Quality Assessment:

Building Reliability and Validity While Diminishing Bias

Kathleen M. Marleneanu

University of Arkansas at Little Rock

Quality Assessment:

Building Reliability and Validity While Diminishing Bias

Assessment is commonly used in education, from the classroom to the national level. According to Popham (2014), educational assessment is a "formal attempt to determine students' status with respect to educational variables of interest" (p. 8), and is used for purposes such as identifying students' strengths and weaknesses, monitoring student progress, assigning grades, and determining instructional effectiveness. High quality assessments effectively fulfill these purposes in a trustworthy manner; and their effectiveness is largely determined by three key criteria: reliability, validity, and bias. In order to develop an assessment of high quality (as is my intention for my assessment concept of measuring the quality of museum public educational programs), reliability and validity need to be built, while bias should be diminished.

Reliability

Reliability is one of the criteria that is present in a quality assessment, and can be thought of as being equal to consistency (Popham, 2014). When a test consistently measures what its goal is to measure, it is considered reliable. The presence of reliability is essential because its lack signifies that the assessment results are variable and undependable, therefore "rendering them potentially meaningless" (Simonson et al., 2012, p. 268). To determine the strength of an assessment's reliability, four types of evidence can be looked at: stability reliability, alternate-form reliability, internal consistency, and interrater reliability.

Stability Reliability

Stability reliability refers to the consistency of results over time; and can be measured by administering the same test to the same learners on two separate occasions while making sure that "no significant performance-influencing events occur during the between-assessments interval" (Popham, 2014, p. 77). To interpret the test-retest results, Popham (2014) suggests two possible approaches: correlational (computing a correlation coefficient between the learners' test results from the two occasions) or classification-consistency (figuring out the percentage of learner classifications that were consistent over time).

Alternate-form Reliability

Alternate-form reliability also involves more than one test administration, but shows the consistency of results among two or more different (yet supposedly equivalent) forms of a test (Popham, 2014). In order to determine alternate-form reliability, these multiple test forms are given to the same learners with very little delay in between. The scores from each form are then compared, similar to stability reliability, by either computing a correlation coefficient or using a classification-consistency approach (Popham, 2014).

Internal Consistency

Unlike stability and alternate-form reliability, internal consistency can be determined by only giving one test one time. This is because it indicates the consistency in the way the items of a particular assessment function (Popham, 2014). Depending on the type of the assessment, different formulae can be used to compute its internal consistency. For example, for tests with dichotomous items (with only right or wrong answers), the Kuder-Richardson procedures are typically used; while for tests with polytomous items (which produce multiple scores), Cronbach's coefficient alpha is the most common approach (Popham, 2014).

Interrater Reliability

Suter (2012) explains that interrater reliability is also a measure of consistency among scores, but is used in evaluation circumstances involving observation, such as performance assessments. High interrater reliability indicates that observations of the same thing made by two or more raters are consistent. Similar to other forms of reliability, interrater reliability can be determined by a correlational approach, with the type of correlation maneuver used depending on the number of raters making observations (Suter, 2012).

Validity

Like reliability, validity is an essential characteristic of a quality assessment. Popham (2014) emphasizes this by stating that it is “hands down, the most significant concept in assessment” (p. 97), and explains that it concerns the accuracy of score-based inferences (i.e., the interpretation of what test results mean). Even when an assessment matches its objectives on the surface, the score-based inferences will be inaccurate if its results are not a true reflection of the learning that has occurred (Simonson et al., 2012). To help determine the truthfulness of this reflection, evidence of validity can be gathered relating to content, criterion, and construct.

Content-Related Evidence

Content-related evidence of validity describes the degree to which an assessment sufficiently represents the curricular aim (i.e., content standard) it is meant to measure (Popham, 2014). One way to gather this evidence is to “employ a set of test-development procedures carefully focused on assuring that the curricular aim’s content is properly reflected in the assessment procedure itself” (Popham, 2014, p. 105). Especially high care should be taken with high-stakes tests. The content validity of a high-stakes test is also often determined by an external review, in which a group of judges is gathered to rate the relationship between the curricular aim and content. Less formal external reviews are often used when assessments are of lower stakes (Popham, 2014).

Criterion-Related Evidence

Another form of validity evidence is called criterion-related, as it is collected when an assessment is used to predict how well learners will do on a subsequent criterion variable. Popham (2014) explains that to determine what percentage of learners’ performance on the criterion variable is meaningfully linked to their predictor test scores, the correlation coefficient between the two must be squared.

Construct-Related Evidence

The last type of validity evidence is construct-related. Suter (2012) explains that a construct is a "label for an abstract trait or ability (such as creativity or intelligence) that is only presumed to exist, since it cannot be measured directly (as weight or height can)" (p. 124). Gathering construct-related evidence of validity involves forming a hypothesis concerning learners' performance on the test in question, and then compiling empirical evidence that the construct is assessed accurately. This is customarily accomplished using an approach such as an intervention study, differential-population study, or related-measure study (Popham, 2014).

Bias

While the presence of validity and reliability in assessment is desirable, the presence of bias is not. Assessment quality decreases when aspects of it offend or unfairly penalize a group of learners due to characteristics such as race, ethnicity, language, religion, disability, or socioeconomic status (Popham, 2014). It is necessary to eliminate sources of bias when developing an assessment, because the validity of the assessment is threatened by potential bias due to its ability to distort the results (Chappuis, Commodore, & Stiggins, 2010).

One way Popham (2014) suggests to diminish assessment bias is by taking a judgmental approach. For high-stakes tests, a bias review panel could be put together, consisting of reviewers representative of those intended to take the assessment. This panel would answer questions to determine if assessment bias is possible on a per-item and overall basis. An empirical approach is especially helpful in high-stake situations involving a large quantity of learners. This method involves gathering "tryout evidence regarding the performance of different groups of students on individual items," and using differential item functioning procedures to appraise items for possible bias “for which there are substantial disparities between the performances of different groups” (Popham, 2014, p. 135).

In assessment situations that are not high-stakes, Popham emphasizes the importance of educators simply becoming sensitive to the existence and need to eliminate bias. Arter (2012) lists assessment traits to use to diminish bias: “Instructions are clear; Task is narrow enough; If choice, options are equivalent; Necessary resources are available to all; Success does not depend on unrelated skills; Success does not depend on cultural experience or language” (p. 33).

Assessment Concept

In the design of the assessments I create, I will use the concepts discussed above to strive to diminish bias while building reliability and validity. The goal of my assessment is to measure the quality of museum public educational programs developed by program assistants. This will be in the form of a performance assessment, with the program assistant creating and presenting an educational program which I (as the hypothetical educational supervisor) will observe and evaluate. Assessment instruments will be a rubric with four specific criteria (i.e., concerning demonstration of knowledge of resource, audience, techniques, and museum mission), as well as a short self-report inventory available for the public to complete at the end of each programs’ test run to help measure the its benefit to the public, including aspects such as audience attitudes, program interest, and understanding.

Reliability

In regard to reliability, I will make sure the rubric provides consistent measurement of program quality over time. To do this, I will gauge stability reliability by the test-retest format by using the same rubric to assess the same programs on two separate occasions (e.g., on back-to-back Saturdays), and then using either the correlation coefficient or classification-consistency approach. If the results are fairly unstable, I will reconsider and modify the rubric. I will also test the rubric's internal consistency by using Cronbach's coefficient alpha.

Validity

In terms of validity, the content of the assessments will ideally represent the content of the curricular aim (i.e., what the programs are meant to do), so I know the score-based inferences I make about the programs are accurate. To collect content-related evidence of validity, I will first develop the rubric and self-report inventory with care, giving deliberate consideration to whether the content reflects the aim. Secondly, I will utilize an informal external review by asking the museum's education directors to look over the items and share their conclusions.

Bias

To help eliminate potential bias such as unfairly penalizing a program assistant based on a personal characteristic (e.g., race, gender), I will make the rubric items specific while paying close attention to what I am creating, and also ask colleagues to review them. Additionally, to diminish bias when scoring the rubric (e.g., generosity, severity, or central-tendency errors), I will be aware of the possibility for such errors, and when creating the rubric, intermittently reverse the order of high/low positions on the scoring scale. I may also ask a qualified colleague to score the same programs, and then compare our scores to detect major differences. To avoid bias in the self-report inventory, I will make it anonymous, use simple and clear wording, and have the public answer by circling or checkmarks. I will also make sure the wording is inoffensive; and like with the rubric, ask colleagues to help review it for possible bias.

Conclusion

Reliability, validity, and bias are all significant aspects to consider when attempting to develop high quality assessments; and these aspects form an important relationship: The presence of validity indicates the presence of reliability and absence of bias (Popham, 2014). A variety of approaches are available to utilize to help create quality assessments at all levels, including those assessments in my concept. These methods emphasize the building of reliability and validity and the diminishing of bias.

References

Arter, J. (2012). *Creating and recognizing quality rubrics: A study guide from pearson assessment training institute*. Retrieved from http://ati.pearson.com/downloads/car_studyguide_forprinting_final.pdf

Chappuis, S., Commodore, C., Stiggins, R.J. (2010). *Assessment balance and quality: An action guide for school leaders* (3rd ed.). Retrieved from http://ati.pearson.com/downloads/chapters/ABQ%20Ch2.pdf

Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Simonson, M., Smaldino, S., Albright, M., & Zvacek, S. (2012). *Teaching and learning at a distance: Foundations of distance education* (5th ed.). Boston, MA: Pearson Education, Inc.

Suter, W. N. (2012). *Introduction to educational research: A critical thinking approach* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.