THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

CSAI Update

Valid and Reliable Assessments

Determining whether an assessment is valid and reliable is a technical process that goes well beyond making sure that test questions focus on material covered in state standards. While both of these terms are used by researchers in association with precise statistical procedures, this brief will define assessment validity and reliability in a more general context for educators and administrators.

Reliability

Reliability is a measure of consistency. It is the degree to which student results are the same when they take the same test on different occasions, when different scorers score the same item or task, and when different but equivalent tests are taken at the same time or at different times. Reliability is about making sure that different test forms in a single administration are equivalent; that retests of a given test are equivalent to the original test, and that test difficulty remains constant year to year. When a student must take a make-up test, for example, the test should be approximately as difficult as the original test. There are many such informal assessment examples where reliability is a desired trait. The main difference is how it is tracked. For informal assessments, professional judgment is often called upon; for large-scale assessments, reliability is tracked and demonstrated statistically. Whether it is high-stakes assessments measuring end-of-course achievement, or assessments that measure growth, reliability is critical for any assessment that will be used to make decisions about the educational paths and opportunities of students.

Types of evidence for evaluating reliability may include:

- Consistent score meanings over time, within years, and across student groups and delivery mechanisms, such as internal consistency statistics (e.g., Cronbach's alpha)
- Evidence of the precision of the assessments at cut scores, such as reports of standard errors of measurement (the standard deviation of errors of measurement that are associated with test scores from a particular group of students)
- Evidence of the consistency of student level classification, such as reports of the accuracy of categorical decisions over time (reliability analyses [e.g., overall, by sub-group, by reportable category])

- Evidence of the generalizability of results, including variability of groups, internal consistency of item responses, variability among schools, consistency between forms, and inter-rater consistency in scoring, such as a discussion of reliability in the technical report for the state's assessments[1]

Reliability is expressed mathematically on a scale from zero to one, with one representing the highest possible reliability. Multiple choice and selected response items and assessments tend to have higher reliability than constructed responses and other open-ended item or assessment types, such as alternate assessments and performance tasks, since there is less scoring interpretation involved.[2] Since reliability is a trait achieved through statistical analysis, it requires a process called equating, which involves statistically adjusting scores on different forms of the same test to compensate for differences in difficulty (usually fairly small differences). Equating makes it possible to report scaled scores that are comparable across different forms of a test.

Validity

One question that is often asked when talking about assessments is, "Is the test valid?" The definition of validity can be summarized as how well a test measures what it is supposed to measure. Valid assessments produce data that can be used to inform education decisions at multiple levels, from school improvement and effectiveness to teacher evaluation to individual student gains and performance. However, validity is not a property of the test itself; rather, validity is the degree to which certain conclusions drawn from the test results can be considered "appropriate and meaningful."[3] The validation process includes the assembling of evidence to support the use and interpretation of test scores based on the concepts the test is designed to measure, known as constructs. If a test does not measure all the skills within a construct, the conclusions drawn from the test results may not reflect the student's knowledge accurately—and thus, pose a threat to validity.

To be considered valid, "an assessment should be a good representation of the knowledge and skills it intends to measure," and to maintain that validity for a wide range of learners, it should also be both "accurate in evaluating students' abilities" and reliable "across testing contexts and scorers."[4]

Types of evidence for evaluating validity may include:

- Evidence of alignment, such as a report from a technically sound independent alignment study documenting alignment between the assessment and its test blueprint, and between the blueprint and the state's standards
- Evidence of the validity of using results from the assessments for their primary purposes, such as a discussion of validity in a technical report that states the purposes of the assessments, intended interpretations, and uses of results
- Evidence that scores are related to external variables as expected, such as reports of analyses that demonstrate positive correlations with 1) external assessments that measure similar constructs, 2) teacher judgments of student readiness, or 3) academic characteristics of test takers

---

[1] CCSSO. (2013). *Criteria for procuring and evaluating high-quality assessments*. Washington, DC: Author. Retrieved March 16, 2018 from https://www.ccsso.org/sites/default/files/2017-10/CCSSO%20Criteria%20for%20High%20Quality%20Assessments%2003242014.pdf

[2] RAND Corporation. (1997). Criteria for comparing assessments: Quality and feasibility. In *Using alternative assessments in vocational education*. Retrieved March 16, 2018 from https://www.rand.org/content/dam/rand/pubs/monograph_reports/MR836/MR836.chap4.pdf

[3] Caffrey, E. (2009). *Assessment in elementary and secondary education: A primer*. Congressional Research Service. Retrieved March 16, 2018 from https://fas.org/sgp/crs/misc/R40514.pdf

[4] Darling-Hammond, L., Herman, J., Pellegrino, J., et al. (2013). Criteria for high-quality assessment. Stanford, CA: Stanford Center for Opportunity Policy in Education.

High Quality Assessments

Validity and reliability (along with fairness) are considered two of the principles of high quality assessments. Though these two qualities are often spoken about as a pair, it is important to note that an assessment can be reliable (i.e., have replicable results) without necessarily being valid (i.e., accurately measuring the skills it is intended to measure), but an assessment cannot be valid unless it is also reliable. Other principles of high quality assessments are fairness—that an assessment is free from bias, and coherence—that each assessment is used in a manner consistent with its intended purpose.

Resources

USED created this non-regulatory guidance document for states on the peer review process, which includes examples of evidence for determining validity and reliability.

This Center on Standards and Assessment Implementation (CSAI) report provides a framework for understanding types of assessments and descriptions of technical considerations in assessments.

The Council of Chief State School Officers detail criteria for evaluating high-quality assessments in this document.

This CSAI toolkit is made up of modules designed to walk the participant through the assessment design process, and also includes definitions of key terms and concepts in the Introduction to Assessment Design section.

The National Center for Research in Vocational Education created this monograph chapter on comparing the quality and feasibility of assessments.

This article on criteria for high-quality assessment was produced by multiple education research organizations.

Researchers from the Center for Assessment produced this Guide to Evaluating College- and Career-Ready Assessments.

This collection of materials describes the process of evidence-centered design, including a Framework for Collecting Evidence for Test Validation.

Educational assessment design and evaluation are discussed in this technical report.

This instructional module defines standard error of measurement and provides exercises for its application.

Educational Testing Service produced this glossary of standardized testing terms.

The Standards for Educational and Psychological Testing were developed jointly by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education, and are considered a foundational text on this topic.

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION WestEd CRESST

CSAI Update is produced by the The Center on Standards and Assessment Implementation (CSAI). CSAI, a collaboration between WestEd and CRESST, provides state education agencies (SEAs) and Regional Comprehensive Centers (RCCs) with research support, technical assistance, tools, and other resources to help inform decisions about standards, assessment, and accountability. Visit www.csai-online.org for more information.

This document was produced under prime award #S283B050022A between the U.S. Department of Education and WestEd. The findings and opinions expressed herein are those of the author(s) and do not reflect the positions or policies of the U.S. Department of Education.

WestEd

WestEd is a nonpartisan, nonprofit research, development, and service agency that partners with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide, from Massachusetts, Vermont and Georgia, to Illinois, Arizona and California, with headquarters in San Francisco.

For more information, visit WestEd.org; call 415.565.3000 or, toll-free, (877) 4-WestEd; or write: WestEd / 730 Harrison Street / San Francisco, CA 94107-1242.