Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Copyright is retained by the first or sole author, who grants right of first publication to *Practical Assessment, Research & Evaluation*. Permission is granted to distribute this article for nonprofit, educational purposes if it is copied in its entirety and the journal is credited. PARE has the right to authorize third party reproduction of this article in print, electronic and database forms.

Volume 4, Number 2, November, 1994 ISSN=1531-7714

Questions To Ask When Evaluating Tests

Lawrence M. Rudner,
ERIC Clearinghouse on Assessment and Evaluation

The "Standards for Educational and Psychological Testing" established by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, are intended to provide a comprehensive basis for evaluating tests. This article identifies the key standards applicable to most test evaluation situations. Sample questions are presented to help in your evaluations.

TEST COVERAGE AND USE

There must be a clear statement of recommended uses and a description of the population for which the test is intended.

The principal question to ask when evaluating a test is whether it is appropriate for your intended purposes as well as your students. The use intended by the test developer must be justified by the publisher on technical grounds. You then need to evaluate your intended use against the publisher's intended use. Questions to ask:

1. What are the intended uses of the test? What interpretations does the publisher feel are appropriate? Are inappropriate applications identified?

2. Who is the test designed for? What is the basis for considering whether the test applies to your students?

APPROPRIATE SAMPLES FOR TEST VALIDATION AND NORMING

The samples used for test validation and norming must be of adequate size and must be sufficiently representative to substantiate validity statements, to establish appropriate norms, and to support conclusions regarding the use of the instrument for the intended purpose.

The individuals in the norming and validation samples should represent the group for which the test is intended in terms of age, experience and background. Questions to ask:

1. How were the samples used in pilot testing, validation and norming chosen? How is this sample related to your student population? Were participation rates appropriate?

2. Was the sample size large enough to develop stable estimates with minimal fluctuation due to sampling errors? Where statements are made concerning subgroups, are there enough test-takers in each subgroup?

3. Do the difficulty levels of the test and criterion measures (if any) provide an adequate basis for validating and norming the instrument? Are there sufficient variations in test scores?

RELIABILITY

The test is sufficiently reliable to permit stable estimates of the ability levels of individuals in the target group.

Fundamental to the evaluation of any instrument is the degree to which test scores are free from measurement error and are consistent from one occasion to another when the test is used with the target group. Sources of measurement error, which include fatigue, nervousness, content sampling, answering mistakes, misinterpreting instructions and guessing, contribute to an individual's score and lower a test's reliability.

Different types of reliability estimates should be used to estimate the contributions of different sources of measurement error. Inter-rater reliability coefficients provide estimates of errors due to inconsistencies in judgment between raters. Alternate-form reliability coefficients provide estimates of the extent to which individuals can be expected to rank the same on alternate forms of a test. Of primary interest are estimates of internal consistency which account for error due to content sampling, usually the largest single component of measurement error. Questions to ask:

1. How have reliability estimates been computed? Have appropriate statistical methods been used? (e.g., Split

half-reliability coefficients should not be used with speeded tests as they will produce artificially high estimates.)

2. What are the reliabilities of the test for different groups of test-takers? How were they computed?

3. Is the reliability sufficiently high to warrant using the test as a basis for decisions concerning individual students?

4. To what extent are the groups used to provide reliability estimates similar to the groups the test will be used with?

CRITERION VALIDITY

The test adequately predicts academic performance.

In terms of an achievement test, criterion validity refers to the extent to which a test can be used to draw inferences regarding achievement. Empirical evidence in support of criterion validity must include a comparison of performance on the validated test against performance on outside criteria. A variety of criterion measures are available, such as grades, class rank, other tests and teacher ratings.

There are also several ways to demonstrate the relationship between the test being validated and subsequent performance. In addition to correlation coefficients, scatterplots, regression equations and expectancy tables should be provided. Questions to ask:

1. What criterion measure has been used to evaluate validity? What is the rationale for choosing this measure?

2. Is the distribution of scores on the criterion measure adequate?

3. What is the overall predictive accuracy of the test? How accurate are predictions for individuals whose scores are close to cut-points of interest?

CONTENT VALIDITY

Content validity refers to the extent to which the test questions represent the skills in the specified subject area.

Content validity is often evaluated by examining the plan and procedures used in test construction. Did the test development procedure follow a rational approach that ensures appropriate content? Did the process ensure that the collection of items would represent appropriate skills? Other questions to ask:

1. Is there a clear statement of the universe of skills represented by the test? What research was conducted to determine desired test content and/or evaluate content?

2. What was the composition of expert panels used in content validation? How were judgments elicited?

3. How similar is this content to the content you are interested in testing?

CONSTRUCT VALIDITY

The test measures the "right" psychological constructs.

Intelligence, self-esteem and creativity are examples of such psychological traits. Evidence in support of construct validity can take many forms. One approach is to demonstrate that the items within a measure are inter-related and therefore measure a single construct. Inter-item correlation and factor analysis are often used to demonstrate relationships among the items. Another approach is to demonstrate that the test behaves as one would expect a measure of the construct to behave. For example, one might expect a measure of creativity to show a greater correlation with a measure of artistic ability than with a measure of scholastic achievement. Questions to ask:

1. Is the conceptual framework for each tested construct clear and well founded? What is the basis for concluding that the construct is related to the purposes of the test?

2. Does the framework provide a basis for testable hypotheses concerning the construct? Are these hypotheses supported by empirical data?

TEST ADMINISTRATION

Detailed and clear instructions outline appropriate test administration procedures.

Statements concerning test validity and the accuracy of the norms can only generalize to testing situations which replicate the conditions used to establish validity and obtain normative data. Test administrators need detailed and

clear instructions to replicate these conditions.

All test administration specifications, including instructions to test takers, time limits, use of reference materials and calculators, lighting, equipment, seating, monitoring, room requirements, testing sequence, and time of day, should be fully described. Questions to ask:

1. Will test administrators understand precisely what is expected of them?

2. Do the test administration procedures replicate the conditions under which the test was validated and normed? Are these procedures standardized?

TEST REPORTING

The methods used to report test results, including scaled scores, subtests results and combined test results, are described fully along with the rationale for each method.

Test results should be presented in a manner that will help schools, teachers and students to make decisions that are consistent with appropriate uses of the test. Help should be available for interpreting and using the test results. Questions to ask:

1. How are test results reported? Are the scales used in reporting results conducive to proper test use?

2. What materials and resources are available to aid in interpreting test results?

TEST AND ITEM BIAS

The test is not biased or offensive with regard to race, sex, native language, ethnic origin, geographic region or other factors.

Test developers are expected to exhibit a sensitivity to the demographic characteristics of test-takers. Steps can be taken during test development, validation, standardization and documentation to minimize the influence of cultural factors on individual test scores. These steps may include evaluating items for offensiveness and cultural dependency, using statistics to identify differential item difficulty, and examining the predictive validity for different groups.

Tests are not expected to yield equivalent mean scores across population groups. Rather, tests should yield the same scores and predict the same likelihood of success for individual test-takers of the same ability, regardless of group membership. Questions to ask:

1. Were the items analyzed statistically for possible bias? What method(s) was used? How were items selected for inclusion in the final version of the test?

2. Was the test analyzed for differential validity across groups? How was this analysis conducted?

3. Was the test analyzed to determine the English language proficiency required of test-takers? Should the test be used with non-native speakers of English?

RECOMMENDED READING

American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education (Joint Committee) (1985), *Standards for Educational and Psychological Tests,* Washington, DC APA.

Anastasi, A. (1988) *Psychological Testing* New York: MacMillan Publishing Company.

Messick, S. (1989) Validity. In R.L. Linn *Educational Measurement*, Third Edition. New York: MacMillan Publishing Company.

Uniform Guidelines on employee selection procedures (1978) *Federal Register,* 43, 38290-38315.

Descriptors: Ability; *Academic Achievement; *Evaluation Methods; Norms; *Predictive Validity; *Selection; Standards; Test Bias; Test Construction; Test Content; Test Reliability; Test Use; Test Validity; *Tests

Citation: Rudner, Lawrence M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation*, 4(2). Available online: http://PAREonline.net/getvn.asp?v=4&n=2.