Designation: E2849 – 13

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 and health practices and determine the applicability of regulatory limitations prior to use.*

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 *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.

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

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

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

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

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

2.1.6.1 *Discussion*—See *rater reliability*.

2.1.7 *item, n*—scored response unit.

2.1.7.1 *Discussion*—See *task*.

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

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

2.1.9.1 *Discussion*—See *target context*.

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

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

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

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

2.1.13.1 *Discussion*—See *inter-rater reliability*.

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

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

2.1.15 *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.

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

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

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

2.1.19 *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.

[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. Published December 2013. DOI: 10.1520/E2849-13.

2.1.20 *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.

2.1.20.1 *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."

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

2.1.21.1 *Discussion*—Also termed the response trajectory.

2.1.22 *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.

5.1.5 *Automated Scoring*—To verify automated scoring, the test developer shall develop test cases that verify the scoring of a minimum of 95 % of anticipated responses. When items are scored automatically, for the first 100 administrations of the test, the test developer shall verify that the scoring algorithm is scoring responses correctly. Verification may be done by human observation, alternate scoring mechanisms, playback of recorded performance, or audit of collected data. Initial verification shall be performed for at least 5 % of failed items. After 100 administrations, the developer shall verify 1 % of failed items until at least 200 failed items have been checked.

5.1.6 *Item Stimulus Construction*—The item solution space shall enable options that would be used by at least 95 % of practitioners in addressing the problem represented by the item.

NOTE 4—The estimate of the practitioner percentage can be derived empirically from usability studies, use cases, expert panels, observation, or other empirical means.

5.1.7 *Simulation Representation of Reality*—Simulation rules shall represent reality as it is encountered in the target context or accurately abstract essentials of reality in the target context, unless the content of the item is for the candidate to infer the rules of the simulation.

5.1.8 *Access to Help*—Support available to the candidate during the examination shall reflect the support available in the target context, unless the test is designed to predict candidate behavior in an unsupported environment.

5.1.9 *Reconfiguration*—Reconfiguration is so commonplace in many work environments that it shall be taken into account when evaluating the valid range of interpretations of a performance test.

5.1.9.1 If minimal reconfiguration is encountered in the field, requiring the examinee to take the test with the default configuration is acceptable.

5.1.9.2 If field practice normally involves extensive reconfiguration of the tools, then the test shall allow candidates to import their industry standard configurations into the test environment, provided that doing so does not compromise exam security, provide unfair advantage over other candidates, or impact the generalizability of results.

5.1.9.3 The criterion the test developer shall use to determine "minimal reconfiguration" is whether competence measured with the default configuration will predict performance with a reconfigured system.

5.1.10 *Level of Feedback*—Feedback during the test shall reflect feedback available doing similar tasks in the target context.

NOTE 5—Feedback may be time compressed to minimize testing time. Interim results may be omitted if they do not impact success in performing the item.

5.1.11 *American with Disabilities Act (ADA) Accommodations*—Accommodations shall be fair to the candidate, the testing administrator, other candidates, and the potential employer alike, with no interest predominating. Before awarding accommodations, the test administrator shall discuss with the candidate what the candidate feels would be reasonable accommodations and, when feasible, shall allow the methods candidates use for accomplishing tasks in the target context. The candidate shall possess the capability to perform the required test item in full with the agreed upon accommodations. In no case shall a verbal option be given in place of a performance requirement.

5.1.12 *Sensitivity and Bias*—Items shall be developed with sensitivity toward the cultural context within which the candidate will be practicing the skills evaluated. The items shall not include content that would prevent people of equal ability or skill from exhibiting those abilities or skills.

5.1.13 *Item Response Termination*—Item termination methods used shall create an environment in which the examinee's response during a test will best predict performance in the target context.

NOTE 6—In the target context, if an examinee determines completion of the task, then the examinee shall indicate completion of the task on the test. If, in the target context, an external individual determines completion of the task, then an examiner or external indication shall terminate the item.

5.1.14 *Observer Item Effects*—The test developer shall minimize the intrusiveness of the item observer on the process being evaluated at or below the normal level of supervision encountered by the candidate in the target context.

5.1.15 *Item Scoring:*

5.1.15.1 Item scoring shall be both consistent and fair. The scoring rubric shall be applied in the same manner to all examinees' responses. The scoring rubric shall give credit to all correct responses.

5.1.15.2 There shall be a method that allows an auditor to evaluate scored states of the item, evaluate the accuracy of task and item timing, and assess the accuracy of the weighting scheme if one is applied.

5.1.15.3 When the universe of response trajectories is undefined, scoring for a reasonable set of correct paths to the correct answer shall be verified.

5.2 *Test Development:*

5.2.1 *Equivalent Forms:*

5.2.1.1 *Difficulty: IRT*—Test information functions shall have integrals within 2 % of each other and not depart more than 5 % anywhere along the theta range from –3.0 to +3.0.

5.2.1.2 *Difficulty: Classical Test Theory*—Difficulty between forms shall be equated. The recommended range of P-values is from 0.35 to 0.95.

5.2.1.3 *Discrimination: Classical Test Theory*—A minimum acceptable point-biserial for any item is 0.05.

5.2.1.4 *Content Balancing*—Each of the equivalent forms shall conform to the constraints described in the blueprint.

5.2.1.5 *Time:*

(1) Predicted total time for multiple forms shall be within 5 % of each other.

(2) When content constraints require the inclusion of items that do not meet the criteria described in the test blueprint, the offending items shall be flagged and replacement items inserted as soon as is practical.

5.2.2 *Content Definition*—A role delineation study, practice analysis, or job task analysis shall be conducted to verify that the content of the examination reflects current practice in the target context.

NOTE 7—If it can be demonstrated that the tasks on the examination represent a comprehensive spectrum of the tasks required in the target context, a direct determination of the knowledge required to perform those tasks is not necessary. However, verification of knowledge may be required to establish the generalizability of a specific skill which has been demonstrated.

5.2.3 *Construct Validity Required*—Demonstration of construct validity shall be required to evaluate the comprehensiveness of the scope of the exam in assessing knowledge required for generalizeable performance in the target context.

5.2.4 *Cutpoint Setting*—The cutpoint shall be set to demonstrate that the certified practitioner can competently perform the requisite elements of the target context. A psychometrically justified method shall be used to set the standard.

5.2.5 *Quality Assurance (QA):*

5.2.5.1 QA shall be conducted to minimize the chance that any candidate will encounter a response path that has not been adequately submitted to QA.

5.2.5.2 All reasonable successful trajectories through the solution space as defined by an SME panel shall be tested, except that any logically isomorphic paths need not be executed during QA.

5.2.5.3 The set of unsuccessful trajectories tried by a minimum of 5 % of beta test candidates shall be included in the QA trials.

6. Test Administration

6.1 *Audit Trail:*

6.1.1 For every examination, the test shall record sufficient data to reconstruct sufficient parameters of the end state of any item to allow assessment of its correctness and reconstruct any intermediate results that are scored.

6.1.2 For any examination for which the process is evaluated, the test shall record sufficient information to document the process each candidate followed in the performance of the item.

6.2 *Authentication*—At a minimum, the testing body shall require a government-issued photo identification (ID) to verify the identity of anyone taking a professional or vocational test.

NOTE 8—Additional biometric identity confirmation methods may include retinal scan, fingerprinting, voice print, palm vein scan, or others that become available. Authentication techniques requiring sampling of bodily fluids are not reasonable.

6.3 *Test Security*—Test security practices shall be sufficient to assure the generalizability of the interpretation of the examinee's responses.

6.4 *Test Security Precautions*—When test administration has determined beyond a reasonable doubt that the examinee is taking the test not to obtain certification but to audit the testing process, the test administration has the right to use whatever measures are necessary to protect the authenticity of the test, including terminating the test prematurely, presenting unscored items, and presenting items that do not appear on the actual test.

6.5 *Retest Interval*—The retest interval shall balance the candidate's need for certification with the expense of test administration and the cost of developing alternate forms.

6.6 *Observational Independence*—The test item observer(s) shall not have an incentive to either pass or fail a candidate on any item.

7. Test Measures

7.1 *Reliability*—Reliability shall be measured by a method appropriate to a multidimensional instrument. If a unidimensional measure of reliability is reported, the developer shall report empirical measures that confirm the appropriateness of the unidimensionality assumption.

7.2 *Reporting to Candidates*—Reporting for failed examinations shall be sufficient to provide a guide for studying to remedy skill deficiencies.

7.3 *Test Length:*

7.3.1 Unless fatigue is a factor measured by the test, a test shall be short enough to discount fatigue as a factor.

NOTE 9—Fatigue may be measured through such measures as sequence effects on item pass rate.

7.3.2 If fatigue is not a factor evaluated by the test and fatigue is suspected of impacting score results, sections of the test shall be administered in varying order to measure empirically sequence and fatigue effects on test scores.

7.4 *Measuring Efficiency:*

7.4.1 If a test is speeded, then the test developer shall establish that the level of speededness on the test is also required in the target context.

7.4.2 A test may combine accuracy and response time in a function that reflects efficiency if efficiency is required in the target context.

7.5 *Gating Items*—When the test is stopped prematurely as a result of an examinee failing a gating item, items before the gating item may be treated as missing data.

7.6 *Measurement Error:*

7.6.1 When a test is comprised of fully compensatory domains, the standard error of measurement (SEM) at the cutpoint may be measured by classical methods.

7.6.2 When a test is comprised of gating items, no SEM at the cutpoint may be measured, since there are multiple cutpoints.

7.6.3 When a test is comprised of both compensatory and conjunctive domains, separate SEMs shall be reported for the compensatory domains and for each conjunctive domain.

7.6.4 *Measures of Test Goodness*—When test domains are optional dependent on the requirements of a job site, test reliability shall be reported for the base test and a combination of the base test and each domain offered separately.

NOTE 10—If time constraints allow for optional administration of the entire test, the reliability for the combination of the base test and all optional domains shall be reported as well.

8. Keywords

8.1 authentic testing; evaluation; performance assessment; performance test; performance testing; psychometrics

APPENDIX

(Nonmandatory Information)

X1. GUIDLINES FOR AN IDEAL STANDARD

X1.1 In constructing this practice, we observed the following guidelines for what comprised an ideal standard:

X1.1.1 *Standard:*

X1.1.1.1 An auditor shall be able to determine unambiguously whether a standard is met.

X1.1.1.2 A standard shall assure the integrity of the test for all stakeholders.

X1.1.1.3 A standard shall balance the candidate's right to a fair test with the developer's need for cost-effective test production.

X1.2 These standards were written to focus attention on aspects of performance testing that are either overlooked in most conventional testing standards or are improperly treated in conventional references for multiple-choice testing.

X1.3 Notes Regarding Vocabulary

X1.3.1 *Target Context:*

X1.3.1.1 The target context is most frequently on the job for professional certifications. But for educational performance tests, the target context is often the next level of education. For aptitude testing, the target context may actually be job training.

X1.3.1.2 A test is designed to predict performance in a specific target context, most frequently defined by the scope of the certification.

X1.3.1.3 *Validation*—Finally, interpretation is the key to Cronbach's requirements for testing. However, all testing is validated for interpretation in a specific context. When the target context changes, then the test shall be revalidated for interpretation in the new context.