

1. Validity

Validity is arguably the most important criteria for the quality of a test. The term validity refers to whether or not the test measures what it claims to measure. On a test with high validity the items will be closely linked to the test's intended focus. For many certification and licensure tests this means that the items will be highly related to a specific job or occupation. If a test has poor validity then it does not measure the job-related content and competencies it ought to. When this is the case, there is no justification for using the test results for their intended purpose. There are several ways to estimate the validity of a test including content validity, concurrent validity, and predictive validity. The face validity of a test is sometimes also mentioned.

2. Reliability

Reliability is one of the most important elements of test quality. It has to do with the consistency, or reproducibility, of an examinee's performance on the test. For example, if you were to administer a test with high reliability to an examinee on two occasions, you would be very likely to reach the same conclusions about the examinee's performance both times. A test with poor reliability, on the other hand, might result in very different scores for the examinee across the two test administrations. If a test yields inconsistent scores, it may be unethical to take any substantive actions on the basis of the test. There are several methods for computing test reliability including test-retest reliability, parallel forms reliability, decision consistency, internal consistency, and interrater reliability. For many criterion-referenced tests decision consistency is often an appropriate choice.

3. Fairness

The fairness of an exam refers to its freedom from any kind of bias. The exam should be appropriate for all qualified examinees irrespective of race, religion, gender, or age. The test should not disadvantage any examinee, or group of examinees, on any basis other than the examinee's lack of the knowledge and skills the test is intended to measure. Item writers should address the goal of fairness as they undertake the task of writing items. In addition, the items should also be reviewed for potential fairness problems during the item review phase. Any items that are identified as displaying potential bias or lack of fairness should then be revised or dropped from further consideration.



4. Legal Defensibility

For an exam program to have legal defensibility there must be evidence as to the test's quality that would stand up in a court challenge. You will need to be able to provide evidence that sound, professionally recommended guidelines were followed throughout the design, development, and maintenance of the exam program. Professional guidelines for testing are offered by the American Psychological Association (APA), American Educational Research Association (AERA), and the National Council on Measurement in Education (NCME). Studies should also be conducted to investigate and confirm that the test has reasonable degrees of validity, reliability, and fairness. Among the most important elements that courts look for are a well-conducted job analysis and strong content validity (that is, the items need to have a high degree of "job relatedness"). Finally, good documentation of the design, development, and analysis of the exam program should be collected and maintained.

