

STANDARDS

*for Educational and
Psychological Testing*

**Essential Guidance and Key Developments
in a New Era of Testing**

www.aera.net/standards

#TestStandards

September 12, 2014 | Washington, DC

Standards for Educational and Psychological Testing: **Historical Notes**

Wayne Camara, Senior Vice President,
Research, ACT

Management Committee for Revising the
Standards for Educational and Psychological Testing

Editions

- The 2014 *Standards* are the sixth edition and the fifth “Joint Standards” by AERA, APA and NCME.
- Initial *Standards* issued in 1954 by APA (Technical Recommendations for Psychological tests and Diagnostic Techniques) and 1955 by AERA and NCME (Technical Recommendations for Achievement Tests).
- In 1966 a Joint Committee was formed to consolidate and revised the separate documents into a single set of *Joint Standards*.
- Revisions followed in 1974, 1985 and 1999.

Purpose of the *Standards*

- Guide the sound and ethical use of tests and evaluate the quality of tests and testing practices.
- Definitive technical, professional and operational standards for all forms of assessments that are professionally developed and used in a variety of settings.
 - Apply to professional test developers, sponsors, publishers and users – providing criteria for the evaluation of tests, testing practices, and the effects of test use.

Influence and Citation

- Not directly established in response to an expressed governmental or regulatory need, nor in response to legislation or judicial decisions.
- Although the *Standards* have been cited in Supreme Court and lower court decisions regarding assessment practices, as well as regulatory guidance.
- Today nearly all state assessment programs cite the *Standards* in development and use of their educational assessment programs and practices.

Joint Committee

- Members of AERA, APA and NCME with expertise in substantive areas of measurement and assessment.
- Members provide depth and breadth of knowledge in assessment practice and measurement research.
- Members appointed by Management Cmte and are not representing AERA, APA or NCME.

Major Changes in the 2014 *Standards*

Barbara Plake, Professor Emeritus,
University of Nebraska

Lauress Wise, Principal Scientist, Human
Resources Research Organization

Co-Chairs of the Joint Committee for Revising the
*Standards for Educational and Psychological
Testing*

Major Changes

- Consolidated fairness topics into a unified chapter placed into a foundations section parallel to validity and reliability.
- Addressed and updated treatment of topics related to new and emerging technologies such as:
 - Automated test development, scoring, and reporting
 - Adaptive or multi-stage testing
- Expanded discussion of use of tests in accountability

Other Key Changes

- Shifted focus from reliability alone to reliability and precision in order to generalize the consideration of score consistency.
- Integrated employment testing and credentialing standards to clarify the similarities and differences between these applications of test results for making employment and credentialing decisions.
- Organized standards topically within each chapter, parallel to topics covered in the background sections to aid user understanding.

Treatment of Validity

Largely Unchanged

- But validity concerns were made more clearly central to key standards in each of the other chapters.
- Clarification of the role of consequences in evaluating the theory of action behind test uses.

Treatment of Fairness in the 2014 *Standards*

Barbara Plake, Professor Emeritus,
University of Nebraska

Co-Chair of the Joint Committee for Revising the
*Standards for Educational and Psychological
Testing*

New Chapter on Fairness

- Consolidated fairness topics into a unified chapter
 - Providing access for all examinees in the intended population
 - Identifying and removing irrelevant sources of performance
 - Supporting appropriate reporting of results

Location of Fairness Chapter

- First section of the *Standards*, “Foundations”
 - Equal positioning with Validity and Reliability

Distinguishing “Accommodations” and “Modifications”

- Accommodations: retains original, intended construct; results in comparable scores to the non-accommodated assessment.
- Modification: change in the original, intended constructs; results may not be comparable to score from non-modified assessment.
- These adaptations of the original test may be needed to provide fair assessments for examinees with disabilities or limited English language proficiency.

Minimizing Barriers to Access

- Use of test design and development procedures to reduce barriers; “universal design.”
- Attempt to remove issues that might restrict an examinee’s ability to demonstrate what they know and can do because of structural issues with the test.
- No Child Left Behind requires increased testing of students; need to ensure each student has an appropriate means to demonstrate what they know and are able to do.

Expanded Examples

- Individual with disabilities
- Individuals from diverse language and cultural backgrounds
- Young children
- Older adults
- Gender
- Race/ethnicity

Opportunity to Learn

- High-stakes testing applications
- Consider opportunity to learn as a causal factor in score interpretation
- Promotes fair test score interpretations

Treatment of Validity in the 2014 *Standards*

Lauress Wise, Principal Scientist, Human
Resources Research Organization
Co-Chair of the Joint Committee for Revising the
*Standards for Educational and Psychological
Testing*

Validity

- Validity evidence is required to support interpretations of test scores for intended uses.
- The 2014 standards for supporting validity are essentially the same as the 1999 standards.
 - If anything, validity is a more central concern in each of the other chapters of the 2014 *Standards*.

What makes a test valid?

- A test is not valid in itself, but specific interpretations of the test scores are valid for particular uses.
- Evidence required to demonstrate validity varies with interpretation and use. For example,
 - Evidence of the alignment of test content to content standards is essential if test scores are interpreted as indicating mastery of the targeted content.
 - Correlational studies are needed to support interpretations of test scores as predictive of future success (i.e., college and career ready).
 - Internal and external correlational studies may be needed to support diagnostic interpretations of sub-scores.

What is the role of the consequences of test use in validity?

- The *Test Standards* distinguish between:
 - The validity of test score interpretations for particular uses
 - Evaluation of the theory of action behind these uses
- Analyses of intended and unintended consequences is required to support claims for test use.

How much validity evidence is required?

- Validation is an open-ended process
 - Validity evidence should be collected prior to initial test use and further data analyzed as the test continues in operational use.
- Higher levels of evidence are required when test use has important consequences for individuals or for society.
- Professional judgment is needed to balance evidence supporting or contradicting test score interpretations for particular uses.

Treatment of Accountability in the 2014 *Standards*

Laura Hamilton, Senior Behavioral
Scientist, RAND Corporation

Joint Committee for Revising the *Standards for
Educational and Psychological Testing*

Why was new material on accountability introduced?

- Use of tests for accountability has expanded dramatically in K-12 education
 - State and federal legislation (NCLB, RttT)
 - Advances in data systems and analytic methods
 - Growing emphasis on teacher effectiveness
- Tests are also used for accountability in other contexts (e.g., behavioral health, postsecondary education)
- Accountability policies can affect not just test takers but those who provide services to them

Where is accountability addressed in the *Standards*?

- Chapter on Program Evaluation, Policy Studies, and Accountability is most relevant
- Chapter on educational testing is applicable (e.g., testing for promotion or graduation)
- All of the foundations and operations chapters have material that applies to accountability testing

How do the *Standards* address measures of educator effectiveness?

- Value-added measures (VAM) and other indicators of teacher or principal effectiveness are considered “accountability indices”
 - Indices combine information from multiple tests or other sources, or involve complex statistical modeling
 - Standards apply not just to raw scores, but to these indices
- Chapter on workplace testing also applies; VAM can be used for tenure, placement, etc.

Are interim/benchmark tests covered by *Standards*?

- Many schools use interim/benchmark assessments to gauge student progress
 - New CCSS-aligned assessment systems include these
 - Primary purpose is to inform instruction
- Application of *Standards* depends on purpose and context
 - There should be evidence to support specific uses, including low- and high-stakes uses
 - Teacher-developed classroom assessments face different requirements than commercial tests

How do the *Standards* apply to technology-based assessment?

- Assessments increasingly use technology for administration, scoring, and reporting
- The 2014 *Standards* committee discussed technology-based assessment extensively and sought input from outside experts.
- The *Standards* addresses some specific issues related to technology-based testing
 - Test security
 - Automated scoring of open-ended items
 - Automated score reporting

STANDARDS

*for Educational and
Psychological Testing*

**Essential Guidance and Key Developments
in a New Era of Testing**

www.aera.net/standards

#TestStandards

September 12, 2014 | Washington, DC