Designation: ~~E2849 – 13~~ E2849 – 18

An American National Standard

Standard Practice for Professional Certification Performance Testing[1]

This standard is issued under the fixed designation E2849; 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 This practice covers both the professional certification performance test itself and specific aspects of the process that produced it.

1.2 This practice does not include management systems. In this practice, the test itself and its administration, psychometric properties, and scoring are addressed.

1.3 This practice primarily addresses individual professional performance certification examinations, although it may be used to evaluate exams used in training, educational, and aptitude contexts. This practice is not intended to address on-site evaluation of workers by supervisors for competence to perform tasks.

1.4 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate ~~safety~~ safety, health, and ~~health~~environmental practices and determine the applicability of regulatory limitations prior to use.*

1.5 *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. Terminology

2.1 *Definitions*—Some of the terms defined in this section are unique to the performance testing context. Consequently, terms defined in other standards may vary slightly from those defined in the following.

2.1.1 *automatic item generation (AIG), n*—a process of computationally generating multiple forms of an item.

2.1.2 *candidate, n*—someone who is eligible to be evaluated through the use of the performance test; a person who is or will be taking the test.

2.1.3 *construct validity, n*—degree to which the test evaluates an underlying theoretical idea resulting from the orderly arrangement of facts.

2.1.4 *differential system responsiveness, n*—measurable difference in response latency between two systems.

2.1.5 *examinee, n*—candidate in the process of taking a test.

2.1.6 *gating item, n*—unit of evaluation that shall be passed to pass a test.

2.1.7 *inter-rater reliability, n*—measurement of rater consistency with other raters.

2.1.7.1 *Discussion*—

See *rater reliability*.

2.1.8 *item, n*—scored response unit.

2.1.8.1 *Discussion*—

See *task*.

[1] This practice is under the jurisdiction of ASTM Committee E36 on Accreditation & Certification and is the direct responsibility of Subcommittee E36.80 on Personnel Performance Testing and Assessment.

Current edition approved ~~Dec. 1, 2013~~Nov. 1, 2018. Published ~~December 2013~~November 2018. Originally approved in 2013. Last previous edition approved in 2013 as E2849 – 13. DOI: ~~10.1520/E2849-13.~~10.1520/E2849-18.

2.1.9 *item observer, n*—human or computer element that observes and records a candidate's performance on a specific item.

2.1.10 *on the job, n*—another term for "target context."

2.1.10.1 *Discussion*—

See *target context*.

2.1.11 *performance test, n*—examination in which the response modality mimics or reflects the response modality required in the target context.

2.1.12 *power test, n*—examination in which virtually all candidates have time to complete all items.

2.1.13 *practitioners, n*—people who practice the contents of the test in the target context.

2.1.14 *rater reliability, n*—measurement of rater consistency with a uniform standard.

2.1.14.1 *Discussion*—

See *inter-rater reliability*.

2.1.15 *reconfiguration, n*—modification of the user interface for a process, device, or software application.

2.1.15.1 *Discussion*—

Reconfiguration ranges from adjusting the seat in a crane to importing a set of macros into a programming environment.

2.1.16 *reliability, n*—degree to which the test will make the same prediction with the same examinee on another occasion with no training occurring during the intervening interval.

2.1.17 *rubric, n*—set of rules by which performance will be judged.

2.1.18 *speeded test, n*—examination that is time-constrained so that more than 10 % of candidates do not finish all items.

2.1.19 *target context, n*—situation within which a test is designed to predict performance.

2.1.20 *task, n*—unit of performance requested for the candidate to do; a task can be scored as one item; a task may also be comprised of multiple components each of which is scored as an item.

2.1.21 *test, n*—sampling of behavior over a limited time in which an authenticated examinee is given specific tasks under specified conditions, tasks that are scored by a uniformly applied rubric.

2.1.21.1 *Discussion*—

A test can also be referred to as an assessment, although typically "assessment" is used for formative evaluation. This practice addresses specifically certification and licensure, as stated in 1.3. A test is designed to predict the examinee's behavior in a specified context, the "target context."

2.1.22 *trajectory, n*—candidate's path through the solution to a single item, task, or test.

2.1.22.1 *Discussion*—

Also termed the response trajectory.

2.1.23 *validity, n*—extent to which a test predicts target behavior for multiple candidates within a target context.

3. Significance and Use

3.1 This practice for performance testing provides guidance to performance test sponsors, developers, and delivery providers for the planning, design, development, administration, and reporting of high-quality performance tests. This practice assists stakeholders from both the user and consumer communities in determining the quality of performance tests. This practice includes requirements, processes, and intended outcomes for the entities that are issuing the performance test, developing, delivering and evaluating the test, users and test takers interpreting the test, and the specific quality characteristics of performance tests. This practice provides the foundation for both the recognition and accreditation of a specific entity to issue and use effectively a quality performance test.

3.2 Accreditation agencies are presently evaluating performance tests with criteria that were developed primarily or exclusively for multiple-choice examinations. The criteria by which performance tests shall be evaluated and accredited are ones appropriate to performance testing. As accreditation becomes more critical for acceptance by federal and state governments, insurance companies, and international trade, it becomes more critical that appropriate standards of quality and application be developed for performance testing.

4. Candidate Preparation

4.1 *Number of Practice Items*—A candidate shall be given access to sufficient practice items that the novelty of the item format shall not inhibit the examinee's ability to demonstrate his or her capabilities.

4.2 *Scoring Rubric Available to Candidates:*

4.2.1 Candidates shall have sufficient information about the scoring rubric to be able to appropriately prioritize their efforts in completing the item or test.

4.2.2 The examinee shall not be provided so much information about the scoring rubric that it diminishes the ability of stakeholders to generalize the examinee's skills from his or her test score.

4.3 *Practice Tests:*

4.3.1 There are two types of practice tests: one for gaining familiarity with the user interface of the test items and the other to allow the candidate to self-evaluate mastery of the content.

4.3.1.1 *User Interface Preparation*—A practice test or tests to familiarize candidates with the user interface shall be made available to the candidate at no charge. The practice test shall be sufficient to assure adequate candidate practice time so that the degree of familiarity with the user interface does not impair the validity of the test.

4.3.1.2 *Content Self-Assessment*—Practice tests that evaluate content mastery may be made available at no charge or for a fee. There is no obligation on the part of the test provider to provide a self-assessment practice test to evaluate content mastery.

NOTE 1—If a practice test is provided, it shall sample test content sufficiently to allow the candidate to predict reasonably success or failure on the test.

4.3.2 Candidates shall know specifically which type of practice test they are requesting.

4.3.3 Both types of practice test shall help candidates understand how their responses are going to be scored.

5. Procedure

5.1 *Item Development*—All requirements in Section 5 may be superseded by empirical, logical, or statistical arguments demonstrating that the practices of a certification body are equivalent to or superior to the practices required to meet this practice.

5.1.1 *Item Time Limits:*

5.1.1.1 When items or test sections can be accessed repeatedly, no item time limit is required to be enforced or recommended to the candidate.

5.1.1.2 When items can be accessed only once, item time limits shall be either suggested or enforced, with a visual timekeeping option for the examinee.

5.1.1.3 For a power test, item time limits shall be set using a standard practice such as the mean item response time measured in beta testing plus two standard deviations for successful candidates within the calibration sample. When sufficient data have been collected from test administrations, the item time shall be recalibrated to reflect performance on the actual test

5.1.1.4 For a speeded test, item time limits shall be determined by measuring minimum acceptable time limits in the target context.

5.1.2 *Differential System Responsiveness*—Differential system responsiveness may be due to variance in network bandwidth, network latency, random-access memory (RAM), storage speed, operating systems, computer processing unit (CPU) count and performance, bus speed, or other factors.

NOTE 2—It is the obligation of the test developer to attempt to measure differences in latency and system responsiveness whenever possible and, if possible, to compensate appropriately for these variations.

5.1.2.1 There shall be compensation in test scoring for variances in the hardware and software environment to assure that all examinees are scored fairly.

NOTE 3—Compensation may be in adjusting item time limits, item latency scoring factors, or other compensatory variables.

5.1.2.2 An examinee taking a test under one set of conditions shall receive the same score as if he or she took the test under any admissible alternative set of conditions.

5.1.3 *References/Citations*—When possible, codes, guidelines, industry standards, application source code, or other evidence shall be sufficient to establish the correctness of scoring a procedure. Where such documentation does not exist, correct responses may be documented as standard practice by a vote of the subject matter expert (SME) advisory panel for the test.

5.1.4 *Rater Reliability*—When human raters are involved in assessing item success, rater reliability shall correlate with an established performance standard greater than 0.80.

5.1.4.1 When multiple raters are used to rate a single performance, inter-rater reliability shall correlate higher than 0.80.