## **EPSY 8225: Operational Measurement Introduction to Testing Standards**

*Standards for Educational & Psychological Testing* (2014).

# **Standards for Validity**

## Standard 1.0

## Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.

## Standard 1.1

## The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

# **Standards for Reliability/Precision**

## Standard 2.0

## Appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use.

# **Standards for Reliability/Precision**

## Specifications for replications of the testing procedure

## Evaluating reliability/precision

## Reliability/Generalizability coefficients

## Factors affecting reliability/precision

## Standard errors of measurement

## Decision consistency

## Reliability/precision of group means

## Documenting reliability/precision

# **Standards for Fairness**

## Standard 3.0

## All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended populations.

# **Standards for Test Design and Development**

## Standard 4.0

## Tests and testing programs should be designed and developed in a way that supports the validity of interpretations of the test scores for their intended uses. Test developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for intended uses for individuals in the intended examinee population.

## Standard 4.18

## Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical for extended-response items such as performance tasks, portfolios, and essays.

## Standard 4.20

## The process for selecting, training, qualifying, and monitoring scorers should be specified by the test developer. The training materials, such as the scoring rubrics and examples of test takers’ responses that illustrate the levels on the rubric score scale, and the procedures for training scorers should result in a degree of accuracy and agreement among scorers that allows the scores to be interpreted as originally intended by the test developer. Specifications should also describe processes for assessing scorer consistency and potential drift over time in raters’ scoring.

# **Standards for Scores, Scales…**

## Interpretation of scores

## Norms

## Score linking

## Cut scores

# **Standards for Scores, Scales, Norms, Score Linking, and Cut Scores**

## Standard 5.0

## Test scores should be derived in a way that supports the interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.

## Standard 5.12

## A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.

## Standard 5.21

## When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

# **Standards for Test Administration, Scoring, Reporting, and Interpretation**

## Standard 6.0

## To support useful interpretation of score results, assessment instruments should have established procedures for test administration, scoring, reporting, and interpretation. Those responsible for administering, scoring, reporting, and interpreting should have sufficient training and supports to help them follow the established procedures. Adherence to the established procedures should be monitored, and any material errors should be documented and, if possible, corrected.

# **Standards for Supporting Documentation**

## Contents of Test Documents:

## Appropriate use

## Test development

## Test administration and scoring

## Timeliness of delivery of test documents

# **Standards for Supporting Documentation**

## Standard 7.0

## Information relating to tests should be clearly documented so that those who use tests can make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret test scores.

## Standard 7.13

## Supporting documents (e.g., test manuals, technical manuals, user’s guides, and supplemental materials) should be made available to the appropriate people in a timely manner.

# **Standards for Test Users’ Rights and Responsibilities**

## Validity of interpretations

## Dissemination of information

## Test security and protection of copyrights

**2009 BOTA Letter, National Academies**

## In 2009, the National Academies published their letter[[1]](#footnote-1) to the US Department of Education regarding the proposed regulations for the Race to the Top initiative. The National Academies consists of the nation’s leading advisors on Science, Engineering, and Medicine, and also includes the Board on Testing and Assessment (BOTA) – some of the preeminent experts and thought leaders in testing and assessment. In that letter, the BOTA urged the Department to be consistent with measurement theory and particularly aware of how some regulations can promote or prevent valid interpretations and uses of test scores. A number of sections in the letter are strong reminders of core measurement principles – a few are reviewed here. Many of these comments can also be found in their report: *Lessons Learned about Testing*.[[2]](#footnote-2)

# **MCA Technical Manual, 2013-2014**[[3]](#footnote-3)

## Minnesota’s MCA Technical Manual provides important guidance regarding test use and score interpretation (in addition to comprehensive details regarding test development, scoring, scaling, and administration). In the Purpose statement of the manual, it asserts a purpose for the educational assessment program.

# **MCA Tech Manual: Test Purpose Statement**

## *Improved student learning is a primary goal of any educational assessment program. This manual can help educators use test results to inform instruction, leading to improved instruction and enhanced student learning. In addition, this manual can serve as a resource for educators who are explaining assessment information to students, parents, teachers, school boards and the general public.* (p. 9)

# **MCA Tech Manual: Regarding score interpretation**

## *As with any large-scale assessment, the Minnesota Assessments provide a point-in-time snapshot of information regarding student achievement. For that reason, scores must be used carefully and appropriately if they are to permit valid inferences to be made about student achievement. Because all tests measure a finite set of skills with a limited set of item types, placement decisions and decisions concerning student promotion or retention should be based on multiple sources of information, including, but not limited to, test scores.* (p. 72)

# **MCA Tech Manual: Regarding score use**

## *The tests in the Minnesota Assessment System are designed primarily to determine school and district accountability related to the implementation of the Minnesota standards. They are summative measures of a student’s performance in a subject at one point in time. They provide a snapshot of the student’s overall achievement, not a detailed accounting of the student’s understanding of specific content areas defined by the standards. Test scores from Minnesota assessments, when used appropriately, can provide a basis for making valid inferences about student performance. The following list outlines some of the ways the student scores can be used.*

## *Reporting results to parents of individual students*

## *The information can help parents begin to understand their child’s academic performance as related to the Minnesota standards.*

## *Evaluating student scores for placement decisions*

## *The information can be used to suggest areas needing further evaluation of student performance. Results can also be used to focus resources and staff on a particular group of students who appear to be struggling with the Minnesota standards. Students may also exhibit strengths or deficits in strands or substrands measured on these tests. Because the strand and substrand scores are based on small numbers of items, the scores must be used in conjunction with other performance indicators to assist schools in making placement decisions, such as whether a student should take an improvement course or be placed in a gifted or talented program.* (p. 77)

# **MCA Tech Report: Score use for individuals**

## *Individual student test scores must be used in conjunction with other performance indicators to assist in making placement decisions. All decisions regarding placement and educational planning for a student should incorporate as much student data as possible.* (p. 78)

# **MCA Tech Report: Cautions for score use**

## *When interpreting test scores, it is important to remember that test scores always contain some amount of measurement error. That is to say, test scores are not infallible measures of student characteristics. Rather, some score variation would be expected if the same student tested across occasions using equivalent forms of the test. This effect is due partly to day-to-day fluctuations in a person’s mood or energy level that can affect performance and partly a consequence of the specific items contained on a particular test form the student takes. … Nevertheless, measurement error must always be considered when making score interpretations.* (p. 80)

# **MCA Tech Report: Use of strand information**

## *Strand or substrand level information can be useful as a preliminary survey to help identify skill areas in which further diagnosis is warranted. The standard error of measurement associated with these generally brief scales makes drawing inferences from them at the individual level very suspect; more confidence in inferences is gained when analyzing group averages. When considering data at the strand or substrand level, the error of measurement increases because the number of possible items is small. In order to provide comprehensive diagnostic data for each strand or substrand, the tests would have to be prohibitively lengthened. Once an area of possible weakness has been identified, supplementary data should be gathered to understand strengths and deficits.* (p. 81)

# **Board of Testing and Assessment, National Research Council: *LESSONS LEARNED***

## In many situations, standardized tests provide the most objective way to compare the performance of a large group of examinees across places and times.

## A test score is an estimate rather than an exact measure of what a person knows and can do.

## High-stakes decisions about individuals should not be made on the basis of a single test score.

## Tests should not be used for high-stakes decisions if test takers have not had an opportunity to learn the material on which they will be tested.

## States, districts, and schools should aim to maximize the participation of English-language learners and students with disabilities in large-scale tests.

## Teachers need professional development that helps them better understand core principles of assessment and how to apply these to their regular instruction and testing.

## In the design of tests, form must follow function.

## The design process must ensure that test score interpretations are valid.

## The design process must ensure that the test results are reliable and fair.

## Testing professionals should consider the relationships among cognition, observation, and interpretation—the “assessment triangle”—when evaluating the soundness of current educational tests or designing new ones.

## Advances in the cognitive sciences and measurement offer opportunities to develop educational assessments that better support learning.

## The people who design and mandate tests must be constantly vigilant about equity concerns, including opportunity to learn, cultural bias, or adverse impact.

## In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.

## Test results may be invalidated by teaching narrowly to a particular test.

## New testing programs should build in an evaluation component.

## Test developers and policy makers should clearly explain to the public the purpose for a test and the meaning of different levels of test performance.

## When test results are reported to students, teachers, and the public, the limitations of the test should be explained clearly to a lay audience.

1. http://www.nap.edu/catalog.php?record\_id=12780 [↑](#footnote-ref-1)
2. http://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse\_082766.pdf [↑](#footnote-ref-2)
3. http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/ [↑](#footnote-ref-3)