Designation: F3516 – 22

Standard Guide for Testing Interpreting Performance[1]

This standard is issued under the fixed designation F3516; 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:*

1.1.1 This guide describes factors to be considered for the development and use of language interpreting performance tests, referencing the Interagency Language Roundtable (ILR) scale. It is intended to help people commission, develop, or select assessment tools for the evaluation of interpreting skills.

1.1.2 The purpose of any test developed following this guide is to rate a candidate's interpreting skills according to the Interagency Language Roundtable Skill Level Descriptions for Interpreting Performance (ILR SLDs for Interpreting). Any pass/fail rating assigned should reference the specific ILR level at which the candidate has tested.

1.1.3 The objectives for all tests should be clearly defined and convincing evidence presented to justify any claims, inferences, and decisions.

1.1.4 This guide focuses on two types of assessment; one is for screening candidates, and the other is for evaluating actual interpreting skills. It also outlines the appropriate characteristics and uses of each.

1.1.5 When evaluating actual interpreting skills, it should be noted that according to ILR, it is at the Professional Performance Level 3 that all necessary skills align to enable a reasonably accurate, reliable, and trustworthy interpretation.

1.2 *Limitations:*

1.2.1 This guide is not intended to address test development for use in the following areas:

1.2.1.1 Translation,

1.2.1.2 Audio Translation,

1.2.1.3 Transcription/Translation,

1.2.1.4 Diagnostic Assessments,

1.2.1.5 Less-commonly tested languages, and

1.2.1.6 Other job-specific language performance tests.

1.2.2 This guide also does not purport to prescribe definitive descriptions of every possible approach for testing interpreting performance, nor does it prescribe the exact parameters that must be used in a valid and reliable test of interpreting skills. It does, however, suggest approaches to help test designers and users determine whether the use of a test is appropriate and justifiable.

1.2.3 This guide is not intended to address ancillary processes and procedures governing how organizations provide interpreting services.

1.3 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

F2089 Practice for Language Interpreting

F2889 Practice for Assessing Language Proficiency

3. Terminology

3.1 *Definitions:*

3.1.1 *adaptive tests, n*—tests in which the selection of the next item depends upon the rating assigned to previously taken items.

3.1.1.1 *Discussion*—In computer-adaptive tests, for example, candidates who do not show mastery at one level may not be asked to respond to higher-level prompts, but may be given lower ones to determine their ability. In human-delivered adaptive tests (such as Oral Proficiency interviews), testers select the next prompt based upon how well or badly they believe the candidate handles a previous prompt.

3.1.2 *analytic scoring, n*—results in the assignment of particular values to each individual element of the candidate's performance, which may or may not result in a final overall rating; analytic rating breaks down performance into discrete features and assigns separate ratings or values to each.

3.1.3 *decision-tree guidelines, n*—describe the paths a candidate may take through an adaptive test; they suggest which items should be delivered next, based on measurements of current performance.

[1] This guide 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 April 1, 2022. Published April 2022. DOI: 10.1520/F3516-22.

[2] 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.

3.1.4 *holistic scoring, n*—requires raters to assign a rating based on the overall quality of the candidate's performance, based upon a set of criteria describing typical performance at a particular level; it results in a final rating which does not provide individualized feedback on the discrete elements of performance; contrast with analytic scoring.

3.1.5 *interpreting, n*—the process of first fully understanding, analyzing, and processing a spoken or signed message and then faithfully rendering it into another spoken or signed language.

3.1.5.1 *Discussion*—Interpreting is different from translation which results in the creation of a written target text.

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

3.1.6.1 *Discussion*—Language proficiency encompasses a person's ability to read, write, speak, or understand a language.

3.1.7 *performance, n*—the ability of candidates to perform particular tasks, usually associated with job or study requirements.

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

3.1.8.1 *Discussion*—Quality Assurance (QA) applies when (*1*) a new test is being created, (*2*) an existing test is being repurposed or revised, or (*3*) new personnel is being trained to develop or administer a test, the latter in accordance with uniformly acceptable standards.

3.1.9 *quality control, n*—the system of post-development evaluations used at the point of product acceptance and following product use to determine whether the test and testing practices implemented by an organization continue to meet and adhere to all established standards and relevant testing policies; Quality Control (QC) is part of the test maintenance process.

3.1.9.1 *Discussion*—Quality Control (QC) is used at the point of product acceptance and any time after product use. QC verifies the continued validity and reliability of the test and demonstrates that the test is being used in an appropriate manner on an ongoing basis.

3.1.10 *reliability, n*—the consistency with which a test measures a skill or activity throughout the life of the test or the degree to which it does so without deviation each time it is used (repeatability).

3.1.10.1 *Discussion*—Consistency is the essential idea in classical reliability. Reliability is defined as the extent to which separate measurements (items, scales, test administrations, and interviews) yield comparable results under the same or similar conditions. Test items measuring the same construct should yield similar results when administered to the 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

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

3.1.11.1 *Discussion*—Validity is a judgment of the degree to which the evidence (arguments) supports the conclusions, interpretations, uses, and inferences of test scores. 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 and justification of test uses, effects, audiences, and content. Different statistical procedures can be applied to estimate the validity of a test. Such procedures generally seek to determine what the test measures, and how well it does so. The rigor and strength of the validity argument should increase as the stakes associated with the test (consequences for the individual or organization, or both) increase.

4. Significance and Use

4.1 *Intended Use:*

4.1.1 This guide is intended to assist in the design or evaluation of screening and interpreting tests, or both.

4.1.2 This guide also satisfies the need for testing interpreting performance identified in other relevant ASTM standards (see Practice F2889 and Practice F2089).

4.2 *Compliance with the Guide:*

4.2.1 Compliance requires the user to identify which sections of this guide apply to their specific use and circumstances. The decision to not adhere to any sections should be fully explained.

5. Overarching Considerations

5.1 This guide combines expertise from the fields of language testing and interpreting and describes best practices from each.

5.2 *Test Purpose:*

5.2.1 An interpreting performance test developed in accordance with this guide should place candidates within the range of interpreting performance described in the ILR Skill Level Descriptions for Interpreting Performance.

5.3 *Reliability (See also 3.1.10):*

5.3.1 Without measurement consistency, decisions based on test scores or ratings may be incorrect. Any assessment developed should include an explanation of how reliability will be ensured.

5.4 *Validity (See also 3.1.11):*

5.4.1 A test is considered valid to the extent that it measures what it is intended to measure. A screening test should measure whether a candidate possesses some or all of the prerequisite abilities required of interpreters. It is considered valid if it effectively excludes candidates who do not possess the interpreting skills required for the interpreting assessment.

5.4.2 An interpreting test is considered valid if it measures the interpreting ability of a candidate accurately. It can be developed for use in a specific area of interpreting or it can be intended for more general use.

5.4.3 It is important that the test be used in a manner consistent with what it actually measures. For example, it may not be valid to use a test designed to assess interpreting ability in the medical domain to infer ability in the legal domain. Any validity argument should be rigorous enough to justify the decisions made on the basis of test ratings and the potential consequences of those decisions.

5.5 *Practicality:*

5.5.1 The development of valid, reliable tests requires that resources be allocated for the development, administration, and periodic evaluation and improvement of the assessment. Necessary resources may include the following:

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

5.5.1.2 Funding for assessment development, the compensation, training, and maintenance of raters and administrators, ongoing improvements, and operations and security management; and

5.5.1.3 Sufficient time to plan and execute test development and maintenance processes.

5.5.2 During the test design and development phases, it is often necessary to make tradeoffs between the validity and reliability of a test, and the practical constraints of time, money and other resources. In such cases, it is important to recognize the extent to which validity or reliability, or both, may be compromised.

5.6 *Technical Documentation:*

5.6.1 Technical documentation covering the entire test lifecycle includes, but is not limited to, the following:

5.6.1.1 Needs Analysis,

5.6.1.2 Test Specifications,

(1) Test use,

(2) Test design, and

(3) Test scoring/rating.

5.6.1.3 Test Validation,

5.6.1.4 Test Administration,

5.6.1.5 Test Security, and

5.6.1.6 Test Refreshment.

5.6.2 Documentation should serve to assure interested parties of the applicability and rigor of the approach, processes, methodologies, findings, decisions, and deliverables at each stage of the lifecycle.

5.7 *Ethics:*

5.7.1 This guide addresses the ethical considerations that must be part of any assessment of interpreting performance in keeping with good testing practice. Several organizations have created ethical codes of practice designed to safeguard the rights of test takers by focusing on professional test development, administration, and rating practices. These instruments can also serve as guides to ethical behavior in interpreting performance testing.

5.7.2 The development and use of an interpreting test entail ethical responsibilities for contracting agencies, testing organizations, test developers, and test users who must consider the implications of their own actions as well as those of others during all phases of testing.

6. Test Planning

6.1 *Test Types:*

6.1.1 Based on the results of the Needs Analysis, tests can be used to measure general language or be domain-specific.

6.1.2 Screening tests are easier to administer and may prove cost-effective by eliminating candidates with little or no chance of attaining the desired level on the ILR scale for interpreting performance.

6.1.3 The purpose of the screening tests is to identify individuals who are unlikely to perform well on interpreting performance tests. The tests can be used to assess the source or the target languages, or both. While language proficiency is a prerequisite, it is not enough to ensure a successful interpreting performance.

6.1.4 Regardless of the nature of the screening test, it is critical that empirical evidence be provided demonstrating that the screening test is an effective indicator of how well a candidate will perform on an Interpreting Performance Assessment.

6.1.5 While the method of test delivery is largely irrelevant as long as it does not affect test validity, a written test format for the screening test can be justified for practical considerations.

6.1.6 An Interpreting Performance Test should require that candidates demonstrate that they can interpret effectively in the interpreting mode required.

6.2 *Screening Assessments:*

6.2.1 A screening test measures whether or not a candidate possesses some of the prerequisite skills required of interpreters. Ideally, it may test language proficiency using the ILR Skill Level Descriptions for Proficiency in the following:

6.2.1.1 Speaking,

6.2.1.2 Listening comprehension,

6.2.1.3 Reading comprehension,

6.2.1.4 Writing,

6.2.1.5 American Sign Language (ASL) comprehension, and

6.2.1.6 American Sign Language (ASL) production.

6.2.2 There are additional elements which may be assessed:

6.2.2.1 Written translations,

6.2.2.2 Grammar and vocabulary,

6.2.2.3 Specialized terminology,

6.2.2.4 Interpreting protocols,

6.2.2.5 Ethics, and

6.2.2.6 Situational decision-making.

6.3 *Interpreting Assessments:*

6.3.1 An interpreting assessment measures the candidate's integrated ability to interpret, conveying meaning and exhibiting the conduct appropriate to the level(s) being tested. Depending on the interpreting mode being assessed, the test should evaluate both receptive (listening, reading, or ASL comprehension) and productive skills (speaking or signing).

6.3.2 One or multiple modes of interpreting (simultaneous interpreting, consecutive interpreting, and sight translation) may be tested, either unidirectionally or bidirectionally.

6.4 *Test Planning Requirements:*

6.4.1 Prior to test development, a series of planning steps should be considered to produce a document which will be used to guide the development and use of an assessment. It would include the following elements:

6.4.1.1 Needs Analysis,

6.4.1.2 Test Specifications,

(1) Test use,

(2) Test design, and

(3) Test scoring/rating.

6.4.1.3 Test Validation,

6.4.1.4 Test Administration,

6.4.1.5 Test Security, and

6.4.1.6 Test Refreshment.

6.5 *Needs Analysis:*

6.5.1 The development, commissioning, or selection of an interpreting test should be based on the needs of the organization commissioning or selecting the test. To ensure that the test is appropriate for its intended use, the organization should perform a Needs Analysis.

6.5.2 The Needs Analysis should include input from an appropriately broad range of stakeholders. A test may be general in nature or be designed to evaluate interpreting skills for particular domains or a specific requirement.

6.5.3 The Needs Analysis should cover the following:

6.5.3.1 The interpreting requirements of the organization(s) that will use the test and those of their clients (including domain, mode, ILR level, working direction, language pairs, or dialects, if applicable);

6.5.3.2 The type of decisions that will be made on the basis of test scores;

6.5.3.3 How many examinees will take the test, which will dictate how many test forms may be needed;

6.5.3.4 How often an examinee may be tested;

6.5.3.5 The facilities to be used for testing; and

6.5.3.6 The location of test candidates (in-person or remote).

6.6 *Test Specifications:*

6.6.1 Test specifications should justify and explain the rationale for test use, design, content, and scoring/rating. They should be easily available to the public.

6.6.1.1 *Test Use*—The specifications should clearly state the purpose of testing and define the construct(s) to be measured, making specific reference to the ILR interpreting performance guidelines, and identifying the domains and modes of interpreting to be tested.

6.6.1.2 *Test Design*—The specifications should contain the following:

(1) An explanation of how test design reflects the Needs Analysis,

(2) A description of test format and delivery method, and

(3) Detailed specifications for types of test items, content coverage, and the number and nature of items by level/domain.

6.6.2 *Test Scoring/Rating:*

6.6.2.1 The scoring/rating section of test specifications should explain how scores are calculated and how ratings (referencing the ILR SLDs for Interpreting) are assigned. Guidance for test scoring/rating should be provided in the following areas:

(1) Scoring specifications explaining in detail how both raw and scaled scores are generated (as applicable), and how cut scores are set and interpreted;

(2) Partial credit scoring models and criteria for evaluating and rating constructed responses by human raters should be described in detail (as applicable);

(3) Rating specifications should include explanations of how raters are trained and the rating scale being used for rating;

(4) Any key used to assist in the generation of scores or ratings should be described in detail; and

(5) Individual testing and reporting of each modality.

6.7 *Test Validation:*

6.7.1 A test is valid if it tests what it purports to test. Accordingly, care should be taken that the tests asks candidates to perform authentic tasks which closely mimic the types of interpretations they will have to perform in the real world.

6.7.2 Ideally, test validation is performed by an independent party. Whether it is done independently or by the organization responsible for test development, results must be published in a document which justifies the ILR ratings assigned, and the types of decisions being made based on those ratings.

6.7.3 It is incumbent on the users of the test to determine the legitimacy of the validation approaches used.

6.7.4 As part of the Test Validation process, the test may be piloted to help determine whether the test functions as intended in the real world. Relevant factors to be considered may include the scale of piloting, the population to be used, the range of acceptable scores for item performance (generally used in screening tests), and the range of ratings which would constitute the acceptable or unacceptable performance of the test (in both screening and interpreting tests).

6.7.5 Processes should be implemented to ensure that the test remains valid and reliable over time and evidence preserved, which may include the following:

6.7.5.1 A list of the documents comprising reliability and validity evidence to be preserved in anticipation of future reviews and audits;

6.7.5.2 Information describing how test and item performance will be evaluated;

6.7.5.3 Specification of how often evaluations will be performed; and

6.7.5.4 The metrics used to determine item or test life cycle or both; exposure to a certain number of examinees, time elapsed, or some combination thereof.

6.8 *Test Administration:*

6.8.1 Test specifications should describe standard test administration conditions and procedures. Examples of administration guidelines include, but are not limited to, the following:

6.8.1.1 The physical testing environment or setting;

6.8.1.2 Time allotted for test administration;

6.8.1.3 Selection of test administration personnel, including any qualification and training requirements; and

6.8.1.4 Documents, materials, tools, and equipment required by test takers or test administrators.

6.9 *Test Security:*