Designation: F2889 – 11

Standard Practice for Assessing Language Proficiency[1]

This standard is issued under the fixed designation F2889; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 *Purpose*—This practice describes best practices for the development and use of language tests in the modalities of speaking, listening, reading, and writing for assessing ability according to the Interagency Language Roundtable (ILR)[2] scale. This practice focuses on testing language proficiency in use of language for communicative purposes.

1.2 *Limitations*—This practice is not intended to address testing and test development in the following specialized areas: Translation, Interpretation, Audio Translation, Transcription, other job-specific language performance tests, or Diagnostic Assessment.

1.2.1 Tests developed under this practice should not be used to address any of the above excluded purposes (for example, diagnostics).

2. Referenced Documents

2.1 *ASTM Standards:*[3]

F1562 Guide for Use-Oriented Foreign Language Instruction

F2089 Practice for Language Interpreting

F2575 Guide for Quality Assurance in Translation

3. Terminology

3.1 *Definitions:*

3.1.1 *achievement test, n*—an instrument designed to measure what a person has learned within or up to a given time based on a sampling of what has been covered in the syllabus.

3.1.2 *adaptive test, n*—form of individually tailored testing in which test items are selected from an item bank where test items are stored in rank order with respect to their item difficulty and presented to test takers during the test on the basis of their responses to previous items, until it is determined that sufficient information regarding test takers' abilities has been collected. The opposite of a *fixed-form test*.

3.1.3 *authentic texts, n*—texts not created for language learning purposes that are taken from newspapers, magazines, etc., and tapes of natural speech taken from ordinary radio or television programs, etc.

3.1.4 *calibration, n*—the process of determining the scale of a test or tests.

3.1.4.1 *Discussion*—Calibration may involve anchoring items from different tests to a common difficulty scale (the theta scale). When a test is constructed from calibrated items then scores on the test indicate the candidates' ability, i.e. their location on the theta scale.

3.1.5 *cognitive lab, n*—a method for eliciting feedback from examinees with regard to test items.

3.1.5.1 *Discussion*—Small numbers of examinees take the test, or subsets of the items on the test, and provide extensive feedback on the items by speaking their thought processes aloud as they take the test, answering questionnaires about the items, being interviewed by researchers, or other methods intended to obtain in-depth information about items. These examinees should be similar to the examinees for whom the test is intended. For tests scored by raters, similar techniques are used with raters to obtain information on rubric functioning.

3.1.6 *computer adaptive test, n*—a test administered by a computer in which the difficulty level of the next item to be presented to test takers is estimated on the basis of their responses to previous items and adapted to match their abilities.

3.1.7 *construct, n*—the knowledge, skill or ability that is being tested.

3.1.7.1 *Discussion*—The construct provides the basis for a given test or test task and for interpreting scores derived from this task.

3.1.8 *constructed response, adj*—a type of item or test task that requires test takers to respond to a series of open-ended questions by writing, speaking, or doing something rather than choose answers from a ready-made list.

3.1.8.1 *Discussion*—The most commonly used types of constructed-response items include fill-in, short-answer, and performance assessment.

[1] This practice is under the jurisdiction of ASTM Committee F43 on Language Services and Products and is the direct responsibility of Subcommittee F43.04 on Language Testing.

Current edition approved May 1, 2011. Published June 2011. DOI: 10.1520/F2889-11.

[2] Interagency Language Roundtable, Language Skill Level Descriptors (http://www.govtilr.org/Skills/ILRscale1.htm).

[3] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

3.1.9 *content validity, n*—a conceptual or non-statistical validity based on a systematic analysis of the test content to determine whether it includes an adequate sample of the target domain to be measured.

3.1.9.1 *Discussion*—In order to achieve content validity, an adequate sample involves ensuring that all major aspects are covered and in suitable proportions.

3.1.10 *criterion-referenced scale, n*—a graduated and systematic description of the domain of subject matter that a test is designed to assess; (or) a rating scale that provides for translating test scores into a statement about the behavior to be expected of a person with that score and/or their relationship to a specified subject matter.

3.1.10.1 *Discussion*—A criterion-referenced test is one that assesses achievement or performance against a cut score that is determined as a reflection of mastery or attainment of specified objectives. Focus is on ability to perform tasks rather than group ranking.

3.1.11 *cut score, n*—a score that represents achievement of the criterion, the line between success and failure, mastery and non-mastery.

3.1.12 *dichotomous scoring, n*—scoring based on two categories, e.g., right/wrong, pass/fail. Compare to *polytomous scoring*.

3.1.13 *equated forms, n*—two or more forms of a test whose test scores have been transformed onto the same scale so that a comparison across different forms of a test is made possible.

3.1.14 *expert panel, n*—a group of target-language experts who take a test under test-like conditions and provide comments about any problem areas.

3.1.14.1 *Discussion*—An expert panel should include at least 8 members. Panel members receive training before they take the test in order to ensure that their comments will be helpful.

3.1.15 *face validity, n*—the degree to which a test appears to measure the knowledge or abilities it claims to measure, based on the subjective judgment of an observer.

3.1.16 *fixed-form test, n*—a test whose content does not vary in order to better accommodate to the examinee's level of knowledge, skill, ability or proficiency. The opposite of an *adaptive test*.

3.1.17 *genre, n*—a type of discourse that occurs in a particular setting, that has distinctive and recognizable patterns and norms of organization and structure, and that has particular and distinctive communicative functions.

3.1.18 *ILR scale, n*—a scale of functional language ability of 0 to 5 used by the Interagency Language Roundtable.[2]

3.1.18.1 *Discussion*—The range of the ILR scale is from 0—no knowledge of a language to 5—equivalent to a highly educated native speaker.

3.1.19 *indirect test, n*—a test that measures ability indirectly, rather than directly.

3.1.19.1 *Discussion*—An indirect test requires examinees to perform tasks that are not directly reflective of an authentic target-language use situation. Inferences are drawn about the abilities underlying the examinee's observed performance on the indirect test.

3.1.20 *interpretation, n*—the process of understanding and analyzing a spoken or signed message and re-expressing that message faithfully, accurately and objectively in another language, taking the cultural and social context into account.

3.1.20.1 *Discussion*—Although there are correspondences between the skills of interpreting and translating, an interpreter conveys meaning orally, while a translator conveys meaning from written text to written text. As a result, interpretation requires skills different from those needed for translation.

3.1.21 *inter-rater reliability, n*—the degree to which different examiners or judges making different subjective ratings of ability agree in their evaluations of that ability.

3.1.22 *intra-rater reliability, n*—the degree to which an individual examiner or judge renders consistent and reliable ratings.

3.1.23 *item, n*—one of the assessment units, usually a problem or a question, that is included on a test.

3.1.23.1 *Discussion*—Test items provide a means to measure whether a test taker can perform a task and are scorable using a scoring rubric or answer key. Successful or unsuccessful performance on an item contributes information to the test taker's overall score. Examples of item types include: multiple choice, constructed response, cloze, matching and essay prompts.

3.1.24 *item response theory (IRT), n*—the theory underlying statistical models that are used to describe the relationship between a student's ability level and the probability of success on a test question.

3.1.24.1 *Discussion*—IRT encompasses latent trait theory; logistic models; Rasch models; 1, 2, and 3 parameter IRT; normal ogive models; Generalized Partial Credit models; and Samejima's Graded Response model.

3.1.25 *language proficiency, n*—the degree of skill with which a person can use a language for communicative purposes.

3.1.25.1 *Discussion*—Language proficiency encompasses a person's ability to read, write, speak, or understand a language and can be contrasted with language achievement, which describes language ability as a result of learning. Proficiency may be measured through the use of a proficiency test.

3.1.26 *operational validity, n*—the extent to which item tasks, items, or interviewers on a test perform as intended and function to create an accurate score in a real world setting, as opposed to a setting involving an experiment, a simulation or training.

3.1.27 *performance test, n*—a test in which the ability of candidates to perform particular tasks, usually associated with job or study requirements, is assessed using "real-life" performance requirements as a criterion.

3.1.28 *polytomous scoring, n*—a model for scoring an item using a scale of at least three points.

3.1.28.1 *Discussion*—Using a polytomous scoring model, for example, the answer to a question can be assigned 0, 1, or

2 points. Open-ended questions are often scored polytomously. Also referred to as scalar or polychotomous scoring. Compare to *dichotomous scoring*.

3.1.29 *predictive validity, n*—the degree to which a test accurately and reliably predicts future performance in the domain being tested.

3.1.30 *protocol, n*—a standardized method or procedure for executing a given task, often formalized in documents.

3.1.31 *quality assurance, v*—the process of ensuring that the test planning and development phases are executed properly and satisfy the needs of all stakeholders.

3.1.31.1 *Discussion*—Quality assurance (QA) applies *(1)* when a new test is being created, *(2)* when a test that already exists is being repurposed or revised, *(3)* during certain aspects of the implementation process of the test (that is, replenishment of test items), *(4)* during item replenishment to ensure that new test items and prompts that will be used in the test conform to the original specifications that were used in creating the original items of that type, and *(5)* to train new personnel to administer the test to the same standards that were specified for the first testing personnel.

3.1.32 *quality control, v*—the system of post-development evaluations used at and after product acceptance to determine whether the test and testing practices used by an organization continue to meet and adhere to all standards and relevant testing policies.

3.1.32.1 *Discussion*—Quality control (QC) is used at and any time after product acceptance. QC verifies the continued validity and reliability of the test and shows the test is being used in an appropriate manner on an ongoing basis. Quality control (QC) is part of the test maintenance process.

3.1.33 *rater, n*—a suitably qualified and trained person who assigns a rating to a test taker's performance based on a judgment usually involving the matching of features of the performance to descriptors on a rating scale.

3.1.34 *rating, v*—to exercise judgment about an examinee's performance on a given task.

3.1.35 *rating scale, n*—a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged.

3.1.36 *reliability, n*—the consistency of a test in measuring what it is intended to measure across the life of the test or the degree to which an instrument measures the same way each time used; reproducibility.

3.1.36.1 *Discussion*—Consistency is the essential notion of classical reliability. Reliability is defined as the extent that separate measurements (for example, items, scales, test administrations, and interviews) yield comparable results under the same or similar conditions. For example, test items measuring the same construct should yield similar results when administered to same group of test-takers under comparable testing situations. Simply put, reliability is the extent to which an item, scale, procedure, or test will yield the same value when administered under similar or dissimilar conditions.

3.1.37 *scoring rubric, n*—a standardized method or procedure used by a rater in assigning a score to an examinee's performance on a given task.

3.1.37.1 *Discussion*—A scoring rubric is a detailed document that is used by trained raters to assess test taker performance. Correct interpretation and application of the scoring rubric requires training.

3.1.38 *selected response, adj*—any item which requires the examinee to choose between response options which are provided to the examinee, including, but not limited to true/false and multiple-choice items.

3.1.39 *skill modality, n*—any one of the four receptive and productive language skills of listening, reading, speaking, writing as defined in the ILR.

3.1.40 *specifications, n*—a detailed description of the characteristics of a test, including what is tested, how it is tested, details such as number and length of papers, item types used, etc.

3.1.41 *task, n*—an activity performed by a test taker in order to demonstrate functions and other proficiency criteria stated in the ILR Skill Level Descriptors.

3.1.42 *test-retest reliability, n*—an estimate of the reliability of a test as determined by the extent to which a test gives the same results if it is administered at two different times under the same conditions with the same group of test takers.

3.1.42.1 *Discussion*—Test-retest reliability is estimated from the coefficient of correlation that is obtained from the two administrations of the test. An assessment should provide a stable measurement of a construct across multiple administrations, especially when the time interval in between the administrations limits the potential for the amount of the underlying proficiency to change. There are three components of the test-retest reliability method: *(1)* two measurements with the instrument at two separate times for each test taker; *(2)* computation of a correlation between the two separate measurements; and *(3)* assumption that no change has occurred in the underlying trait or construct.

3.1.43 *translation, n*—process comprising the creation of a written target text based on a source text in such a way that the content and in many cases, the form of the two texts, can be considered to be equivalent.

3.1.44 *validity, n*—the degree to which a test measures what it is intended to measure, or can be used successfully for the purpose for which it is intended.

3.1.44.1 *Discussion*—Validity is a judgment of the degree to which the evidence (arguments) supports the conclusions, interpretations, uses and inferences of test scores.[4] A validity argument demonstrates the appropriateness and defensibility of a test's conclusions, interpretations, and inferences for a specific use in a given situation. The validity argument is based on the fact that a test is developed for specific uses and users and includes, but is not limited to, a description of and justification for test uses, impacts, audiences, and content. A number of different statistical procedures can be applied to a test to estimate its validity. Such procedures generally seek to determine what the test measures, and how well it does so. The

[4] Cook, T. D. and Campbell, D. T., *Quasi-Experimentation: Design and Analysis for Field Settings*, Rand McNally, Chicago, Illinois, 1979.

rigor and strength of the validity argument should increase as the stakes associated with the test (consequences for the individual and/or organization) increase.

4. Significance and Use

4.1 *Intended Use:*

4.1.1 This practice is intended to serve the language test developer, test provider, and language test user communities in their ability to provide useful, timely, reliable, and reproducible tests of language proficiency for general communication purposes. This practice expands the testing capacity of the United States by leveraging commercial and existing government test development and delivery capability through standardization of these processes. This practice is intended to be used by contract officers, program managers, supervisors, managers, and commanders. It is also intended to be used by test developers, those who select and evaluate tests, and users of test scores.

4.1.2 Furthermore, the intent of this practice is to encourage the use of expert teams to assist contracting officers, contracting officer representatives, test developers, and contractors/vendors in meeting the testing needs being addressed. Users of this practice are encouraged to focus on meeting testing needs and not to interpret this practice as limiting innovation in any way.

4.2 *Compliance with the Practice:*

4.2.1 Compliance with this practice requires adherence to all sections of this practice. Exceptions are allowed only in specific cases in which a particular section of this practice does not apply to the type or intended use of a test. Exceptions shall be documented and justified to the satisfaction of the customer. Nothing in this practice should be construed as contradicting existing federal and state laws nor allowing for deviation from established U.S. Government policies on testing.

5. Overarching Considerations

5.1 The purpose of a test is to provide useful information about examinees or programs. To build a useful test, developers and stakeholders must participate in an ongoing development and evaluation process, shown in Fig. 1 as the life cycle of a test and described further in Sections 6 – 10. Along with the processes of the life cycle, there are several interconnected elements that contribute to the usefulness of the information. These are validity (5.3), reliability (5.4), practicality (5.5), quality assurance (5.6), quality control (5.7), technical documentation (5.8), and ethics (5.9). This section provides general considerations about the life cycle and the elements as an overview, with Sections 6 – 10 providing more specific information about each phase of the life cycle.

5.2 *Test Life Cycle*—See Fig. 1.

5.2.1 The test life cycle is an iterative process, with new test development beginning with the plan for the test (to include a needs assessment, the creation of test framework and test specification documentation, followed by a plan for test maintenance). Test planning is described in Section 6. Following the acceptance of the planning stage, test development occurs (see Section 7). During this phase, qualifications are established and development teams hired, items are developed, scoring and rating is outlined, and validity evidence is collected. When the stakeholders agree that the test meets the expected standards, the test is accepted (see Section 8).

FIG. 1 Test Life Cycle

5.2.2 The test life cycle continues with test administration, ensuring standards for delivery, proctoring, scoring and rating, reporting of scores, and arbitration are met (see Section 9). The next stage in the test life cycle is test maintenance, which includes refreshment of test content (see Section 10). During this phase, new items are written and validated and testing documentation is updated to reflect current realities. When the test is determined to no longer meet the needs of the organization, it is retired.

5.3 *Validity:*

5.3.1 The validity argument begins at test creation and continues throughout the life of the test. The validity argument integrates multiple sources of data and brings elements from each stage of the life cycle as evidence for the goodness of fit between the test and its intended purpose. This is particularly important when a test has been developed for a specific use or audience and an organization wishes to use it for a different purpose or audience. When any test is developed, a test framework shall include an explanation of how the validity evidence will be gathered. As any part of the test use—such as the audience, purpose, administration, scoring or content—changes, the original test validity argument shall be replaced with a new or supplemental argument. The rigor of the validity argument should be sufficient to justify the consequences of the use of its scores or ratings, such that as the stakes to test takers and organizations increase, the rigor and strength of the validity argument should increase.

5.4 *Reliability:*

5.4.1 Without consistency and stability of measurement as indicated by reliability, decisions made from test scores or ratings are biased or potentially erroneous. Items, tests, raters, and procedures shall yield reliable measurements and have psychometric merit to be a useful basis for judgments or

inferences of knowledge, skill, or proficiency. Data that are unreliable are, by definition, unduly affected by error, and decisions based upon such data are likely to be quite tenuous at best and completely erroneous at worst. As the stakes of the test increase, reliability shall be more rigorously assessed. When any test is developed, a test framework shall include an explanation of how the reliability will be ensured. Although validity is considered the most important psychometric measurement property, the validity of an assessment is undermined if the construct or content domain cannot be measured accurately or consistently.

5.5 *Practicality:*

5.5.1 Practicality underlies the entire life cycle, as it is the extent to which appropriate resources are available for test development, operations, administration, and ongoing improvement. Necessary resources include:

5.5.1.1 Personnel to develop, administer, rate, score, report results, ensure security, and provide ongoing improvement;

5.5.1.2 Funds to develop the test, pay raters and administrators, support ongoing improvements, and manage test operations and security; and

5.5.1.3 Materials, including paper-based test booklets, scoring systems, tape recorders, and computers or computer software necessary for test administration, operations, scoring, security assurance, and ongoing improvement.

5.6 *Quality Assurance (QA):*

5.6.1 The application of QA to the creation of a new language proficiency test requires that a needs assessment be undertaken and executed correctly, and that input is received from all stakeholder groups. The needs assessment document is the first in a series of documents that guide the subsequent steps in the planning and development phases.

5.6.2 QA does not end when the test is created. Documentation that those original standards are being applied to new item creation and training shall be created during the process of new item creation or training.

5.7 *Quality Control (QC)*—Quality control is an essential component of the test maintenance process since it verifies the continued validity and reliability of the test and shows the test is being used in an appropriate manner on an ongoing basis. Documentation that supports the validity and reliability of the test and that the original standards and other relevant testing policies continue to be fulfilled shall be created and/or collected during quality control evaluations.

5.8 *Technical Documentation:*

5.8.1 All tests shall include technical documentation that covers the test life cycle from initial planning and development through ongoing test use. The technical documentation shall include sufficient information and evidence to evaluate the appropriateness and rigor of the approach, process, methodology, findings, decisions, and deliverables as appropriate to each stage of the test life cycle.

5.8.2 The documentation of test protocols and procedures, such as the test administration manual or the test security instructions, shall be provided and shall include sufficient information for the intended audience to perform their roles and responsibilities. Documentation shall meet professional standards for presenting information and evidence as appropriate to the specific stage of the test life cycle. The documentation can be provided as a series of individual reports for each stage or as a single report for the entire life cycle.

5.8.3 Documentation shall be periodically updated and supplemented as the test is either modified or extended to additional uses, populations, or contexts. These updates can be provided as supplemental reports or updates to the original reports.

5.9 *Ethics:*

5.9.1 At the highest level, ethics is a form of QA and QC. Ethics encompasses both standards of practice and moral obligations. Unethical behavior, whether intentional or unintentional, can result in considerable harm and be very costly to the organizations and individuals affected. Unethical behavior negatively affects the quality of the information provided by the test and reflects poorly on organizations, casting the professionals who create, use, or rely on test data in a poor light. Furthermore, the perceived value of language tests depends upon ethical practice and decisions made on the basis of test scores assume ethical practice.

5.9.2 In the development and operationalization of a language test, contracting agencies, testing organizations, test developers, and test users have ethical responsibilities. It is the responsibility of these organizations and individuals to determine, communicate, and document any local responsibilities and obligations that may not be known to others involved in the development and administration of a test. In all phases of a testing project, it is the responsibility of all participants to consider the ethical implications of their own and other's actions.

5.9.3 In addition to the standards included in Section 6, other sections of this practice address ethical considerations in language testing, since practicing ethical behavior is a part of good testing practice. Several organizations[5] have created ethical codes of practice in educational measurement designed to safeguard the rights of test takers by focusing on professional test development practices that could negatively impact examinees. These documents can also serve as guides to ethical behavior in language testing.

5.9.4 *Publication and Distribution of Accurate Information*—Test information provided to testing organizations, test developers, test users, and test takers shall be true and accurate. It is unethical to knowingly misrepresent information about a test.

5.9.5 *Copyright and Proprietary Materials*—Authorization for reproduction and distribution of secure test materials shall follow procedures established during the development process. All authorized reproduction shall be documented. Test developers and testing organizations shall respect copyright laws. Test materials subject to copyright may include, but are not limited to, test forms, items, ancillary materials, answer sheets, scoring templates, and conversion tables.

[5] For example, International Language Testing Association, ILTA Code of Ethics (http://www.iltaonline.com/images/pdfs/ILTA_Code.pdf), and Joint Committee on Testing Practices, Code of Fair Testing Practices in Education (http://www.apa.org/science/programs/testing/fair-testing.pdf).

5.9.5.1 If required by law, test developers shall ensure copyright permissions are obtained for any materials used in the test.

5.9.5.2 When required by law, testing organizations shall obtain consent of the owner before reproducing copyrighted or proprietary test materials.

6. Test Planning

6.1 Test planning is a phase of the test life cycle that begins with resource planning (6.3) and needs analysis (6.4) and guides the production of a series of key documents including the product acceptance plan (6.5), the test framework (6.6), test specifications (6.7), the test maintenance plan (6.8), the test refreshment plan (6.9), and the test security plan (6.10). All of these documents shall be developed in accordance with 5.8 and shall be revisited throughout the life cycle of testing to ensure continued relevance.

6.2 The test planning documents are related and inform each other. The resource planning and test security documents will evolve as additional needs are brought to light through the other documents. The needs analysis document is the first in a series of documents that guide the subsequent steps in the planning and development phases. The needs analysis guides the creation of the framework document. These two documents together guide the creation of the test specifications document.

6.3 *Resource Planning*—Without resources, a test cannot be developed. Because there are so many components to planning, development, administration, maintenance, refreshment, and security, organizations that wish to have tests shall develop a plan for resource allocation. This plan will change as test planning and development progresses: for example, after the needs analysis is funded, it may reveal the need for a level of statistical analysis that was not foreseen. Nevertheless, beginning with a plan for the resources known to be needed at the time, as well as a plan for revisiting resource needs, is crucial for the ultimate success of the test project. The resource plan shall address, at a minimum:

6.3.1 Personnel to plan, develop, analyze, produce, administer, rate, report, maintain, refresh, and provide adequate security for the test;

6.3.2 Funds to provide infrastructure such as test item banks, computer-adaptive algorithms, test centers, and secure servers;

6.3.3 Materials for development, production, and security;

6.3.4 Contingency funds for security breaches; and

6.3.5 Mechanisms for revising resource allocation as new needs become apparent through the planning, development, and maintenance process.

6.4 *Needs Analysis*—An organization's development, commissioning, or selection of a language test shall be based on the language use needs of the personnel to be tested by the organization. The ultimate responsibility for determining and evaluating the suitability of a test for a particular use rests with the organization using the test, not with the organization that developed the test. To ensure that the test is appropriate for its intended use, the organization shall perform a needs analysis before developing, commissioning, or selecting any language test. Then, the findings can be compared with the scope, design, tasks, purpose, and Interagency Language Roundtable (ILR)[2] level(s) of any proposed test to determine the ability of that test to meet the organization's current assessment needs.

6.4.1 *Repurposing of Existing Tests*—If an existing test is proposed for use in a situation that was unanticipated by its original designers or developers, the organization proposing the repurposing of the test shall evaluate its suitability for use in the new situation. While the results of the original needs analysis may have been useful in determining the suitability of an existing test for its originally intended use, they might not be sufficient evidence to justify the use of that test in a situation for which it was not intended, especially if high-stakes decisions will be made.

6.4.2 *Scope of Input*—The needs analysis should include input from the wider community of potential users to maximize opportunities for coordination and minimize duplication of effort. By having a needs analysis done, the organization will be able to determine the degree of fit between the ILR scale and the language skills needs of potential examinees who use language skills in their work. The organization should also recognize that the degree of fit may vary by the type of job or position within the organization. Thus, no single test may fit all situations in which a test is needed. In some situations, a needs analysis may reveal that an ILR-based test is appropriate for the whole potential testing population. In other situations, a needs analysis may reveal that a performance test or a test of language for specific purposes would be more appropriate for at least some segments of the potential testing population.

6.4.3 *Results*—Whenever possible, the results of the needs analysis study shall be shared with the group responsible for developing or selecting the test. When it is not possible, it is incumbent on the organization that will use the test to use the results of the study to specify the desired language skills to be assessed.

6.4.4 *Intended Use*—The organization that will use the test also shall consider the type of decisions that will be made on the basis of the test scores. Scores used to make high-stakes decisions require the selection or development of a test with a high degree of reliability and validity. Thus, indirect measures of the desired skills might not be suitable without strong evidence to support their use.

6.4.5 *Minimum Requirements*—As a minimum requirement, the results of the needs analysis shall provide the organization that will develop or supply the test with the following information:

6.4.5.1 The language requirements of the organization(s) that will use the test (including if applicable, variants of scripts, fonts, accents, and dialects),

6.4.5.2 The ILR level(s) that are needed to fulfill the language proficiency requirements of the organization(s) that will use the test,

6.4.5.3 The type of decisions that will be made on the basis of test scores,

6.4.5.4 How many examinees will take the test,

6.4.5.5 How often each examinee will be tested, and

6.4.5.6 The facilities available or planned for testing.

6.4.5.7 The circumstances under which a documentation audit (see Section 10) may be requested, and by whom.

6.4.6 *Documentation*—Needs analysis shall be documented in accordance with 5.8.

6.5 *Product Acceptance Plan:*

6.5.1 For a test to be used operationally, it shall be accepted by the relevant stakeholders. The organization or organizations that will use the test and the test development organization together shall develop a product acceptance plan that reflects the needs of stakeholders and developers for the particular testing program. In some cases, the stakeholders will not be involved until final acceptance of the test; in others, they may need to see interim products, such as the framework document or the results of field testing, to feel comfortable accepting the final product. The product acceptance plan shall include, at a minimum:

6.5.1.1 A list of the points in the planning and development process at which stakeholder acceptance is required (for example, the stakeholders might want to approve the framework document or the categories of people who can be examinees for field testing);

6.5.1.2 A list of the documents representing those points that the stakeholders will receive for approval (for example, the framework document, a list of examinees, and statistical reports on item quality);

6.5.1.3 A timeframe for acceptance (when the test developer shall submit materials to stakeholders and when stakeholders shall finalize their acceptance decision for each stage); and

6.5.1.4 A set of criteria by which stakeholders will judge acceptability (for example, they require the framework document to be readily understood by non-specialists).

6.5.2 As the planning, development, maintenance, and refreshment of a test progresses, the needs and priorities of the stakeholders may change, and it is legitimate to revise the list of points of acceptance and criteria for acceptance; however, these revisions shall be documented and agreed to by all involved, so that the acceptance process remains transparent and consistent across the testing program. Any agreed-upon revisions shall be fully funded and shall include appropriate revisions to project timelines and deliverable schedules.

6.6 *Framework Document:*

6.6.1 *Purpose*—A framework is an essential document that provides the rationale for the test design. It is the bridge between the needs analysis and the test specifications. It justifies and explains test design decisions. A framework document is useful for clarifying consequences of test use and providing an underpinning for test specifications. The more important the consequences of decisions based on the test scores, the more important it is for the framework document to be comprehensive and explicit. For ILR-based tests in particular, it is important to make clear the interpretation of the ILR and the aspects of the ILR that are considered important for the construct of the particular test in question. The framework document can then be used as a basis for making decisions about what new research needs to be conducted to justify using the test for different populations or using the test scores in a new way. The framework document shall be developed in accordance with 5.8. See 6.6.3 for more specific guidance.

6.6.2 *Process*—Test developers shall develop a framework document in close coordination with test users and other relevant stakeholders with input from outside testing experts as needed. At the beginning of a testing project, test developers shall inform stakeholders of the usefulness of a framework document and request that such a document be created before test development begins. In the event that stakeholders reject the request, test developers shall develop the framework document concurrently with the test specifications and the test items. The document should be updated in accordance with 5.8 as new research is conducted or new issues concerning test use arise. For existing tests that are being adopted for the testing of ILR-based proficiency, the organization that will use the test is responsible for creating a framework document, with the cooperation of the original developers if possible, preferably before the test begins to be used.

6.6.3 *Content*—The framework document shall contain the following:

6.6.3.1 The decisions to be made on the basis of test scores (for example, hiring, placement, and retention);

6.6.3.2 The intended consequences of test use (for example, eligibility for training courses, reassignment of personnel, or determination of operational readiness);

6.6.3.3 An interpretation of the relevant sections of the ILR skill level descriptions and how they are to be operationalized (for example, taking the phrase "speakers can make themselves understood to native speakers who are in regular contact with foreigners" and defining or exemplifying who those native speakers are and how this characteristic is assessed in the test);

6.6.3.4 An interpretation of the relevant sections of the ILR skill level descriptions and how they are to be operationalized (for example, taking the phrase "speakers can make themselves understood to native speakers who are in regular contact with foreigners" and defining or exemplifying who those native speakers are and how this characteristic is assessed in the test);

6.6.3.5 A justification of the links between test scores and their interpretations, uses, and consequences; and

6.6.3.6 An explanation of the research that has been done to support the links above and identification of areas in which more research is needed. This section would likely change as the test is used. Before the test is developed, research would presumably focus on previous types of tests, with a discussion of how the current test is similar or different, and this section would primarily outline predictive or concurrent validity studies that are planned for the test. Once the test is operational, the results of those validity studies would be incorporated. Any updates to the framework document shall be in accordance with 5.8.

6.7 *Test Specifications Document*—The test specifications is an essential document that provides detailed specifications regarding the construct, design, content, administration, scoring, reporting, and intended use of the test. The test specifications shall be sufficiently detailed to guide the day-to-day work of test development and serve as a standard against which the completeness of that work can be measured. The

more important the consequences of decisions based on the test scores, the more important it is for the test specifications document to be comprehensive and explicit. For existing tests that are being used for new purposes, the organizations using the test are not responsible for obtaining or generating specifications for test design (6.7.5). The other sections of the specifications shall be obtained from the original test designers or written by the organization using the test to reflect the intended use, scoring or rating, reporting, and administration requirements of the test in its new use. The test specifications document shall be developed in accordance with 5.8.

6.7.1 *Intended Test Use*—The specifications shall clearly state that the purpose of the test is to measure general proficiency as defined by the ILR scale. The skill domain(s) covered by the test (listening, reading, speaking, or writing) shall be specified, as shall the range of ILR levels.

6.7.2 *Construct Definition*—The specifications shall clearly define the construct(s) to be measured with specific reference to the ILR skill level descriptions.

6.7.3 *Intended Score Use(s)*—The intended score use(s) and limitations in the application or interpretation of scores shall be clearly stated. The consequences of decisions based on test scores shall be clearly stated.

6.7.4 *Intended Test Taker Population*—The specifications shall describe the intended test taker population for the test. If the population is diverse, the specifications should indicate how the diversity of the population is taken into account in the test design and how it is taken into account in the way that items are written or tasks constructed or both.

6.7.5 *Test Design:*

6.7.5.1 Test design specifications shall include a general description of the test format (for example, interactive oral interview, non-interactive oral presentation, passage-based interview, selected response, constructed response) and the delivery model (for example, fixed-form, computer-adaptive, human-adaptive), as well as detailed specifications for item types, content coverage, and test form composition. Item and test form specifications shall take test security into account by emphasizing item types and test form compositions that discourage memorization and cheating.

6.7.5.2 Item specifications shall include a general description of each item type in the test, along with a detailed description of scoring attributes (for example, dichotomous, polytomous, partial credit), prompt attributes (what the examinee will encounter, including the directions for taking the test and responding to the items), response attributes (what the examinee is expected to do in response to the prompt and what will constitute failure or success), scoring rubrics or protocols or both, and a sample item for each item type, including sample response attributes and sample rubrics/protocols, if applicable.

6.7.5.3 Content specifications shall describe guidelines for content coverage and balance.

6.7.5.4 Test form specifications shall provide specific guidelines for test form construction, including number of items per passage, stage, and level (as applicable).

6.7.5.5 Test form specifications shall include guidelines for the development of tasks to ensure that such tasks are developed in a standard and replicable manner.

6.7.5.6 Specifications for adaptive tests shall include decision-tree guidelines or rubrics or both for human testers or adaptive algorithms for computer-adaptive tests.

6.7.6 *Scoring, Rating, and Reporting:*

6.7.6.1 Scoring specifications shall explain in detail how both raw and scaled scores are generated (as applicable) and how cut scores are set and interpreted.

6.7.6.2 Partial credit scoring models and criteria for evaluating and rating constructed responses by human raters shall be described in detail (as applicable).

6.7.6.3 Rating specification shall include explanations for how raters are trained and the rating scale being used for rating.

6.7.6.4 Reporting specifications shall describe how test scores and ratings are reported to test takers, test users, and other stakeholders (as applicable).

6.7.7 *Administration and Technological Requirements:*

6.7.7.1 The test specifications shall describe standard test administration conditions and procedures. The descriptions should include required training and qualification information for any test administration personnel and any materials or technology needed to administer the test under standard conditions. If these descriptions are particularly complex, they should be described, in detail, in a separate document and the document referenced in the test specifications. Examples of administration and technological requirements include, but are not limited to, the following (see 9.2 for specific requirements):

(1) The physical testing environment or setting;

(2) Time allotted to test administration;

(3) Test administration personnel, including any training and qualification requirements;

(4) Documents, materials, and tools required by test takers or test administrators, including printing and binding requirements of any published materials; and

(5) Hardware and software, including version, bandwidth, and security requirements.

6.7.7.2 The test specifications shall describe circumstances under which the standard test administration procedures may be modified and the extent to which they may be modified without affecting the validity and reliability of the test.

6.7.7.3 If technology is used, the specifications shall describe how the technology interfaces with the specifications. When there is an interface between the technology to be used and the types of items that will be written, then this shall be indicated in the specifications.

6.8 *Test Maintenance Plan*—Maintenance means ensuring and documenting that the test remains valid and reliable. Organizations planning to use a test shall have a plan for ensuring that the test continues to provide useful information. The test maintenance plan shall be developed in accordance with 5.8. This plan shall include the following elements:

6.8.1 A list of the documents comprising reliability and validity evidence that will be maintained in anticipation of reviews and audits;

6.8.2 Specifications for how test performance will be evaluated (impact data, item performance data, test and rater reliability data, conformity to specifications, and so forth);

6.8.3 A list of the processes that will be used to review the items and test, conduct statistical analyses of operational items and tests, retrain raters, and recertify raters;

6.8.4 A specification of how often each of these processes will be performed over the life cycle of the test;

6.8.5 The metrics used to determine item or test life cycle or both: exposure to a certain number of examinees, time elapsed, or some combination. The metrics shall take test security into account by acknowledging the value of limiting exposure rates;

6.8.6 A recommendation for what is to be done with the results of the maintenance review; and

6.8.7 An estimate of the resources (money, contracts, and personnel) needed to perform test maintenance.

6.9 *Test Refreshment Plan*—An anticipated outcome of a test maintenance review is that test content (items, training materials, and scoring and rating protocols) will need to be replaced. The organization planning the test shall have a test refreshment plan. The test refreshment plan shall be developed in accordance with 5.8. This plan shall include the following:

6.9.1 A specification of the circumstances under which changes will be allowed and those under which changes will be mandatory, for example, exposure to a certain number of examinees, time elapsed, amount of change in item statistics, impact data outside of a particular range of what was expected, slippage in ILR level, and unfavorable review of materials;

6.9.2 A specification of the mechanisms for refreshment, for example, whether whole forms will be replaced or a certain percent of items will be replaced and how new cut scores will be generated following refreshment of items;

6.9.3 A specification of the circumstances under which cut scores may be changed in the absence of changes to the composition of the test;

6.9.4 If the test uses testers or raters or both, a specification of how much change in the tester/rater pool is allowable, for example, whether it is acceptable to retire all testers/raters from the pool and replace them with new raters at once or whether a core of existing testers/raters needs to continue as new testers/raters are brought on;

6.9.5 A specification of the statistical requirements for inclusion of new items in the test, for example, whether they need to be calibrated on the same scale as existing items before being inserted in the operational test; and

6.9.6 A specification of if and how new items are to acquire statistical information, for example, by being administered but not scored, administered in a separate testing session, or have item parameters estimated based on item content characteristics.

6.10 *Test Security Plan*—Test security encompasses all areas of test development, production, administration, scoring, rating, and reporting. In the test planning stage, a test security plan shall be developed to ensure that, from the very beginning, resources are allocated and good test security practices are followed. In section 6.10.1, the requirements for the overall test security plan are outlined; in 6.10.2, the security breach contingency plan as a separate document is addressed. See Appendix X3 for additional information about test security plans. The test security plan shall be developed and maintained in accordance with 5.8.

6.10.1 *Overall Test Security Plan:*

6.10.1.1 The test security plan shall include, at a minimum, the following:

(1) A description of the roles and responsibilities of personnel required to ensure security;

(2) A list of the test security documents that will be generated or appropriated for the test, to include instructions for development personnel, test security nondisclosure forms, instructions for proctors, examinees, and raters, and policy statements;

(3) A description of the methods to be used to train personnel on test security;

(4) A list of physical and electronic security requirements;

(5) A description of the methods to be used for monitoring for compliance with security policies; and

(6) A security breach contingency plan.

6.10.1.2 Many of the components of the security plan are described elsewhere in this practice (see, in particular, 7.9 and Section 9). Because the security breach contingency plan is primarily a planning document, it is described in more detail in 6.10.2.

6.10.2 *Security Breach Contingency Plan*—Aside from routine maintenance and refreshment, there may be a need for changes to a test arising from a security breach. The organization using the test shall document a plan for actions in response to specific types of security breaches. The plan shall identify the different types of security breach that might arise (for example, the loss of an answer key, the posting of an item on a student website, the theft of a scoring protocol) and, for each type of breach, specify what changes to the test, if any, will result (for example, reordering of answer choices, removal of an item and recalculation of cut scores, replacement of an item, withdrawal of a test form). The plan shall also specify whether the test developers shall develop enough extra items or forms to hold in reserve so that compromised tests can be immediately replaced or whether item replacement as a result of compromise will take place on an ad hoc basis.

7. Test Development

7.1 Test development is guided by the test purpose and intended use as documented in Section 6. In this phase, qualifications of test development teams are addressed (7.2), a test administration manual (7.3.1) is created, and test specifications are implemented through item development (7.4) and scoring and rating (7.5). Best practices are outlined for item analysis (7.6), form comparability (7.7), and cut score setting (7.8). Wrapping up this section is a discussion of test security (7.9).

7.2 *Qualifications of Developers and Reviewers*—The test development process shall rely on qualified personnel who work together in teams as appropriate. This section addresses qualifications (7.2.1.1 and 7.2.2) and training (7.2.3) of these personnel.

7.2.1 *Test Development Teams:*

7.2.1.1 Language test developers shall compose a team of experts in the following four areas:

(1) Language testing experts knowledgeable in the theory of testing who can ensure that specifications are met and who

possess a thorough understanding of the entire test development process and life cycle;

(2) Language experts who can ensure that content is accurate and appropriate;

(3) Psychometric experts who can ensure that items are functioning properly; and

(4) Item writers who understand how to elicit useful examinee responses.

7.2.1.2 The team shall also include programming and software expertise as required by the specifications document. It is essential to have members with expertise in all four language-testing areas, though a single member may qualify in more than one area. Two types of reviewers are needed: one with language expertise and the other with psychometric expertise. A reviewer may have both types of expertise. All team members shall have language proficiency in the working language of the team that would allow them to communicate efficiently and effectively with the other members of the test development team.

7.2.2 *Preferred Qualifications:*

7.2.2.1 Testing experts shall have qualifications encompassing many aspects of testing so that they can reasonably supervise the construction of a test. Examples of relevant qualifications include:

(1) A masters degree or higher in a relevant field (for example, language testing, applied linguistics),

(2) At least three years experience working as a testing expert on language test development projects of a similar scale, and

(3) Published papers on test theory or practices in a peer-reviewed publication.

7.2.2.2 The language expert's qualifications may include:

(1) Proficiency in the target language that is equal to or higher than the maximum ILR being assessed in the test in the relevant skill(s) and

(2) Training in the linguistic aspects of the target language.

7.2.2.3 The psychometric expert's qualifications may include:

(1) A masters degree or higher in a relevant field (for example, statistics, educational measurement),

(2) At least three years experience working on psychometric aspects of language test development projects of a similar scale, and

(3) Published papers on statistical measures and analyses in a peer-reviewed publication.

7.2.2.4 The item writer's qualifications may include:

(1) Experience or training or both in language test item development.

7.2.3 *Training:*

7.2.3.1 Test development team members shall undergo training on the test project, including all areas of the needs analysis, framework, and test specifications documents.

7.2.3.2 Test development team members should also familiarize each other with concepts from their own specialized areas relevant to the project, such as requirements of a specific type of item development (for example, multiple-choice items, cloze items, and essay items), issues particular to the language(s) involved, and psychometric constraints and limitations. The areas to be covered in training team members shall include:

(1) Relevant language testing principles,

(2) ILR skill level descriptions,

(3) Passage selection and development,

(4) Item development,

(5) Elicitation techniques,

(6) Evaluation processes, and

(7) Test security.

7.2.3.3 Training shall include a combination of theory, review, discussion of previously administered tests, and practice using unofficial tests.

7.3 *Supporting Materials*—Test developers shall produce materials to support the test, including a test administration manual (7.3.1), training materials (7.3.2), and scoring and rating information (7.3.3 and 7.3.4).

7.3.1 *Test Administration Manual*—The test developer shall provide a test administration manual in accordance with 5.8 describing the mechanics of delivering the test, including an outline of the process, and an explanation of the scoring rubrics and rating forms needed to administer and rate the test. The manual shall address the following:

7.3.1.1 Method of delivery (electronic, paper and pencil, and so forth),

7.3.1.2 Timing of the test,

7.3.1.3 Proctoring needs,

7.3.1.4 Personnel or technology or both involved (if technology enhanced or technology based proctoring is used),

7.3.1.5 Security features,

7.3.1.6 Method of determining score,

7.3.1.7 Score adjudication,

7.3.1.8 Method of delivering the score to sponsor/examinee, and

7.3.1.9 Appeal process.

7.3.2 *Training Materials*—The test developer shall produce materials that clearly describe the training process and the evaluation criteria required for proper administration and scoring/rating for the test. The test developer shall specify the following:

7.3.2.1 Training materials,

7.3.2.2 Duration of training,

7.3.2.3 Type of delivery (face-to-face training, online training),

7.3.2.4 Criteria that constitute satisfactory completion of training, and

7.3.2.5 Other training outcomes.

7.3.3 *Scoring Rubrics*—The test developer shall provide a scoring rubric for determining the conversion of raw test data into meaningful scores.

7.3.3.1 *Scope and Content*—The scoring rubric should have enough breadth and depth to permit the rater to obtain sufficient information to assign a score. The scoring rubric shall specify levels of performance/proficiency and factors to be rated and provide detail on the characteristics of performance at each level. Descriptions of levels shall be clearly defined and operationalized.