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 Guide for Language Interpretation Services

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).

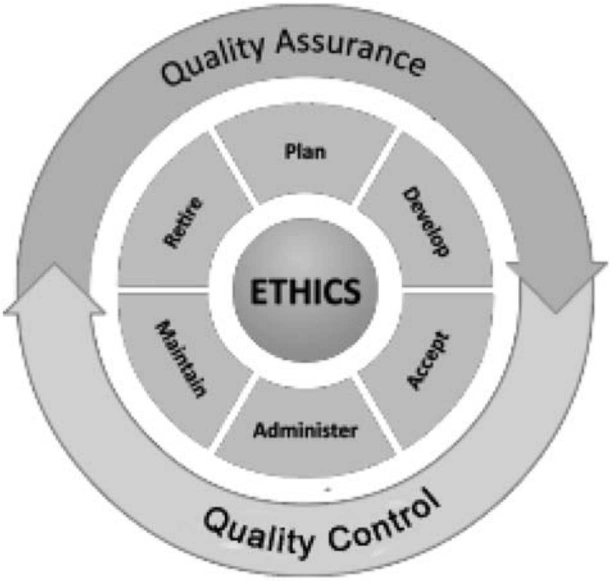

FIG. 1 Test Life Cycle

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.

7.3.3.2 *Alignment*—The scoring rubric shall provide information on how the scores align with the ILR skill level descriptions for the relevant skill(s) (listening, speaking, reading, writing).

7.3.4 *Rater Forms/Reviewer Forms:*

7.3.4.1 The test developer shall provide forms for raters of the test and reviewers of the raters to use.

7.3.4.2 The forms shall contain categories pertinent to determining a valid score on the test.

7.3.4.3 The forms shall be representative of the complexity of the test and cover all categories required for effective scoring according to the ILR scale.

7.4 *Item Development:*

7.4.1 Generating tasks to which the examinee responds is called item development. Item development includes (but is not limited to) multiple-choice items; prompts, scoring rubrics, and/or expected responses for constructed-response items; and tasks eliciting speaking or writing samples. Development of all item and task types requires strict adherence to test specifications and rigorous review processes.

7.4.2 Test developers, whether writing items, prompts, or tasks, shall adhere to the test specifications and shall submit all components of each item or prompt, including the expected response, to a rigorous review process.

7.4.3 *Item Development Teams*—Test developers shall ensure the participation of two groups of experts in item development teams: target-language experts, whose input as content specialists is critical to the development of test items that measure real-world language proficiency and professional item writers, whose expertise in item writing and in the ILR skill level descriptions is critical to the development of items that align to the ILR standards. The team shall include technical expertise as required by the specifications document. In the ideal case, the item development team has professional item writers who are also target-language experts. If this is not possible, it is highly beneficial to have the target-language experts and the item writers working together on a team, so that they may inform each other's work directly. Also acceptable is a model in which the two groups are not working together as a team, but have access to each other for consultation. Test developers shall demonstrate that they have used an approach to item development that allows these two groups of experts to contribute equally to the development of the highest quality test items possible.

7.4.4 *Item Development Process*—The test developer shall follow a rigorous and documented process for item development, QC, and QA. Items shall be written in accordance with the test specifications (see 6.7). Where required by the test specifications, scoring rubrics shall be developed in conjunction with the items. Although there are differences among types of items (see 7.4.5), the item development and review process for all items shall encompass at least four stages as described in the following.

7.4.4.1 *Text Typology*—For receptive skills, authentic texts shall be selected and analyzed by a team of experts using a text typology protocol such as Child's classification of text modes.[6] All texts shall be analyzed and rated based on a number of text-linguistic features such as mode, genre, text source, topic, syntactic structures, discourse features, language functions, and the ILR skill level required to comprehend the core of the text. For productive skills, expectations about the responses to each task or its accompanying prompt or both shall be documented and linked to the ILR skill level descriptions specifying expectations about mode, genre, topic, syntactic structures, discourse features, and language functions.

7.4.4.2 *QA*—A documented peer-review process shall require both item writers and target language experts to review items produced by other writers. (For production of dynamic tasks, see 7.4.5.1.) Each draft item will be reviewed using an established procedure to check for such things as accuracy, clarity, consistency, and conformity to item specifications. There shall be a standard process for dealing with items that are problematic, for example, tester retraining or a rewriting process for prepared items. QA documentation shall indicate that the following considerations have been taken into account where applicable for the type of items (see section 7.4.5 for specifications of which considerations are applicable to which types of items):

(1) The interaction among complexity of task, expected response, stimulus material, ILR level, and difficulty of the item;

(2) Whether what constitutes a successful response is made clear to the examinee (for example, whether examinees are to deliver a formal speech or an information briefing, how much detail is required in the response, and whether the response requires a literal translation or an inference);

(3) Whether the indicated correct response(s) is/are the only correct response(s) from among the answer choices;

(4) How accurate and complete an examinee's response shall be, relative to what is in the scoring rubric to receive credit;

(5) Whether the range of possible acceptable answers have been accounted for in the protocol/rubric; and

(6) In a dynamic test, the implications of the response for future items in the test: if the examinee comes close to completing the task successfully is the tester to provide another task at the same level, a task at a lower level, or repeat the same task type?

7.4.4.3 *Quality Control for Items*—The test developer shall document and put in place a plan to ensure ongoing scrutiny of items. For example, testers might be required or invited to fill out comment forms on their perceptions of the effectiveness of the items they used, or routine analysis might be conducted of the tasks selected for tests and any correlations with those tasks and level ratings or tester behavior.

7.4.4.4 *Communication*—The test developer shall put in place a system for communicating issues with particular items

[6] Child, J., "Language proficiency and the typology of texts," in *Defining and Developing Proficiency: Guidelines, Implementation, and Concepts*, H. Byrnes and M. Canale, Eds., National Textbook Co., Lincolnwood, IL, 1987.

or item types so that all involved have access to the same information about the items they are using.

7.4.5 *Requirements for Specific Item Types*—The following are the basic types of items and which considerations in 7.4.4.2 apply to them.

7.4.5.1 *Dynamic Items*—Dynamic items are given in an interactive environment, such as on an interview test. Dynamic items are of two types: (1) prepared items, which are developed before the test is administered and delivered with a predetermined wording (for example, some role-play situation descriptions) and (2) extemporaneous items, which are based on guidelines developed for tester training but are produced by the tester during the test (for example, prompts eliciting past narration based on information provided by the examinee, or follow-up questions). Development and review of extemporaneous items shall focus on the guidelines and training given to testers and on evaluation of tester performance. Tester training and maintenance shall include practice sessions in which testers demonstrate their ability to create extemporaneous items and they can critique each other's items. Relevant considerations from 7.4.4.2 for prepared items are items (1) to (5) and relevant considerations for extemporaneous items are (1), (2), and (6).

7.4.5.2 *Static Items*—Static items are given in a non-interactive environment, for example, on a traditional multiple-choice receptive skills test or on an asynchronously rated productive skills test. The quality of the information provided by the items depends on having the items perform consistently from one examinee to another, assuming the examinees have equal proficiency, and to focus only on the relevant skill. Within the category of static items, there are two relevant subtypes: selected response and constructed response items. Selected response static items include multiple-choice items, drag-and-drop items, matching items, sorting items, and any item type in which the range of possible answers is completely constrained. Selected response items can have one correct answer, multiple correct answers, or a partial-credit model in which some answers are given more weight than others. Relevant considerations from 7.4.4.2 for selected-response items are (1), (2), and (3), and relevant considerations for constructed-response items are (1), (2), (4), and (5).

7.4.6 *Scoring Information*—Scoring information is a required part of the item, as is documentation of what distinguishes a successful from an unsuccessful response. Items include the definition and identification of the correct response.

7.4.7 *Documentation*—Each step in the above item development process, whether for productive or receptive skill test development projects, shall be documented in accordance with 5.8.

7.5 *Scoring and Rating:*

7.5.1 *Scoring:*

7.5.1.1 Scoring of items refers to the use of a scoring key to assign a value to an examinee response to a test item in which all acceptable responses are predetermined. The scorer or scoring computer program determines the credit merited by the response from a predetermined system and calculates the final score based on established formulae. Scoring is typically used for selected response items.

7.5.1.2 Scoring of tests refers to the aggregating of individual item values to a total test or subtest score. The test score can be reported as a raw score, a converted score, or equivalent based on score conversion tables.

7.5.1.3 Scoring does not require trained technical personnel. The process can be conducted either by a machine or a human, but in either case, requires little to no judgment by the scorer. Test developers/administrators shall ensure that training is delivered as necessary to ensure scoring procedures are applied consistently. See 9.5.1.

7.5.2 *Rating:*

7.5.2.1 Rating refers to the assignment of a value to an examinee response to a test item though a process that requires human judgment by a rater. In some cases, rating may be done by a specialized computer rating system that requires training on a set of correct and incorrect responses. In such cases, users of rating software shall check the consistency of computer rating against human raters to ensure adequate reliability. The remainder of this section will deal only with issues regarding human raters. In rating, the test response may be compared to a key, but the credit for the response is determined by the rater. The final rating may be a holistic evaluation, a cumulative score based on established formula, or a combination of the two. Rating is typically used when test items have open responses, such as on constructed response tests, essay tests, or oral proficiency interviews.

7.5.2.2 Rating requires trained technical personnel to make evaluations on the quality of test responses based on established criteria and expert judgment. To be able to make such expert judgments, raters may be required to have prerequisite skills, such as target language ability. Additionally, raters shall meet certain training and accreditation requirements before rating actual tests. Since rating relies at least in part on human judgment, it is advisable that multiple independent raters evaluate examinee responses before a final rating is given. The appropriate number of independent ratings for a given test should be determined using appropriate statistical analyses. Periodically, raters shall undergo quality control and quality assurance to ensure rater reliability. Such periodic retraining shall be required of all active raters.

7.5.2.3 Criteria for test rating shall be carefully selected, and valid for the test use. Raters are required to have an in-depth understanding of how to apply the criteria to the test performances. Raters shall have a clear understanding of how the test performances should be evaluated to determine the rating.

7.6 *Information about Item Analysis:*

7.6.1 *Item Analysis Process:*

7.6.1.1 Item analysis shall be conducted and documented in accordance with 5.8 to gather information pertaining to:

(1) Item difficulty and

(2) Item discrimination (to ensure that items can discriminate between more and less proficient examinees).

7.6.1.2 Depending on the skill modality being tested, the stakes of the test, the size of the testing population, and the availability of examinees during the development stage, different means of gathering information will be appropriate. Test developers shall use at least one of the methods listed in the following and ideally two to include elements of both quantitative and qualitative analysis:

(1) Cognitive laboratories,

(2) Expert panel,

(3) Statistical analysis, and

(4) Item calibration (for example, item response theory [IRT] calibration).

7.6.1.3 Test developers shall document the method used specifying details of how the method was carried out, including number of examinees or panelists or both, and providing all materials given to examinees or panelists or both for instruction.

7.6.2 *Results:*

7.6.2.1 Regardless of the method of obtaining information about item characteristics, developers shall document how that information was used to revise or discard items.

(1) Review—When possible, item content shall be reviewed in conjunction with the information gathered from examinees about item characteristics. Mismatches between what would be expected from content and what was seen from examinee information should be carefully considered.

(2) Analysis—An analysis report shall be produced listing all relevant statistics for each item, summarizing qualitative and quantitative information gathered from examinees, and commenting on any significant trends or concerns.

(3) Revision/Discarding—Based on the review and analysis, items shall be flagged for revision or discarding. General criteria for flagging should be documented. Information shall be obtained and documented about the item characteristics of revised items.

7.6.2.2 When research shows differential item functioning across age, gender, racial/ethnic, cultural, disability, linguistic groups, and/or other structural groups in the population of test-takers, the organizations using the tests shall consider whether this differential item functioning might introduce bias that would impact the functioning of the test and/or the interpretation of the scores.

7.7 *Form Comparability:*

7.7.1 *Procedures:*

7.7.1.1 *Fixed-Form Tests: Equating*—If there is to be more than one form of the test, the forms shall be equated and the test developer shall perform form-to-form reliability analyses. The test developer shall document the results of those analyses. The process, analysis, and results of form comparability shall be documented according to 5.8.

7.7.1.2 *Computer-Delivered Adaptive Tests*—For computer-delivered adaptive test designs, the test developer shall document the design of the adaptive component of the test, for example, whether it is item adaptive or a multistage adaptive test and what the criteria are for moving to a different difficulty level.

7.7.1.3 *Interactive Tests*—The test developer shall document the criteria for moving to a different difficulty level, as well as the design features that ensure comparability across examinees, testers, and raters. Procedures for checking comparability, for example, how and how often inter-rater reliability checks are performed and what is done if reliability slips below acceptable rates, shall be documented. Studies shall be conducted to show consistency of performance with the same examinee and different raters, as well as consistency of raters with respect to following the test design. All study results shall be documented.

7.7.1.4 *Small-Scale Procedures*—For tests with small examinee populations, it may not be possible to perform quantitative equating, form-to-form reliability analyses, or rater reliability analyses. In such cases, test developers will document the content, level, and task characteristics of each test and indicate what steps have been taken to ensure that the forms or tests are as parallel as possible.

7.7.2 *Documentation*—Test developers shall document the equating methods and the rationale for using those methods. Results of the equating shall also be documented, noting population size and characteristics as well as interpretations of any statistics used.

7.7.3 *Accommodations*—In some cases, it may be necessary to produce alternate forms of a test to accommodate examinee disabilities. If alternate forms are produced, the test developer shall document the nature of the accommodation, the rationale for the accommodation, and evidence that the accommodation produces a form equivalent to the base forms.

7.8 *Cut-Score Setting:*

7.8.1 *Background*—The ILR scales provide operational definitions of the construct of general, unrehearsed language proficiency in each of the four skill modalities: listening, speaking, reading, and writing. Each of those construct descriptions also establishes the construct that the ability levels it describes form a hierarchical array of levels in which each higher level of the scale describes a set of task, condition/context, and accuracy criteria that is more complex than the lower levels. When testing the receptive skills of listening and reading, the condition/context component is expanded to include author purpose, text types, and accuracy considerations. Because the ILR scale defines proficiency using a hierarchy of level-specific criteria, and the criteria for a given level must be satisfied before that level can be assigned, tests based on the ILR scales are criterion-referenced tests. Therefore, the cut-score setting processes used with ILR-based proficiency tests (as well as the test design and development steps leading up to cut-score setting) shall be based on information or judgments or both about the examinees' ability to perform relative to the ILR, rather than on information or judgments about examinees' ability to perform relative to each other. Cut-score setting processes shall also be informed by the standards laid out in Chapter 14, "Testing in Employment and Credentialing," in *Standards for Educational and Psychological Testing*.[7]

7.8.2 *Test Design Expectations:*

7.8.2.1 The documentation from the foregoing sections shall be taken into account in cut-score setting whether preparing scenarios for ILR oral proficiency interview (OPI) testing, prompts for writing proficiency (WP) tests, or test items to assess listening and reading comprehension. Test developers shall also document:

[7] American Educational Research Association, *Standards for Educational and Psychological Testing*, American Educational Research Association, Washington, DC, American Psychological Association, National Council on Measurement in Education, 1999.

(1) Where each test item fits into the table of design specifications;

(2) The criteria and process used to link each item to the content, communication task, and accuracy expectations associated with a specific level in the ILR scale;

(3) The statistical process used to confirm that each item is performing as designed; and

NOTE 1—The item statistics may be classical item statistics such as facility and discrimination indices or, preferably, IRT item characteristic curves or criterion-referenced B-index statistics.

7.8.2.2 These elements are required regardless of the type of test, although the specific information will vary with test type.

7.8.3 *Cut-Score Setting Procedures:*

7.8.3.1 General procedures for establishing and documenting test reliability and validity are described elsewhere in this practice. Note that, with criterion-referenced tests, test validation is inseparable from the process of setting cut scores. Unless a test is valid, accurate cut scores cannot be set; and unless it can be demonstrated that the cut scores are accurate, a test is not valid for the purpose for which it was intended. ILR test developers shall document the procedures used to establish criterion-referenced cut scores. In general, cut-score setting processes fall into one of three categories:

(1) Test-centered analyses (such as Angoff, modified Angoff, book marking, and so forth);

(2) Work-sample-centered analyses (such as body of work, borderline group, or similar analysis); and

(3) Comparison analyses (such as a contrasting groups approach or double testing of candidates using the new test and an established criterion measure). The criterion measure may be an accepted test with known properties for the examinee population, or a measure created especially for the purpose of setting cut scores for the new test. In either case, the validity of the criterion measure shall be documented as part of the cut-score-setting process.

7.8.3.2 For single-level tests, the criteria used to designate the passing score shall be explained. For multi-level tests, the test developer shall document the method used to establish cut scores for assigning each level and this shall be documented. This documentation shall stipulate the approach used to establish the cut scores and describe how that process was applied. This documentation shall be of sufficient detail that others can replicate the processes used and compare their results with the results reported by the original test developers.

7.8.4 *Validating the Consistency of the Test's Cut Scores:*

7.8.4.1 Once cut scores are established for a test, the consistency of those cut scores shall be statistically documented. For speaking and writing tests, the culminating statistical evidence used to validate any cut-score criteria used shall report the extent to which the prompts used generate ratable language samples from which a rater consistently assigns identical ratings on the ILR scale (intra-rater reliability) and the extent to which those ratings agree with ratings assigned by other independent raters (inter-rater reliability). Because relative or rank-order agreement is insufficient to demonstrate categorical agreement, these rater correlations should be calculated using intra-class rather than product-moment correlations.

7.8.4.2 For listening and reading tests, the consistency of the test's cut scores shall be documented in a report showing the statistical accuracy around each of the cut scores (using standard error of measurement calculations, cluster analyses, or related statistical procedures) when each cut score is applied to the target population. For ease of interpretation, this report shall also present the operational impact of implementing each cut score. This impact shall be presented in the form of an expectancy table, which shows the percentage of individuals for whom there is exact agreement between the test and the criterion assessment techniques, the percentage of false positives (individuals given a higher rating by the test than by the criterion measure), and the percentage of false negatives (individuals given lower ratings on the test than on the criterion measure). For cut-score-setting options that are not based on external criterion of comparison groups, a similar table can be constructed using confidence interval data.

7.9 *Test Security:*

7.9.1 Security refers to both the security of test materials themselves and the security of examinee outcomes. Security of test materials ensures that all examinees are equally advantaged in terms of access to test items. Security of outcomes ensures that the privacy and rights of individual examinees are protected.

7.9.2 Test materials that shall be secured include, but are not limited to: test layout forms, item pools, operational or field test books, test questions or test book sections, answer documents, and test administrator manuals. Test materials may be in paper or electronic format.

7.9.3 Test security procedures for materials and outcomes shall be documented in accordance with 5.8.

7.9.4 *Securing of Test Materials*—All secure test materials, whether paper or electronic, such as test booklets, blank answer sheets, and answer keys shall be kept in a secure location in accordance with the test security plan. Access to tests and test materials shall be limited to personnel who have a legitimate need, and procedures shall be established to determine who may access test materials. A list of individuals with access to the test materials shall be kept current.

7.9.5 *Access to Secure Test Materials*—Persons with access to secure test materials shall not use their access to those materials for personal gain. Personnel with access to test materials shall not disclose the content of secure tests by discussing specific test questions or information contained within the test with unauthorized colleagues, outside organizations, or test takers unless authorized and required to do so during the performance of their work responsibilities. Test specifications and test frameworks may be disclosed only if they are explicitly designated by the owners of those documents as public documents. Persons with access to secure test materials shall receive training on test security and what nondisclosure means and how to avoid inadvertently disclosing test content. Any personnel with access to secure test materials shall be bound by nondisclosure agreements.

7.9.6 *Security Considerations for Stakeholders*—The normal course of the test development process is understood to include stakeholder involvement in developing test specifications, item writing and review, experimental form/

section review, bias review, operational form review, and standard setting. During these processes, if stakeholders are asked to view secure test materials they shall be bound by additional nondisclosure agreements.

7.9.7 *Security Considerations for Test Takers*—Test takers shall be made aware of their personal and legal responsibilities to maintain test security. Although this practice cannot regulate the behavior of test takers, organizations using tests shall take measures to prevent test takers from disclosing secure test information and acting in a way that gives them an unfair advantage over other test takers.

7.9.8 *Disclosure of Test Security Breach:*

7.9.8.1 Organizations, test developers, and test users have a responsibility to document and inform relevant stakeholders or law enforcement or both of any breech in test security. In the event of any security breech, corrective and remedial action shall be implemented to address the cause of the breech and ensure continued test validity and reliability.

7.9.8.2 Procedures to handle predictable breeches of security shall be established and documented by all parties responsible for maintaining test security. A periodic review committee should meet to review all security breaches, especially novel and major breeches to test security.

7.9.9 *Test Administration Security Measures*—When a test is administered, proctors shall secure all test-related materials at the end of the testing period. In a paper-based environment, proctors shall check test booklets for marks after each use; the test developer shall establish a procedure for the return and replacement of booklets that contain marks made by examinees and are therefore no longer usable. Examinees shall not be permitted to bring cell phones or other electronic equipment that might be used to photograph or copy test materials into the testing area. During testing, only examinees and proctors shall be present in the area where the test is being administered. Examinees shall not communicate with anyone other than the test proctor.

7.9.10 *Access to Secure Test Results*—Test results shall be made known only to designated authorities at the institution administering the test. Access to completed tests and results shall be restricted to those with permission. The developer shall outline a policy for length of time that completed tests (paper versions of listening, reading, and writing tests; recordings of oral tests) are retained before they are destroyed.

8. Test Acceptance

8.1 The next phase of the test life cycle is test acceptance, whereby a test moves from design and development into operational use. The following sections describe the conditions that shall be met as a prerequisite to test acceptance, including evidence of test validity, evidence of test reliability, and full documentation that the product acceptance plan has been followed and all criteria for acceptance have been met.

8.2 *Building the Validity Argument and Establishing Reliability:*

8.2.1 *Establishing Validity:*

8.2.1.1 As the test development phase comes to an end, documentation shall be provided in support of the validity argument. The validity argument is not about one specific type of evidence. It rests on the accumulation and documentation of evidence throughout the various stages of the test's life cycle. Validity evidence is developed in the context of test use and with consideration for how the test scores will be interpreted. Validity evidence may be theoretical, empirical, objective, or subjective in nature. Validity arguments include, but are not limited to, explanations of each characteristic described in the following:

(1) The degree to which the test contains an adequate representation of the language skills it is intended to test (content validity);

(2) The degree to which the test measures the underlying construct it is supposed to be measuring (construct validity); and

(3) The degree to which the test appears acceptable, appropriate, and useful to the stakeholders described in the framework (face validity).

8.2.1.2 The aforementioned items cover most types of validity evidence but are not meant to constitute an exhaustive or exclusive list. Test developers shall provide sufficient information for test users to understand the appropriate inferences that can be drawn from the test scores and the types of decisions that can be supported by those scores. This information shall include an explanation of the reliability of the test scores and the level of confidence that can be associated with those scores.

8.2.2 *Evidence of Validity*—Evidence of validity for the intended use and population of the test shall be provided to support the validity argument. The selection of specific validation techniques shall be determined on a case-by-case basis considering the purpose of the test, the type of test, the existing constraints, and the current state of the art for validity evidence. Reasonable attempts shall be made to assess and document validity rigorously and conform to prevailing professional standards for validity evidence in psychometrics. Evidence of validity shall be provided in accordance with 5.8.

8.2.3 *Establishing Reliability:*

8.2.3.1 Evidence of reliability can only be empirical and quantitative. Reliability evidence is collected and presented in the context of test use and with consideration for how the test scores will be interpreted. When verifying the reliability of a test using a single test administration (per examinee), a test developer shall use an appropriate type of study or method, such as one of the following, according to the nature of the test and purpose of the reliability study:

(1) The degree to which results by examinees across subsets of items (or measures) within the test are comparable (internal consistency) and

(2) The degree to which results generated by different raters rating the same performance(s) by the same examinee(s) are comparable (inter-rater consistency and agreement).

8.2.3.2 When verifying the reliability of a test using multiple test administrations (per examinee), a test developer shall use an appropriate type of study or method, such as one or more of the following, according to the nature of the test and purpose of the reliability study:

(1) The degree to which results by the same examinees who take the same form of the test under the same conditions within a limited period of time are comparable (test-retest),

(2) The degree to which results by examinees on different forms of the test or interview are comparable (equating of forms), and

(3) The degree to which results by the same examinees on different delivery modes (for example, web, telephonic, paper, and video) of the test or interview are comparable (equating of delivery modes).

8.2.4 *Evidence of Reliability*—Evidence of reliability for the intended use of the test shall be demonstrated and documented. The selection of specific reliability study design parameters and analysis techniques (that is, statistics used and reported) shall be determined on a case-by-case basis considering the purpose of the reliability study, the type of test, the existing constraints, and the current state of the art for reliability studies and analysis. Reasonable attempts shall be made to assess reliability rigorously and conform to prevailing professional standards for reliability studies in psychometrics. Evidence of reliability shall be provided in accordance with 5.8.

8.3 *Test Acceptance*—Test acceptance is a stage in the product acceptance process. Once the validity argument has been built and documented to the satisfaction of the test stakeholders, the test is ready for acceptance. At the test acceptance stage, the test developer shall provide documentation that the product acceptance plan has been followed and all criteria for acceptance have been met (see 6.5).

9. Test Administration and Scoring: Procedures and Policies

9.1 Once the test has been accepted into operational use, the next phase is ongoing test administration and scoring. Test administration shall conform to the test administration manual to ensure standardization of procedures. Guidelines for test administration (consistent with 6.7.7 and 7.3.1) and scoring shall be articulated in detail and documented in accordance with 6.7. These guidelines shall address the delivery conditions (9.2), test proctoring (9.3), the role of testers (9.4), rating and scoring (9.5), procedures and materials (9.6), reporting of test results (9.7), policies for arbitration (9.8), retest conditions (9.9), score expiration (9.10), and records management (9.11).

9.2 *Delivery Conditions*—Test conditions shall be standardized to the extent possible, with reasonable accommodations, to ensure an equitable test experience for all test takers. Test conditions shall be consistent with the specifications laid out in 6.7.7.

9.3 *Test Proctors:*

9.3.1 *Testing Organization Responsibilities and Proctor Quality Assurance:*

9.3.1.1 In advance of testing, qualifications and profile of acceptable proctors shall be determined. This shall include factors that exclude acceptance as proctor (for example, proctor is closely related to test candidate, proctor has a conflict of interest, proctor is physically or otherwise unable to administer test, and so forth).

9.3.1.2 Independently validate identification of proctor candidate—required qualifications/status of potential proctors.

9.3.1.3 Provide written instructions that include:

(1) Clearly articulated expectations and responsibilities and

(2) Clearly articulated procedures for preparation of test environment, test administration, and concluding test session.

9.3.1.4 Obtain written agreement to comply with expectations, procedures, and protocols (pretest).

9.3.1.5 Ensure technical and content support available during test administration (pretest).

9.3.1.6 Institute controls to ensure proctor is in compliance with test procedures.

9.3.1.7 Deactivate proctors that deviate from approved protocol.

9.3.2 *Proctor Responsibilities:*

9.3.2.1 Ensure integrity of test environment.

9.3.2.2 Validate identification of test taker.

9.3.2.3 Be familiar with and prepared to comply with specified protocols.

9.3.2.4 Ensure integrity of sample.

9.3.2.5 Communicate with test administrators.

9.3.2.6 Conclude testing process is in compliance.

9.3.3 *Technology Enhanced and Technology Based Proctor Protocols*—Tests are proctored to fulfill necessary, standardized test administration and test security functions. Current proctoring practices rely upon trained human proctors; however, alternative nonhuman proctoring methods that exploit advanced technologies are emerging. Any nonhuman proctoring procedures shall perform all of the test administration and test security functions articulated in this section to the same level as a human proctor.

9.4 *Testers:*

9.4.1 Testers are individuals who interact with examinees to deliver dynamic test content. They have a function different from test proctors in that test proctors do not interact directly with the test content—proctors deliver items in their static form, exactly as developed by developers. Testers, on the other hand, have the potential to influence the delivery of test content, affecting both the validity and the reliability of the score. For example, testers may read stimulus material such as prompts for a speaking task. Through tone of voice, pace, and emphasis, they may affect how the prompt is perceived by examinees. Testers may also affect the test by selecting which content is to be administered at particular points in the test.

9.4.2 Individuals who are testers may also be test developers, developing the items that they will deliver (see 7.2 and 7.4) or raters or both, rating examinee performance on items that they or other testers deliver (see 9.5).

9.4.3 *Tester Qualifications and Training:*

9.4.3.1 Minimal qualifications and training criteria shall be established for recruitment of tester trainees.

9.4.3.2 Training processes and qualification requirements shall be clearly defined and agreed to by the trainees.

9.4.3.3 Procedures for how, where, and when testers may administer a test shall be included in tester training.

9.4.3.4 Procedures for the selection and method of delivery of test items shall be included in tester training. This aspect of training shall include elicitation techniques as well as instruction on how to avoid common pitfalls, such as being distracted by an examinee's accent in a speaking test.

9.4.3.5 Tester qualification shall be based on an individual's demonstrated ability to follow elicitation and item selection standards.

9.4.3.6 Testers shall sign tester conduct agreements and nondisclosure agreements.

9.4.3.7 Testers shall maintain confidentiality regarding test materials, test security, and test taker information.

9.4.3.8 Testers shall maintain ethical standards in testing.

9.4.3.9 Testers shall participate in periodic retraining activities.

9.4.4 *Evidence of Performance:*

9.4.4.1 Testers shall give evidence of reliable performance. They shall undergo evaluation of their ability to deliver test items. Reliability standards shall be specified in the test specification document.

9.4.4.2 Evidence shall be reevaluated periodically. If at any point it is determined that testing personnel are not performing within specified guidelines, personnel may be placed on inactive status.

9.5 *Rating and Scoring:*

9.5.1 *Human Rating and Scoring:*

9.5.1.1 Human rating is defined as the human assessment of a test taker's response to a test item when one or more of the following conditions exist:

(1) There are more correct or partially correct answers than can be realistically specified and listed in an answer key,

(2) More than one point may be awarded to a single item (for example, 0, 1, or 2), or

(3) When the rating protocols use a criterion-referenced scale (for example, the ILR scale).

9.5.1.2 Rating may require the rater to be able to reread or listen to a test taker's response multiple times. Examples of rating include a rater assessing an essay or spoken performance (if the performance can be replayed by the rater) using the eleven-point ILR scale. (Contrast with hand scoring [9.5.1.3] and live scoring [9.5.1.4]).

9.5.1.3 Hand scoring is defined as the human assessment of a test taker's response to a test item using an answer key in which all acceptable responses are predetermined. Hand scoring may also be used to verify the accuracy of machine scoring. An example of hand scoring is a scorer using an answer key to assess responses to a paper-and-pencil multiple-choice test. (Contrast with rating [9.5.1.2], live scoring [9.5.1.4], and machine scoring [9.5.4]).

9.5.1.4 Live scoring is defined as the human scoring of a test during or immediately following test administration (with no opportunity for the scorer to replay a test taker's spoken performance or re-read a test taker's written performance) using a clearly defined scoring protocol with unambiguous right, partially correct, and wrong answers. Examples of live scoring include scoring a test taker's speaking performance immediately following the test taker's response to a speaking item. However, since scoring does not require significant judgment on the part of the scorer, because all possible responses are specified in a scoring guide, raters' speaking and writing tasks are generally assessed by raters rather than scorers (contrast with rating [9.5.1.2]).

9.5.2 *Qualifications and Training for Human Raters and Scorers:*

9.5.2.1 Minimal qualifications and training criteria shall be established for recruitment of rater and scorer trainees.

9.5.2.2 Training processes and qualification requirements shall be clearly defined and agreed to by the trainees.

9.5.2.3 Procedures for how, where, and when raters and scorers record ratings and scores to individual test items, sections, and the test as a whole shall be included in rater and scorer training.

9.5.2.4 Rater and scorer qualification shall be based on an individual's demonstrated ability to follow rating and scoring protocols and score reliably.

9.5.2.5 Raters and scorers shall sign rater and scorer agreements and nondisclosure agreements.

9.5.2.6 Raters and scorers shall agree to maintain confidentiality regarding test materials, test security, and test-taker information.

9.5.2.7 Raters and scorers shall agree to maintain ethical standards in rating, scoring, and reporting.

9.5.2.8 Raters and scorers shall agree to participate in periodic retraining activities.

9.5.2.9 Raters and scorers shall be adequately supervised during the rating and scoring processes.

9.5.3 *Evidence of Performance:*

9.5.3.1 Raters and scorers shall give evidence of reliable performance. They shall undergo evaluation of their ability to rate/score examinee performance. Reliability standards shall be specified in the test specification document.

9.5.3.2 Evidence shall be reevaluated periodically. If at any point it is determined that testing personnel are not performing within specified guidelines, personnel may be placed on inactive status.

9.5.4 *Machine Scoring or Rating*—Machine scoring or rating is defined as the automatic scoring or rating of a test item using technology. Examples of machine scoring include grading a pencil-and-paper multiple-choice test with a Scantron® machine[8] or a computer program automatically scoring a computer-based test. Machine rating is an emerging technology and is commonly used to assess speaking and writing performances. (Contrast with rating [9.5.1.1], hand scoring [9.5.1.2], and live scoring [9.5.1.3]).

9.5.4.1 *Requirements for Scoring Machines:*

(1) The minimum technical requirements for the scoring technology shall be specified.

(2) The accuracy of scoring machines shall be periodically verified by hand scoring or rating.

9.6 *Procedures and Materials:*

9.6.1 *Procedures:*

9.6.1.1 Procedures for transporting test materials to the scoring locations shall consider the security of the test materials.

9.6.1.2 The physical and electronic locations where test materials will be stored before, during, and after scoring shall consider the security of the test materials.

[8] Registered trademark of the Scantron Corp., 1313 Lone Oak Rd., Eagan, MN 55121.

9.6.1.3 Procedures for how, where, and when scorers record scores to individual test items, sections, and the test as a whole shall be established.

9.6.1.4 The location where a live scorer will observe the test shall be established and shall consider the needs and comfort of the test taker, test administrator, and live scorer.

9.6.1.5 Procedures for scoring shall be established and these procedures strictly followed during scoring.

9.6.1.6 Periodic reviews and checks that scoring is following the prescribed protocols shall be conducted.

9.6.1.7 If scorers are expected to perform any mathematical calculations to arrive at a section or overall test score or both, procedures or technology or both to verify the accuracy of scorers' calculations shall be established.

9.6.2 *Materials:*

9.6.2.1 An answer key, rating guide, and rating scale (as appropriate) shall be established by the test developer and provided to scorers.

9.6.2.2 Lists of items required to score the test shall be provided. For example, scorers may use only red or green pens to score the test.

9.6.2.3 Lists of items or activities that are prohibited during scoring shall be provided. For example, scorers may not use pencils or pens with blue or black ink to score the test.

9.7 *Reporting of Test Results:*

9.7.1 There should be the appropriate timeframe for reporting the results to the agency/entity requesting the test. The agency may/should have its own reporting deadline policies in place.

9.7.2 The information included in the report should be clear, concise, descriptive, and usable by the agency and the examinee. The results should indicate the ILR levels obtained by the examinee and, if required, the raw scores or any other relevant information.

9.7.3 Only the authorized and designated parties should receive the results of the test, most likely the test administrators of concerned agencies. All reports should be consistent with privacy regulations.

9.8 *Arbitration, Grievances, and Appeal*—The organization using the test shall publish clearly defined policies describing the appeal process; the conditions under which an appeal can be requested; and the roles, responsibilities, and timeline for the appeal process.

9.9 *Retest Conditions:*

9.9.1 The organization using the test shall publish clearly defined policies for retesting, including conditions, the waiting period, the waiver, and appeal procedures.

9.9.2 When administering a retest, an alternate version of the test shall be administered. If it is a humanly rated test, different testers/raters shall conduct/rate the test.

9.9.3 The organization using the test shall publish a clearly defined exemption policy.

9.10 *Score Expiration*—Language proficiency can decrease/increase over time depending on a speaker's use, experience, and training in the language. For this reason, expiration dates of proficiency scores are typically short ranging from six months to two years. Expiration terms of a score are defined by the needs assessment and shall be reasonable. Ultimately, the customer shall determine the score expiration period. It is important that a rationale for the score expiration be clearly articulated and published.

9.11 *Records Management*—The highest priority shall to be given to security when establishing a records management system.

9.11.1 A highly secure, encrypted database is a requirement.

9.11.2 Employees with access to the data shall to be vetted and highly trained to maintain the confidentiality of candidate information and scores.

9.11.3 All active test versions, prompts, and so forth shall be kept secure to ensure that test items are not compromised.

10. Maintenance and Refreshment

10.1 Once the test has been operationalized and administered to test takers, the test maintenance (10.2) and test refreshment (10.3) plans developed during the planning cycle are implemented.

10.2 *Test Maintenance:*

10.2.1 Maintenance means ensuring and documenting that the test remains valid and reliable. The test shall be maintained. Validity and reliability evidence shall be kept up to date. Documents that comprise the validity and reliability evidence for the initial operational use of a test (reliability studies, item analysis, standard setting, alignment, test specifications, training materials, benchmarks, related samples, and sample performances) shall be available, if feasible, for reference to maintain the test. All documentation supporting the test shall be reviewed and updated in accordance with 5.8.

10.2.2 Maintenance can be done by the developer, the distributor, the client organization using the test, or any combination thereof. Test developers shall not be responsible for maintenance in the event that the client organization using the test chooses to use the test for purposes for which it is not intended or does not follow the standards for testing or rating protocols or both. There shall be a plan for what processes will be used to maintain the quality of the test and how often those processes shall be implemented (see 5.8).

10.2.3 Regular, periodic reviews shall be conducted to ensure compliance with the same standards that are required during the development and original implementation of the test. The organization maintaining the test shall review items to check whether content has become dated, the items conform to current standards for item quality, and the test materials or scoring materials or both are consistent with the most current interpretation of the ILR skill level descriptions.

10.2.4 For interactive tests in which testers have a choice among items or prompts to administer, the organization maintaining the test shall obtain information about which items or prompts are most and least frequently selected and solicit feedback from testers concerning why they believe some items or prompts are more useful than others. For example, a role-play situation in a speaking test may appear reasonable to reviewers, but testers may feel that it rarely produces a useful response.

10.2.5 The organization maintaining the test shall conduct statistical analyses periodically to check whether item statistics are consistent with the original item statistics and whether reliability, of whatever type relevant to the test, is maintained. The frequency of analyses depends on the type of test and analysis. Inter-rater reliability analyses for human-rated tests typically need to be performed more frequently than do item analyses of machine-scored tests.

10.2.6 For human-rated tests, the organization rating the tests shall provide periodic rater refresher training and recertification assessments.

10.3 *Test Refreshment*—Refreshment means providing replacement test items or test forms. Refreshment shall take place on a regular basis as test items reach the end of their life cycle, that is, become overexposed or obsolete. There shall be a plan for how items are resupplied consistent with the original purpose of the test (see 6.8). Refreshment shall also take place in the event of a breach of test security.

10.3.1 *Items*—Test items shall be refreshed, revised, and/or replaced according to criteria outlined in the refreshment plan.

10.3.2 *Training Materials*—Training materials shall be updated to ensure their effectiveness.

10.3.3 *Scoring and Rating*—Scoring and rating protocols shall be refined based on statistical analysis, content review, or feedback from scorers or raters.

10.3.4 *Cut-Score Changes*—If items are replaced, cut scores shall be recalculated. If the refreshment plan allows for cut-score changes in the absence of replacement of items, for example, based on impact data or a new standard-setting study, cut scores shall be changed in these circumstances also.

11. Documentation Audits

11.1 Documentation audits (see 6.4.5.7) occur outside of the testing life cycle. They allow inspection of the life cycle and verification of necessary procedures required for proper QC and QA of tests developed under this practice.

11.2 *Documents*—Documentation audits encompass two sets of documents. The first set includes those documents that were created during the planning and development stages that establish the intended purpose of the test, as well as the results of the latest evaluation cycle. This first set of documents shall be provided to auditors in time for a thorough review before any evaluation of the test begins. The second set includes the documentation that shows how the test is being used in its current state. The second set of documentation shall be provided to the auditors during any evaluation occurring after development is complete and the test is accepted.

11.3 Documentation of planning and development includes but is not limited to:

11.3.1 Those documents that are the product of the original needs analysis (6.4),

11.3.2 Test specifications document (6.7),

11.3.3 The framework document (6.6),

11.3.4 Developer-provided materials (7.3),

11.3.4.1 Test administration manual (7.3.1),

11.3.4.2 Training process outline (7.3.2),

11.3.4.3 Rubrics (7.3.3), and

11.3.4.4 Rater/reviewer forms (7.3.4).

11.4 Documentation of test use and performance includes but is not limited to the following, as applicable:

11.4.1 Training requirements for all testers, raters, and reviewers, scorers, and so forth (see 7.2.3, 9.4.3.6, and 9.5.2.5);

11.4.2 Documented test audits/reviews since last evaluation of the same type;

11.4.3 Documentation indicating that prepared test items for dynamic tests are being reviewed on a regular basis by the testers administering them, individually for effectiveness, collectively for standardization (see 7.4.7);

11.4.4 Documentation indicating that static test items are analyzed statistically on an ongoing basis to find those inconsistent with applicable specifications (see 7.4.7);

11.4.5 Documentation indicating that constructed-response static items are reviewed by personnel as specified in 7.4.7;

11.4.6 Documentation of periodic retraining of human raters to ensure rater reliability (see 9.5.2);

11.4.7 Documentation indicating that all proctors that were used in testing conformed to standards established in advance of testing, including a copy of instructions given to them before testing, access to their signed written agreements to comply with established procedures and controls instituted to ensure proctor compliance (see 9.3);

11.4.8 Current list of individuals with access to the test;

11.4.9 The protocols used in hand scoring (see 9.5.1);

11.4.10 Documentation indicating the minimum qualifications and training criteria for scorers, procedures for scoring, and access to signed scorer agreements and nondisclosure agreements (see 9.5.2);

11.4.11 A copy of the procedures for live scoring and documentation indicating the periodic reviews that show live scoring is following prescribed protocols (see 9.5);

11.4.12 Documentation verifying the accuracy of scoring machines (see 9.5.4);

11.4.13 Documentation indicating the minimum qualifications and training criteria for raters, procedures for rating, and access to signed rater agreements and nondisclosure agreements (see 9.5.2);

11.4.14 A copy of all applicable rating and scoring guides (see 9.6.2);

11.4.15 A copy of arbitration, grievance, and appeal policies (see 9.8); and

11.4.16 A copy of retest conditions policy (see 9.9).

11.5 In addition to the above-mentioned documents (if applicable), it may be necessary to provide to the auditors documentation indicating compliance with any and all processes required by local written policies or regulations. That documentation can take any form such as logs, memorandums for record, or any records that are the natural product of a process being undertaken.

11.6 *Documentation Audit Procedures:*

11.6.1 During an audit, the auditors shall be afforded a space in which to work, and copies of guides, requirements, procedures, policies, and protocols shall be placed in that location in advance. Unless it causes a security violation, the auditors shall be given access to those materials in the place they are usually stored.

11.6.2 The auditors shall review all documents and determine whether the documents are in conformance with the test maintenance plan, whether they indicate that there are changes in the test or testing conditions that might affect validity, and whether the documentation is complete. The auditors shall prepare a formal report for the organization using the test.

12. Keywords

12.1 foreign language; ILR scale; Interagency Language Roundtable scale; proficiency

APPENDIXES

(Nonmandatory Information)

X1. SUMMARY OF QC ROLES AND RESPONSIBILITIES

(Code of Fair Testing Practice in Education ©2004 Joint Committee on Testing Practice)

X1.1 See Table X1.1 for QC roles and responsibilities.

TABLE X1.1 QC Roles and Responsibilities

Test Developers	Test Users
Test developers should provide the information and supporting evidence that test users need to select appropriate tests.	Test users should select tests that meet the intended purpose and that are appropriate for the intended test takers.
A-1. Provide evidence of what the test measures, the recommended uses, the intended test takers, and the strengths and limitations of the test, including the level of precision of the test scores.	A-1. Define the purpose for testing, the content and skills to be tested, and the intended test takers. Select and use the most appropriate test based on a thorough review of available information.
A-2. Describe how the content and skills to be tested were selected and how the tests were developed.	A-2. Review and select tests based on the appropriateness of test content, skills tested, and content coverage for the intended purpose of testing.
A-3. Communicate information about a test's characteristics at a level of detail appropriate to the intended test users.	A-3. Review materials provided by test developers and select tests for which clear, accurate, and complete information is provided.
A-4. Provide guidance on the levels of skills, knowledge, and training necessary for appropriate review, selection, and administration of tests.	A-4. Select tests through a process that includes persons with appropriate knowledge, skills, and training.
A-5. Provide evidence that the technical quality, including reliability and validity, of the test meets its intended purposes.	A-5. Evaluate evidence of the technical quality of the test provided by the test developer and any independent reviewers.
A-6. Provide to qualified test users representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports.	A-6. Evaluate representative samples of test questions or practice tests, directions, answer sheets, manuals, and score reports before selecting a test.
A-7. Avoid potentially offensive content or language when developing test questions and related materials.	A-7. Evaluate procedures and materials used by test developers, as well as the resulting test, to ensure that potentially offensive content or language is avoided.
A-8. Make appropriately modified forms of tests or administration procedures available for test takers with disabilities who need special accommodations.	A-8. Select tests with appropriately modified forms or administration procedures for test takers with disabilities who need special accommodations.
A-9. Obtain and provide evidence on the performance of test takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed.	A-9. Evaluate the available evidence on the performance of test takers of diverse subgroups. Determine to the extent feasible which performance differences may have been caused by factors unrelated to the skills being assessed.

X2. TEST SECURITY AND ADMINISTRATION RESPONSIBILITIES

(Adapted from the APA Code of Fair Testing Practices in Education)

X2.1 See Table X2.1 for test security and administration responsibilities.

TABLE X2.1 Test Security and Administration Responsibilities

Ten Rights of Test Takers	Ten Responsibilities of Test Takers	Guidelines for Testing Professionals Test administrators and professionals should:
To be informed of rights and responsibilities as a test taker.	Read or listen or both to rights and responsibilities as a test taker.	Inform test takers of their rights. Ensure test takers know that they have specific responsibilities as a test taker in addition to those rights.
To be treated with respect and impartiality and to be free from discrimination.	Treat others with courtesy and respect during the testing process.	Make test takers aware of any support materials, clearly described in test registration or test materials or both. Ensure test takers provided with reasonable access. Ensure test takers are responsible for their behavior and do not interfere with the rights of others. Ensure test takers do not compromise the integrity of the test in any manner.
Be tested with measures tests that meet professional standards.	Ask questions before testing about test purpose, administration, and results reporting.	Use measures that meet professional standards and are reliable, relevant, useful, and fair to test takers of varying societal groups. Advise test takers of their responsibility to review materials; ask questions; and request more information about administration, content, and results reporting.
To receive explanation before test of test purpose, with accommodations for disabilities or language issues.	Review descriptive information in advance of test and request accommodations for disabilities or language issues.	Give test takers brief description of test purpose, test format, nature of test, and results reporting (including time frame results remain valid). Inform test takers of appropriate use, retest policies, scoring procedures, services/feedback available for a fee, FAQs on administration. Inform test takers of necessary and prohibited materials and equipment. Provide information to facilitate decisions should test takers have options in test format or forms. Advise test takers that they are entitled to request reasonable accommodation (in accordance with ADA requirements). Inform test takers of their right to explanation should their request for accommodation not be granted. Inform test taker of their responsibility to request special testing arrangements (for disability or language issue) and provide necessary documentation.
To be informed of test date, date to expect results, and fees.	Know when and where the test will be given, pay for the test if required, and appear on time with any required materials and be ready to be tested.	Inform test takers of schedule changes, with reasonable alternatives provided. An explanation of fees should be provided in advance. Inform test takers that they are responsible for familiarizing themselves with the appropriate materials and for covering any necessary fees.
To have test administered and results interpreted by trained individuals.	Follow the test instructions you are given and represent yourself honestly during the testing.	Select appropriate tests, provide qualifications of testing professionals if requested, ensure test conditions do not interfere with performance, and provide reasonable time to complete, and safeguard against fraud. Advise test takers of their responsibility to read/listen to direction, follow instructions, and behave honestly.
To understand if test is optional and to understand consequences of refraining from test.	Be familiar with and accept the consequences of not taking the test should you choose not to take the test.	Inform test takers of the purpose of the test, about the consequences on not taking a test should they choose to refrain, and their responsibility to accept such consequences. Engage in testing activities only after they have received informed consent from the test taker.
To received an explanation of test results within a reasonable amount of time after testing and in commonly understood terms.	Inform appropriate persons if testing conditions affected test results.	Provide results, and corrections, within a reasonable amount of time. Interpret test results in light of additional considerations, if relevant. Provide test taker with a copy of the criteria for passing score. Communicate results in an appropriate and sensitive manner.
To have test results kept confidential to the extent allowed by law.	Ask about the confidentiality of test results, if this aspect is of concern to you.	Inform test takers of their responsibility to ask questions about confidentiality. Insure security of records and prohibit access to unauthorized persons. Explain to test takers who has the right to access the information and limit access only to those persons identified before testing. Maintain confidentiality of requests for accommodation.

TABLE X2.1 *Continued*

Ten Rights of Test Takers	Ten Responsibilities of Test Takers	Guidelines for Testing Professionals Test administrators and professionals should:
Present concerns about the testing process or results and receive information about procedures that will be used to address such concerns.	Present any concerns about the testing process or results in a timely, respectful way.	Advise test takers that it is their responsibility to present concerns about the test in timely, respectful manner. Inform test takers of procedures for appealing test results if their test is under investigation and may be canceled/invalidated. In the event that test results are cancelled/invalidated, inform test taker as to why that action was taken.

X3. FOURTEEN TEST SECURITY STANDARDS

(Adapted from the "Caveon Test Security Standards," Caveon, LLC, Midvale, Utah, 84047)

X3.1 Security Plan—Each organization should create and maintain a test security plan, which is a formal, written document that contains the goals of the program, policies and procedures, definitions, roles and responsibilities, the security breach action plan, approvals, and other components and content.

X3.2 Roles and Responsibilities—There are many individuals involved in protecting the security of a program's tests. The roles and responsibilities of these individuals should be identified and communicated to prevent any weakness in overall security.

X3.3 Budget and Funding—Organizations need to establish and maintain a budget and contingency funding for security purposes.

X3.4 Legal Precautions and Agreements—The security of an organization depends to a large extent on its preparation of legal agreements and other precautions taken to secure its legal rights.

X3.5 Test and Item Design—Tests and items should be designed for security purposes. The design should discourage memorization and sharing and make common methods of cheating less effective. They should also limit item exposure to test takers, thereby prolonging the usefulness of items and test results.

X3.6 Test and Item Development and Maintenance—It is important that during the development of items and tests that the content is protected both through the use of agreements as well as sound security procedures.

X3.7 Test Publication—After the test has been created, it is published and distributed. Security measures shall be in place to protect it during this period.

X3.8 Test Administration—Tests need to remain secure immediately before, during, and after test administration. Test administration refers to the process of registering examinees, scheduling, providing physical security measures, presenting the test content, gathering the test results, and communicating results and other information to the organization.

X3.9 Test Scores and Results—Test scores should be subjected to a security analysis to validate their usefulness for subsequent decisions. In addition, the accuracy of the scoring process, from a security perspective, should be verified.

X3.10 Information Security—Digital and physical information related to the organization's testing program shall be stored and transmitted securely at all times.

X3.11 Web and Media Monitoring—With the ubiquity of the internet, it is critical that a high-stakes testing program monitor the web for the disclosure of its copyrighted items and other test information.

X3.12 Security Awareness and Training—The organization should take proper steps to communicate the value of a security plan; the importance of specific aspects of test security; and the importance of confidentiality to its staff, contractors, vendors, and volunteers.

X3.13 Security Breach Action Plan—An organization should have a concrete plan for responding to actual or alleged security breach incidents. This response may involve further investigation and evaluation of compliance with policies and procedures by the individuals or organizations involved and appropriate and measured responses to the actual or alleged incident.

X3.14 Physical Security—The organization has formal policies and procedures to promote the security of items, tests, and related material in the work area, including access to the area and files and careful monitoring of use of materials.

RELATED MATERIAL

Association of Language Testers in Europe, Multilingual glossary of language testing terms, Studies in Language Testing 6, Cambridge: UCLES/Cambridge University Press, 1998a.

Bachman, L. F. and Palmer, A. S., Language Testing in Practice, Oxford: Oxford University Press, 1996.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., and McNamara, T., Dictionary of Language Testing, Studies in Language Testing 7, Cambridge: UCLES/Cambridge University Press, 1999.

Henning, G., A Guide to Language Testing Development, Evaluation and Research, Cambridge, MA: Newbury House, 1987.

Richards, J. C., Platt, J., and Platt, H., Dictionary of Language Teaching and Applied Linguistics, London: Longman, 1992.

Saal, F. E., Downey, R. G., and Lahey, M. A., "Rating the Ratings: Assessing the Psychometric Quality of Rating Data," in Psychological Bulletin, Vol 88, No. 2, 1980.

ASTM International takes no position respecting the validity of any patent rights asserted in connection with any item mentioned in this standard. Users of this standard are expressly advised that determination of the validity of any such patent rights, and the risk of infringement of such rights, are entirely their own responsibility.

This standard is subject to revision at any time by the responsible technical committee and must be reviewed every five years and if not revised, either reapproved or withdrawn. Your comments are invited either for revision of this standard or for additional standards and should be addressed to ASTM International Headquarters. Your comments will receive careful consideration at a meeting of the responsible technical committee, which you may attend. If you feel that your comments have not received a fair hearing you should make your views known to the ASTM Committee on Standards, at the address shown below.