

The guidelines are organized as follows: We begin with definitions of key terms related to assessment validity and fairness. We then discuss critical stages in the planning and development of an assessment of English proficiency for individuals who have learned English in a foreign-language context. Next, we address more technical concerns in the assessment of English proficiency, including issues related to the development and scoring of selected- and constructed-response test items, analyzing score results, and conducting validity research. This discussion is followed by a section which provides guidance for assuring that stakeholder groups are informed of an assessment practice and are given opportunities to provide feedback into the test development process.

Definitions of Key Terms

The following key terms are used throughout this publication:

- **Bias** in assessment refers to the presence of systematic differences in the meaning of test scores associated with group membership. Tests which are biased are not fair to one or more groups of test-takers. For instance, a reading assessment which uses a passage about a cultural event in a certain part of the world may be biased in favor of test-takers from that country or region. An example would be a passage about Halloween, which might favor test-takers from western countries which celebrate the holiday and disadvantage test-takers from areas where Halloween is not celebrated or not well known.
- A **construct** is an ability or skill that an assessment aims to measure. Examples of common assessment constructs include academic English language proficiency, mathematics knowledge, and writing ability. The construct definition of an assessment becomes the basis for the score interpretations and inferences that will be made by stakeholders. A number of considerations (e.g., age of the target population, context of target-language use, the specific language register that is relevant to the assessment purpose, the decisions that assessment scores are intended to inform) should collectively be taken into account in defining a construct for a particular assessment. For example, a construct for an English listening test might be phrased as follows: "The test measures the degree to which students have the English listening skills required for English-medium middle-school contexts."
- **Construct-irrelevant variance** is an effect on differences in test scores that is *not* attributable to the construct that the test is designed to measure. An example of construct-irrelevant variance would be a speaking test that requires a test-taker to read a graph and then describe what the graph shows. If reading the graph requires background knowledge or cognitive abilities that are not

available to all individuals in the target population, score differences observed among test-takers could be due to differences in their ability to read a complex graph in addition to differences in their speaking proficiency—the target construct. The graph-reading ability is irrelevant to measuring the target construct and would be the cause of construct-irrelevant variance. When construct irrelevant variance is present, it can reduce the validity of score interpretations.

- **Reliability** refers to the extent to which an assessment yields the same results on different occasions. Ideally, if an assessment is given to two groups of test-takers with equal ability under the same testing conditions, the results of the two assessments should be the same, or very similar. Different types of reliability are of interest depending on which specific source of inconsistency is believed to threaten score reliability. For example, **inter-rater reliability** demonstrates the degree of agreement among raters. Inter-rater reliability is typically reported when subjectivity is involved in scoring test-taker responses, such as in scoring constructed-response items. **Internal consistency** is another type of reliability that is commonly reported in many large-scale assessments. It refers to the degree to which a set of items measures a single construct, as they were originally designed to. Cronbach's alpha is the most commonly used indicator of internal consistency.
- **Constructed-response and selected-response items** are two broad categories of test items. The distinction between the two categories refers to the type of response expected from the test-takers. **A response** is the answer that a test-taker gives to a test question. A **constructed-response item** requires a test-taker to produce a spoken or written response rather than selecting an answer choice that has been provided. An example would be to write a short essay on a given topic. A **selected-response item** provides answer choices from which the test-taker must choose the correct answer(s). True-false items, multiple-choice questions, and matching items are examples of selected-response items. Multiple-choice questions, the most frequently used item type, consist of two parts: (i) a *stem* that provides a question to be answered and (ii) *response options* that contain one correct answer and several incorrect options. The incorrect options in a multiple-choice question are called *distracters*.
- **Stakeholders** are any individuals or groups that are impacted by the use or the effects of a testing process. Examples of stakeholders in an academic context are test-takers, teachers, test-taker families, schools, and selection committees.
- **Validity** refers to the degree to which assessment scores can be interpreted as a meaningful indicator of the construct of interest. A valid interpretation of assessment results is possible when the target construct is the dominant factor affecting a test-taker's performance on an assessment. There are several different ways to investigate validity, depending on the score interpretations and inferences that an assessment seeks to support. First, **content validity** refers to the extent to which questions and tasks in an assessment represent all important aspects of the target construct. Second, **construct validity** refers to the extent to which inferences can be made about the target construct based on test performance. Third, **concurrent validity** refers to the relationship between test scores from an assessment and an independent criterion that is believed to assess the same construct. Finally, **predictive validity** refers to the extent to which the performance on an assessment can predict a test-taker's future performance on an outcome of interest.

Planning and Developing an Assessment

In the planning and development of an English language proficiency assessment to be administered to international test-takers, the same general principles of good assessment practices used with other types of assessments apply. Most importantly, the purposes for an assessment must be clearly specified in order for valid interpretations to be made on the basis of the scores from the assessment. An assessment may be appropriate for one purpose but inappropriate for another purpose. For example, an international English assessment that focuses on the uses of English in an academic setting would not necessarily be useful for other purposes, such as screening job candidates in the workplace. In the same way, an assessment may be considered appropriate for one group of test-takers but not necessarily for another. For example, an international assessment of English proficiency intended for use with students preparing to study in an academic environment in a country, such as the United States, where English is the primary language, would almost certainly not be appropriate for assessing the English language abilities of individuals interested in using English for communicating with English speakers recreationally via social media or while travelling. An assessment for the first (study abroad) group would require more formal and academic English than one designed for the second (recreational) group.

It is also essential to develop a precise and explicit definition of the construct the assessment is intended to measure. The underlying theoretical rationale for the existence of the construct should be articulated. An assessment that is built on a strong theoretical foundation is one that is more likely to lead to valid interpretations of test scores. In addition, a clear definition of the construct being measured can help clarify the skills associated with that construct. This enables test developers to create tasks for an assessment that will best engage the test-taker's skills and reflect the construct of interest.

In developing an English language assessment, assessment specifications can be used to define the specific language knowledge, skills, and/or abilities that the test aims to measure. Assessment specifications also document basic information about *how* the specified knowledge, skills, and abilities will be measured, providing details about the test purpose and design. Test specifications commonly include sections on the test purpose, the target population, and a test blueprint that outlines the types and quantity of test tasks and how they will be scored. For English language proficiency assessments intended for international test-takers, one major consideration is the choice of which of the different varieties of English should be used in the assessment items and instructions. For instance, should the test include standard North American English or a sampling of standard global English varieties? Such decisions should be made on the basis of the intended purposes for the assessment scores, as well as the intended test-taker population. A panel of experts who are familiar with the purpose of the assessment and the intended population can play an important role in ensuring valid interpretations of the scores from an assessment. The composition of such a panel should include individuals who represent different stakeholder groups, including test-takers and decision makers, to ensure that the design and content of the assessment is not biased in favor of any identifiable group of test-takers.

Because the population of test-takers who take English language proficiency assessments includes a wide range of proficiency levels, test directions and test items should be written to be fully accessible to

the target test-taker population. Test directions should be written at a level of English that is well below the typical proficiency level of the intended test-taker population. Example items should be included as part of the instructions. Test directions should be designed to maximize understanding of the task being presented and to minimize confusion on the part of test-takers as to what they are expected to do. Complex language should be avoided unless it is directly related to the language ability being assessed. Test items should be written using vocabulary and sentence structures that are widely accessible to test-takers.

With regard to the presentation of test materials, assessment developers should take into account formatting considerations (e.g., fonts, font sizes, and the location of line breaks in sentences and paragraphs). Also to be considered carefully is that the use of visual or graphical materials is clear, tasteful, and free from cultural bias for all test-takers. Because of the diversity of cultural and linguistic backgrounds within the population of international English language proficiency test-takers, it is important to consider how the test materials may appear to test-takers who are less familiar with English presentation conventions.

Using Selected-Response Questions

Selected-response questions are widely used in language assessment for two main reasons. First, because they restrict the responses that a test-taker can provide, these questions can be scored quickly and objectively. Selected-response questions are usually scored dichotomously, i.e., right or wrong. Second, well-written selected-response questions can gather information about a broad range of aspects of the target construct within a relatively short time. In this section, we discuss concepts that need to be considered when writing selected-response items, focusing on the most frequently used selected-response question type—multiple choice. Before discussing guidelines for developing questions, recommendations will be made for the creation of reading and listening passages, both of which are typical types of input that test-takers are asked to process in order to answer questions.

Guidelines for writing reading and listening passages

- *Characteristics of the input that need to be considered in writing, reading and listening passages.* There are multiple factors that can influence the comprehension difficulty and cognitive load of a reading or listening passage. Topic, presence or absence of a clear organizational structure, length, vocabulary, grammatical complexity, discourse structure (e.g., monologue, dialogue, or multiparty discussion), and genre (e.g., weather report, academic lecture) are some of the factors that are likely to influence the difficulty of both reading and listening passages. A speaker's rate and rhythm of speech, native accent, volume, and pitch also need to be considered in creating passages for listening assessments. It should be noted that, to the greatest extent possible, any decision about these features of the language input should be based on the target construct to be measured. For example, if the construct is defined as "ability to understand a dialogue that is found in a typical teacher-student conference about school life," the passages used should contain the features that are appropriate for this context.

- *Influence of topical knowledge*. Topical knowledge plays an important role in comprehending a passage. Depending on the way the construct is defined, topical knowledge can be part of the construct or a source of construct-irrelevant variance. For example, in an English assessment whose purpose is to assess test-takers' readiness to major in chemistry in an English-medium university, using passages that assume a certain level of topical knowledge on chemistry is acceptable given that test-takers are expected to have the knowledge. However, if the purpose of an English assessment is to assess general proficiency, it is strongly recommended that the passages and items not assume any topical knowledge on the part of the test-takers. Any information that is required to answer items correctly should be provided within the given passage so as not to disadvantage test-takers who are not familiar with this information prior to taking the assessment.
- *Incorporating visual input.* When visual input is provided along with language input, test developers should first investigate how the input will influence test-takers' answering of questions. Particularly when visuals are intended to help test-takers comprehension by providing information that is relevant to the content of the passage, investigations into how test-takers actually use (or fail to use) the visuals, and the influence of these test-taker behaviors on their test performance, should be conducted.

Guidelines for writing multiple-choice questions

- *Ensuring that skills and knowledge that are unrelated to the target construct do not influence test-taker performance*. Test developers should pay careful attention to what they desire to assess as compared to what the test actually measures. In a reading comprehension assessment, for example, questions about a reading passage are designed to see whether a test-taker has understood what is covered in the passage. Therefore, stems and response options in multiple-choice questions should be written in language that requires much lower level proficiency than the proficiency level that is required to understand the reading passages. Care should also be taken when providing stems and response options in written form in a listening assessment. In such a situation, reading ability, in addition to listening ability, is required for a test-taker to answer the listening comprehension items correctly. Therefore, if test-takers' reading ability is expected to be lower than their listening ability, which is often true for younger and/or less-proficient English learners, the language used for stems and response options should be as simple as possible. Alternative item presentation schemes can be considered in order to minimize the effects of irrelevant abilities on measurement of the test construct. To return to the example of a multiple-choice listening item, providing questions in both written and spoken forms, or providing nonlanguage picture options might reduce the impact of reading ability on items that measure listening ability.
- *Consider providing instructions in the test-taker's first language*. This can be a way to minimize the probability that a test-taker gets a question wrong because the language used in the instructions and question stems is too difficult to understand, even though the test-taker did actually understand the reading/listening passage. However, this will only be practical when a small number of native languages are spoken in the test-taker population. If the first-language

diversity of the target population is large, it may be cost prohibitive to produce many dozens of translations. This may create a situation in which some first-language versions are unavailable, thus raising an equity issue.

- *Ensuring that questions are not interdependent*. The information in one question should not provide a clue to answering another question. This is particularly true in reading and listening comprehension assessments, in which more than one comprehension question is asked about one passage.

Scoring Constructed-Response Test Items

Constructed-response items, which require test-takers to produce a spoken or written response (e.g., write an essay), are also common tasks used to assess English language ability. Scoring constructed-response items pose various challenges which may or may not be encountered with assessments that use dichotomous (right/wrong) scoring procedures, as in many selected-response questions. The guidance provided in this section draws on information found in an existing ETS publication, *Guidelines for Constructed-Response and Other Performance Assessments* (Educational Testing Service, 2005). In this section we focus on scoring issues, including both human and automated scoring processes, for constructed-response test items on English language proficiency assessments.

In developing scoring specifications and scoring rubrics for constructed-response items, a number of important questions need to be answered, including the following:

- *What is the most appropriate scoring approach for scoring responses to each task?* There are a number of commonly used approaches for scoring constructed-response and other performance assessments, and it is important in the scoring specifications to identify the approach that will be used. **Analytical scoring** requires raters to determine whether specific characteristics or features are present or absent in a response. For instance, an analytic scale designed to assess English language speaking ability would contain two or more subscales of speaking ability, such as fluency, pronunciation, communicative competence, or vocabulary. Each of these subscales would be scored on a proficiency scale (e.g., novice, proficient, advanced proficient). When an analytic approach is used, differential weightings can be attached to different subscales, depending on the purpose of the assessment. For example, in a speaking assessment which measures non-native graduate students' readiness to teach undergraduate content classes, an analytic scoring approach in which pronunciation is given more weight than other subscales (i.e., language use, organization, and question handling) might be used. This decision would be made because pronunciation is more closely related to the difficulties that undergraduate students experience in classes taught by non-native teaching assistants. **Holistic scoring** employs a scoring scale and training samples to guide raters in arriving at a single qualitative evaluation of the response as a whole. A holistic English speaking ability scale would not subdivide speaking into subscales. The scale would contain a single description of each proficiency level, and raters would assign test-takers only one general speaking ability score. Another scoring approach that can be used to assess constructed-response

items is **primary trait scoring**. Primary trait scoring involves using a holistic scale that relies on features of the task that test-takers are required to complete. Multiple-trait scoring analogously uses analytic scales which include features of the task. Rating scales for trait scoring emphasize the inclusion or exclusion of task completion. For example, an evaluation of a summary writing task might include how many of the main points in the passage to be summarized are included in the summary. Another scoring approach is **core scoring**, which identifies certain essential core features that must be present plus additional nonessential features that allow the response to be given higher scores. For example, in assessing a summary writing task, the scales might require the summary to include the main point of the text to be summarized to achieve a certain threshold rating. If the main point is not included, ratings cannot reach this threshold regardless of other features of the summary. If the main point is included, ratings can increase based on other features, such as the extent to which words or phrases are copied from the original text. The most important criterion to be considered when selecting a scoring approach should be the purpose of the assessment. If diagnostic information needs to be provided about test-takers' strengths and weaknesses, an analytic scoring approach should be selected over a holistic scoring approach. Other factors that affect the choice of a scoring approach include the required turnaround time for score reports, the qualifications of raters, and the number of qualified raters.

- *How many score categories (or points) should be assigned to each task?* As a general principle, the ability to be assessed should be subdivided into abilities that follow from an accepted theory of language or communication, and there should be as many score categories available as raters can consistently and meaningfully differentiate. The appropriate number of score categories depends on a number of factors: (1) the purpose of the assessment, (2) the task demands, (3) the scoring criteria, and (4) the number of distinctive categories that can be identified among the responses. Conducting pilot testing of sample items or tasks with a representative sample from the test-taker population will help to confirm the number of score categories that is appropriate. For instance, if a one-on-one oral interview is used to assess speaking ability, it might be desirable, based on theory, to assign a score for fluency, pronunciation, communicative competence, vocabulary, and grammar. However, it may be determined from pilot studies that evaluators cannot manage to assign scores for more than four categories, and grammar and vocabulary are highly related — that is, they cannot be meaningfully distinguished by evaluators. A decision to combine grammar and vocabulary and use four subscales might be made.
- *What specific criteria should be used to score each task?* Scoring rubrics are descriptions of test-taker performances which are used to assign a score for a test-taker's performance on a task. For example, a scoring rubric might have five ability bands, ranging from excellent to poor, which describe five different writing ability levels. In developing scoring rubrics, or scales, for a constructed-response item, one should consider the purpose of the assessment, the ability levels of the test-takers, and the demands of the task. The scoring rubric should be aligned with the directions and task to ensure that raters are applying the appropriate scoring criteria and are not influenced by atypical response formats or the presence of extraneous information that could

bias their scoring. For instance, if the purpose of the task is to elicit academic language, the instructions should direct the test-taker to use academic prose, and the rubrics should be designed to assess the extent to which the writing is appropriately academic.

- *How should raters be trained to score constructed responses?* A rater is a person who scores constructed responses based on scoring rubrics. Raters are typically trained and certified, and also retrained after a certain period of certification, by a testing organization. After being certified, a rater scores test-taker responses based on scoring rubrics. The scoring rubrics for the assessment should be the focus of the training and retraining of raters.
- *How many raters should score each response?* If constructed-response items are used on an assessment whose scores are used to make high-stakes decisions, it is critical that the processes developed for scoring responses be made as reliable as possible. A major determinant of the reliability of scores is the number of raters assigned to each item, with increased reliability resulting from the use of additional well-trained raters. There are factors other than the number of raters that affect the reliability of scores assigned to constructed-response items. The type of responses and the number of total constructed-response items in one assessment are two such factors. Some types of responses can be scored very reliably. For example, responses to items for which the scoring criteria are well-specified and the relevant characteristics or features are easily identifiable may be scored reliably without multiple raters. Such items generally require concise, focused answers, such as the age of one of the characters in the passage of a reading comprehension text. In contrast, responses to more complex items, such as writing an essay, require nuanced judgments to score and are likely to require two or more raters in order to achieve acceptable levels of scoring reliability. Deciding the number of raters for each response is also related to the total number of constructed-response items in an assessment. Research in language assessment has found that increasing the number of tasks is often more effective at increasing reliability than increasing the number of raters per task (e.g., Lee, 2006).
- *Is it preferable to use human raters, automated scoring, or a combination of the two?* The automated scoring of constructed responses using specialized software is becoming increasingly common for large-scale assessments of English language proficiency. If automated scoring is used, the scoring criteria and procedures used by the software should be explicated. Automated scoring can be used independently or in conjunction with human ratings. If a combination of automated scoring and human scoring is used, then information on the scoring processes and procedures for combining the scores of the automated and human scoring should be clearly stated. Just as in the design of scoring rubrics for human raters, the design of automated scoring applications should reflect current theories of language ability.

As in the development of specifications for the scoring of constructed-response items, a number of important considerations need to be made regarding the processes for scoring, including the following:

- *What are the necessary qualifications for the individuals who will serve as raters?* In general, the most important qualification to be a rater is whether the individual has had prior experience in

observing and judging the performance being assessed. Familiarity with the range of responses in the test-taker population on a constructed-response item is another important qualification for a rater. In addition, other considerations for choosing raters include their personal demographic characteristics, the geographical region that they represent, and their professional background or education.

- *What responsibilities do scoring leaders have, and what qualifications should they possess?* Scoring leaders are responsible for ensuring that raters are properly trained, guaranteeing that the scoring process is carefully monitored, assisting in planning the scoring process, and ensuring that the entire scoring process is conducted in a manner that leads to the most valid and reliable scores. Therefore, competent and experienced raters are prime candidates to serve as scoring leaders for constructed-response items.
- *How should raters be trained to apply the scoring criteria?* It is of the utmost importance that all raters be carefully trained to accurately and consistently apply the same scoring criteria in the same manner for a given constructed-response item. Ideally, training should be conducted interactively, although it does not need to be done face-to-face in a central location. Raters should be trained to apply the scoring rubrics accurately and consistently, and feedback should be provided during training and in scoring sessions to ensure that each rater is applying the scoring criteria correctly. The use of benchmark responses (exemplars for each score point) is useful in helping raters understand and recognize the features of a response that produced a particular score. Training sessions usually begin with an explanation of the task to be assessed. An explanation of how to use the scales generally follows. Examples of benchmark responses with explanations for scoring followed by practice rating often conclude a rater training session.
- *What procedures can be used to ensure that the raters are scoring accurately and consistently?* The introduction of prescored responses into operational scoring is helpful in monitoring whether raters are maintaining accuracy and consistency during a scoring session. This procedure can also serve as a quality control check on whether raters are meeting the standards for scoring. In addition, providing retraining for the raters regularly (for example, at the beginning of each scoring session) with new sets of prescored responses and/or reviewing training materials, helps to ensure that they are scoring accurately and consistently.
- *What steps can be taken to minimize bias in scoring?* Designing and conducting a scoring process that minimizes inappropriate influences on raters is the best strategy for minimizing bias in scoring. The use of rating scales that are as unambiguous as possible and thorough training sessions for raters can help to minimize bias in scoring. Raters can also be alerted to possible biases that may present themselves in particular contexts. For instance, if some raters evaluating speaking performances share the same first language as the test-takers while others do not, care should be taken to be sure that these two groups of raters both evaluate different accents similarly.
- *How can discrepancies between raters be resolved?* In situations where multiple raters are used, procedures for resolving differences in scores should be clearly stated and should be based on the

goal of maximizing the meaningfulness of the final score. Procedures for requiring additional ratings, and how to use the additional ratings, need to be made clear. For instance, adjudication can be mandated if the size of discrepancy in ratings by different raters exceeds a predetermined allowable limit. If one rater assigns a rating of highly proficient (the highest score possible) and another rater assigns a rating of novice (the lowest score possible) on a five-point scale of speaking ability, it may be determined that the two ratings are not consistent enough to indicate the ability of the test-taker. A third rating might be required, and procedures for determining how to use the third rating need to be explicated. For instance, all three scores could be combined, or the two closest could be averaged. Determining the best procedure for adjudicating scores should be based on the context of the assessment. For instance, if a highly experienced rater is used as a third rater when two less experienced raters assigned unacceptably different scores, the highly experienced rater's assigned score may be the most appropriate. On the other hand, if all raters have similar experience, it may be best to take an average of the scores of the additional rating and the rating to which it is most similar.

- *When automated scoring is used, how should the use of scoring engines be documented?* It is important for testing program administrators, as well as test-takers, to fully understand how automated scoring processes operate. This makes it possible for these important stakeholders to appropriately prepare for the assessments. The use of automated scoring should be as transparent as the scoring processes that depend on human raters. Just as it is essential that stakeholders be made aware of the rating scales and scoring criteria when human raters are used, it is necessary to provide the criteria that computer software packages use when assigning scores to test-takers. For example, scoring criteria used by computer software packages designed to assess a speaking test should be made available.

Statistical Analyses of Test Results

English language proficiency assessments for international test-takers, as well as any other group, should be statistically analyzed to help determine the effectiveness of the assessments at the item level and at the test level. These analyses should be conducted after each testing administration under the direction of individuals trained in interpreting and evaluating statistical results of English language proficiency assessments. Because English language proficiency assessments for international test-takers may differ with respect to item and test formats, modes of delivery, skills assessed, target population, and other characteristics of the assessment, no single set of analyses is appropriate for all assessments. In addition, limitations of resources and available expertise may limit the types of analyses that are conducted.

For assessments of English language proficiency for international test-takers, the analyses to be conducted should ideally include the following:

- *Analysis of overall test performance in terms of average scores and variability of scores for test-takers.* These analyses should be conducted for all test-takers as well as test-taker subgroups of interest, such as those defined by demographic indicators (e.g., sex and age) or other variables of interest (e.g., first language, years of study in a country where English is the primary language, country of origin, grade level, and size and type of school). These analyses can serve to identify

whether the assessment is at an appropriate difficulty level for most test-takers, as well as to identify performance differences among subgroups of test-takers. Analyses of the variability of scores can aid in understanding the full range of abilities and skills present among test-takers. In particular, it may be helpful to determine whether there is a large proportion of test-takers, overall or within particular subgroups, scoring very low (i.e., below the level of performance expected by chance on multiple-choice items) or very high (i.e., a large proportion receive scores near the maximum point value) on the assessment.

- *Analysis of the properties of items and testlets.* Statistical analyses are informative with regard to how well each item on an assessment is functioning. Items should be analyzed with regard to their difficulty level (how hard the item is), discrimination value (how well the item separates high- and low-ability test-takers), and association (correlation) with other items, as well as test section score and total score. Distracter analysis should also be conducted for multiple-choice items, in which the statistical properties of an item's distracters (incorrect options) are assessed to determine how attractive they are to test-takers with differing abilities on the assessment. Distracter analysis can help to identify problems with an item, and, in particular, can point to whether there may be more than one correct answer or whether one of the distracters may be more plausible to high-ability test-takers. These same analyses can also be conducted for testlets (sets of items that share a common prompt or stimulus, such as a set of items associated with a common reading passage). If sample sizes are sufficiently large, these analyses can also be conducted for subgroups of interest. These analyses can be conducted using either classical test theory or item response theory, both of which require training in assessment and statistical analyses.
- *Analysis of the reliability of the assessment and scoring processes.* Reliability, which is a measure of the extent to which test scores are consistent indicators of the ability that the test aims to measure, is a key technical characteristic of any assessment. It is critical to have evidence on the reliability of an assessment as well as on the reliability of the processes used for items that require judgment in scoring. The statistics commonly used to describe the reliability of an assessment are one or more reliability coefficients (e.g., Cronbach's alpha as a measure of internal consistency). In addition, if there are items that require constructed-response scoring using multiple raters, then a measure of the inter-rater agreement between raters is needed as well as an estimate of reliability. While useful, simple percent agreement approaches, in which the number of rating pairs that are in exact agreement and varying levels of approximate agreement are tallied, should not be used independently to estimate reliability. While these approaches generally provide an indication of the reliability of the scores, it is possible that raters use a small portion of the scale (e.g., only a 3 and 4 on a 1 to 5 scale) which would not lead to reliable scores, but would result in high coefficient agreement indices. A more appropriate approach to estimating reliability of performance assessments is to use a correlational approach, such as using Cronbach's alpha, in which each rating is treated as an item. Of course, reporting the combination of a percent agreement and a correlational approach generally provides more useful information about the reliability of an assessment than simply reporting one or the other.

- *Analysis of the standard errors of measurement and/or the test information functions.* Related to the reliability of an assessment are the standard errors of measurement (SEM; when using a classical test theory approach) and/or the test information functions (if item response theory models are employed). SEM indicates an estimate of the accuracy of the score, and it is used to construct a confidence interval within which a test-taker's true score is likely to lie, considering the test-taker's obtained score and the level of error of the assessment. For example, a 95% confidence interval (i.e., constructed by the formula of "obtained score – 1.96×SEM ≤ estimated true score ≤ obtained score + 1.96 × SEM") identifies that there is 95% probability that the true score of a test-taker, free of error, is within the interval. SEM analyses are useful for understanding the precision of test scores, and they make it possible for test users to take more circumspect actions when making high-stakes decisions about test-takers. Precision of test scores is indicated by test information functions in measurement models grouped as item response theory (IRT). The main difference of the test information function from SEM is that test information functions show different measures of precision at different scale points of an assessment, whereas there is only one measure of precision for the entire assessment when SEM is used. For example, it can be observed that an assessment can provide more precise scores with less error for more proficient test-takers, indicated by a high value for the information function at the higher end of the score points, while the same assessment provides less-precise scores at the lower end. This information is important for identifying the parts of the score scale where there is sufficient precision to support the use of the assessment for different purposes, such as for making placement decisions or proficiency classifications.
- *Analysis of item functioning across different subgroups.* When possible, analyses should be conducted to determine whether any of the items exhibit differential item functioning (DIF). DIF is a statistical method that examines if test-takers from different groups have the same probability of answering an item correctly, given that their ability level on the target construct is the same. More specifically, an item displays DIF when the probability of answering it correctly differs across groups, even though their ability levels are the same. DIF causes a concern for fairness because it implies that factors other than the target construct influence performance on a DIF-detected item. DIF analyses should be conducted on the subgroups of interest for a specific English language proficiency assessment; these subgroups may include those defined by demographic indicators or other variables of interest. For instance, story problems on a mathematics test might function differently for English learners than for highly proficient English users. If the goal of the test is to assess mathematics ability, but an English ability level which the English users do not possess is needed to understand the story problems, these items might be biased against English learners. When an IRT approach for DIF analyses is used, we recommend a minimum sample size of 500 test-takers for each subgroup included. If there are theoretical considerations, those should be the primary factors in deciding which subgroup should be identified as the focal group for DIF analyses and which subgroup should be the reference group. DIF analyses usually should not combine multiple-choice items with constructed-response items in computing a total score for matching test-takers. This recommendation arises because there is substantial evidence that

subgroups of test-takers do not always perform similarly on these different types of items. Items that exhibit DIF should be reviewed to identify possible factors that may have produced the results. A review of the linguistic features of items exhibiting DIF can be very useful in gaining an understanding of the possible causes of DIF. For English language proficiency assessments, the causes could be related to the similarity of English to test-takers' first language, to how English language instruction is delivered, or to cultural background. The linguistic review should be conducted jointly with applied linguists and experts in language assessment.

Validity Research

The evaluation of English language proficiency assessments for international test-takers should include studies designed to determine the extent to which the assessments are valid indicators of the abilities they claim to assess. These studies are not a substitute for the item- and test-level analyses described above, but a critical complement. It may not be necessary to conduct all of these studies for every administration; rather, these studies should be considered in the broader context of creating and carrying out a coherent validity research agenda for an assessment.

- *Research on the internal factor structure of the assessment.* A study on the internal factor structure of an assessment is one widely used approach for evaluating construct validity, that is, the extent to which the test measures the ability it is designed to assess. An analysis of internal factor structure is used to evaluate whether the assessment design as identified in the test specifications is consistent with the responses from test-takers. These investigations are often conducted using factor analytic approaches. If the results of these studies indicate that one or more items deviate from the intended test design, then these should be investigated further to determine whether the items are, in fact, measuring the construct that they are designed to assess. For example, if a test is designed to assess the skills of reading, writing, speaking, and listening, but the results indicate that the assessment only measures two distinguishable factors, the items should be investigated to determine what they are actually measuring, and why they are not measuring each of the intended abilities. In addition, research on the internal factor structure of an assessment should include analyses by subgroups of interest, since it is important to confirm that the assessment has the same underlying structure for different groups of test-takers. For English language proficiency assessments, an analysis of the internal factor structure across first language subgroups is especially critical for gaining an understanding as to whether the assessment is measuring the same construct across different language subgroups.
- *Research on the concurrent or predictive validity of the assessment.* A study on the predictive validity of an assessment is used for identifying suitable outcome measures (such as academic tasks or classes that require proficiency in English) and determining the relationship between scores on the assessment and the criterion performance (through correlation or regression analysis). For example, if an assessment aims to assess the ability to communicate orally in an academic context, assessment results could be correlated with results of performances on actual oral academic tasks, such as giving an academic presentation or participation in an academic

small group discussion. It is also useful to investigate whether there is differential prediction by subgroups of interest, such as first language groups, as this is critical evidence for demonstrating that the assessment is functioning similarly for the different groups of test-takers. In a similar fashion, if proficiency classifications of the test-takers are one outcome from the assessment, then a study that investigates whether different levels of proficiency classification are associated with different levels of later performance is also strongly recommended.

- *Research on the cognitive processes used by test-takers to complete the assessment.* Studies of the cognitive processes used by test-takers to respond to items and complete tasks are important for determining the extent to which an assessment is a valid indicator of the abilities it aims to assess. Such a study would entail identifying a small sample of representative test-takers and having them review and respond to items and tasks as they verbalize their response strategies. Think-aloud verbal protocols, in which test-takers verbalize their thought processes while completing the assessment, are commonly used for this purpose. In particular, it is important that the test-takers selected represent the diversity of proficiency levels, first languages, age and grade levels, as well as other significant variables, in the test-taker population. This approach can be very informative in determining the degree to which test-takers understand the items and tasks presented to them and the extent to which they use the processes and abilities related to the construct of interest when selecting or constructing an appropriate response.

Providing Guidance to Stakeholders

It is crucial that all stakeholders (individuals and groups who are potentially affected by an assessment) are provided with appropriate information about the assessment. Whenever feasible, this information should be provided in language that is accessible to all stakeholders. Stakeholders include test-takers (and parents of young test-takers), teachers, teaching institutions that need to prepare their students for an assessment, and decision makers who must refer to test results to make inferences about test-takers.

- Stakeholders should be provided with information about the target population of the test, which makes it clear for whom the test is appropriate and for whom it is not appropriate. For instance, if the test is designed to assess intermediate- to advanced-level academic English reading for adults, it should be made clear that the test results will *not* indicate young novice test-takers' ability to communicate orally in English. Such documents should provide information about important features of the assessment, such as the target language abilities that are measured in the assessment, the target population, the overall structure of the assessment, the format of the items, the scoring procedures, and the uses of the assessment results. It is imperative that such documents be made available in language that is accessible. Stakeholders should have a comfortable familiarity with the test well in advance of administration. Early availability of information to teachers leads to preparation and instruction most appropriate to prepare students for the test and avoid unfairly disadvantaging certain students.

- The information needed to effectively take the test should be made available to stakeholders. The directions for taking the test, accompanied by example items which are representative of the actual item formats and types, should be accessible. This information should be accompanied by information about what the test is designed to measure (e.g., the ability to make inferences from listening passages) accompanied by advice for taking the test. Care should be taken to consider accessibility to all students. Economics, computer access, and the language of communication, among other factors, should be taken into consideration to ensure that all potential test-takers have access to information needed to effectively take the test.
- Stakeholders should be provided with information about how the scores are obtained. For instance, information should be provided about the scoring criteria for all items in an assessment, including scoring rubrics for constructed-responses and criteria for scoring selected-response items (e.g., how nonattempted vs. incorrectly answered items are scored). Also, if raw scores are not used, simple explanations of the scaling techniques should be provided. Whenever feasible, a scaling metric which provides the conversion from raw to scaled scores should be made available to stakeholders.
- Information about test-takers' rights when it comes to having tests rescored, cancelled, or retaken should be made available to stakeholders. Procedures for exercising these rights, including information about who to contact and what to include in the communication, should be made clear.

Decision makers, or test users, are the stakeholder group that actually creates demand for assessment use in real-world situations, which could include state officials, school administrators, or potential employers. In addition to the information provided to all stakeholders, decision makers should be provided with the following:

- *The usefulness of an assessment for a specific context.* Test users should be made aware that each assessment use situation is unique and, therefore, the usefulness of an assessment should be re-evaluated every time the assessment is considered for an additional purpose.
- *Justifying an existing assessment for an additional use*. Test users should be informed that when they plan to use an assessment for a purpose that is not officially endorsed by the test developer, they should investigate whether the use can be justified and what evidence they need to support the use of the assessment for their purpose. They should be informed that if they do not have enough information to make the decision to use the test, they can contact the test developer to provide information to help gauge the appropriateness of the test for the decision maker's intended purpose. Test developers should provide defensible timely replies to requests from the score users and provide suggestions for alternative test use if it is determined that the test may not be appropriate for the desired purpose.
- *Taking precautions before using an assessment for making multiple decisions*. Score users should be informed that using scores from a single test to make multiple decisions can be practical. However, for each individual decision to be made based on the test scores, unique evidence should be collected regarding the appropriateness of the assessment for this purpose.

Giving a Voice to Stakeholders in the Testing Process

All test-takers should be given a voice in the testing process. Whenever possible, at different stages of test development, feedback from test-takers, parents, teachers, school administrators, employers, and other stakeholders should be sought and incorporated for modification or revision of the assessment. It is important that this feedback be gathered before, during, and after the assessment is put to operational use, the stage at which assessment results are used for making decisions about test-takers. Below are some guidelines to provide stakeholders with the opportunity to have their voices heard.

- *Having an open channel of communication for stakeholders*. A process, as well as a mechanism for incorporating this feedback into the testing process, should be made available to all stakeholders. For example, a public web page that requests suggestions for the assessment can be made available.
- *Responding appropriately to stakeholders' feedback*. Test developers should respond appropriately to the feedback by either revising their assessments based on the recommendations or provide justification for not incorporating these suggestions into the assessment. Decisions for making or not making the suggested changes should be based on a balance of satisfying other recommendations for quality assessments discussed in this publication.

Summary

In this publication, we have provided guidelines, based on prior research and experience with testing programs, for best test development practices to ensure validity and fairness for international English language proficiency assessments. This publication adds to the ETS series on fairness and is focused on recommended best practices for English language proficiency assessments taken by international test-taker populations. We have covered a broad range of topics, including definitions of key terms, planning and developing an assessment, scoring selected- and constructed-response items, statistical analyses, validity research, and providing guidance to and receiving feedback from stakeholders. We intend for these guidelines to be widely disseminated and hope that the applications of the guidance contained within will advance and improve the assessment of English language proficiency of candidates worldwide. As with any scientific endeavor, continuous improvement must be a goal, and so we encourage anyone involved in the practice of assessing English language proficiency to share their wisdom and experiences with us so that we can continue to refine and improve this publication in the future.

Bibliography

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Association of Language Testers in Europe. (2001). *The ALTE code of practice*. Retrieved from http://alte.columnsdesign.com/attachments/files/code_practice_eng.pdf

European Association for Language Testing and Assessment. (2006). *EALTA guidelines for good practice in language testing and assessment.* Retrieved from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

Educational Testing Service. (2002). *ETS standards for quality and fairness.* Princeton, NJ: Author.

Educational Testing Service. (2003). *ETS fairness review guidelines*. Princeton, NJ: Author.

Educational Testing Service. (2005). *Guidelines for constructed-response and other performance assessments.* Princeton, NJ: Author.

Educational Testing Service. (2007). *ETS international principles for fairness review of assessments*. Princeton, NJ: Author.

Educational Testing Service. (2009a). *ETS international principles for fairness review of assessments*: *A manual for developing locally appropriate fairness review guidelines in various countries.* Princeton, NJ: Author.

Educational Testing Service. (2009b). *Guidelines for the assessment of English language learners.* Princeton, NJ: Author.

International Language Testing Association. (2007). *Guidelines for practice*. Retrieved from http://www.iltaonline.com/images/pdfs/ILTA_Guidelines.pdf

International Test Commission. (2000). *International guidelines for test use*. Retrieved from http://www.intestcom.org/upload/sitefiles/41.pdf

The Japan Language Testing Association. (n.d.). *The JLTA code of good testing practice*. Retrieved from http://www.avis.ne.jp/~youichi/COP.html (May 23, 2012)

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131–166.

Listening. Learning. Leading.®

www.ets.org

99832-99832 • Y813E.250 • Printed in U.S.A.