Considering Validity in Assessment Design

Yale Center for Teaching and Learning
https://ctl.yale.edu / https://ctl.yale.edu/AssessmentDesignValidity

Validity describes an assessment's successful function and results. Definitions and conceptualizations of validity have evolved over time, and contextual factors, populations being tested, and testing purposes give validity a fluid definition. Because scholars argue that a test itself cannot be valid or invalid, current professional consensus agrees that validity is the "*process of constructing and evaluating arguments for and against the identified interpretation of test scores and their relevance to the proposed use*" (AERA, APA, NCME, 2014). Professional standards recommend a variety of approaches and practices for measuring validity.

Instructors can improve the validity of their classroom assessments, both when designing the assessment and when using evidence to report scores back to students. When reliable scores (i.e. grades) are reported back to students, they must function as accurate feedback if they are to promote future progress or demonstrate degree of mastery. Validity of assessment ensures that accuracy and usefulness are maintained throughout an assessment.

Examples and Recommendations for Validity Evidence

Validity is the joint responsibility of test developers and the individuals that administer tests. Test developers typically suggest appropriate interpretations of scores for a specified population, and provide initial evidence to support their process and arguments. Test users and administrators then examine and gather evidence, making additional arguments suggesting how the interpretation, consequences, and use of the scores is appropriate, given the purpose of the instrument and the population being evaluated. Validity evidence must continually be gathered by both groups as the consequences of the use of the scores become more apparent.

Professional standards outline several general categories of validity evidence, including:

- **Evidence Based on Test Content** - This form of evidence is used to demonstrate that the content of the test (e.g. items, tasks, questions, wording, etc.) is related to the learning that it was intended to measure. For example, a classroom assessment should not have items or criteria that measure topics unrelated to the objectives of the course. Instructors can design a table of specifications for tests to ensure and communicate how the content of a course or unit is being measured. For larger scale assessments, a panel of experts is usually convened to design the table of specifications and review questions, to ensure that they are representative of the field of knowledge being measured.

- **Evidence Based on Response Processes** - This form of evidence is used to demonstrate that the assessment requires participants to engage in specific behavior deemed necessary to complete a task. For instance, if an item is designed to measure reading comprehension, validity addresses if participants are attempting to comprehend the passages, or can find the answer through other test-taking strategies. Instructors can gather evidence based on response processes by analyzing qualitative responses in order to identify how students arrived at answers or by asking students how they approached specific questions or problems. Larger scale testing requires a more systematic interviewing process and often relies on think-aloud protocols.

- **Evidence Based on Internal Structure** - This form of evidence demonstrates how the relationships between scores on individual test items align with the construct(s) that are being measured. For example, if an assessment is measuring both chemical bonding and chemical equilibrium, scores on different chemical bonding items should have a strong relationship with each other, and scores on different chemical equilibrium items should have a strong relationship with each other. Instructors can gather evidence based on internal structure by conducting item level analyses, or by calculating an exploratory or confirmatory factor analysis to determine how well similar items relate to each other.

- **Evidence Based on Relation to Other Variables** - This form of evidence demonstrates that a score measuring a defined construct relates to other scores measuring that same construct (or "convergent") and does not relate to other scores measuring different constructs (or "divergent"). For example, a score representing mathematical problem solving on one test should relate strongly with a score representing mathematical problem solving on another test. Similarly, mathematical problem-solving scores should not relate as strongly to scores that represent reading comprehension. Instructors can gather several different types of data about students' ability or knowledge of a particular construct in order to generate validity evidence based on relation to other variables. When developing a scale or test for educational research purposes, it is important to demonstrate how the scale relates to other established instruments that measure the same or similar constructs.

- **Evidence Based on Consequences of Testing** - This form of evidence describes the extent to which consequences of the use of the score are congruent with the proposed uses of the assessment. For example, an intended consequence of a score on a placement exam would be appropriate placement in introductory courses so that all students have the best opportunity to achieve success. But evidence would need to be gathered to determine that the scores correspond to success in the course. Additionally, unintended consequences such as decreased student motivation or intention to persist in a major could occur for students who score poorly on the initial exam. Instructors can gather evidence based on the consequences of testing by ensuring that scores on their assessments relate to intended future outcomes.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985, 1999, 2014). Standards for educational and psychological testing. Washington, DC.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. (1966, 1974). Standards for educational and psychological tests and manuals. Washington, DC.

Kane, M. (2013). The argument-based approach to validation. School Psychology Review 42(4):448-457.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 17-64). Washington, DC: National Council on Measurement in Education and the American Council on Education.

Kane, M. (1992). An argument-based approach to validity. Psychological Bulletin 112:527-535.
Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist 35(11):1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.