



Reliability & Validity

Nathan A. Thompson Ph.D.

Whitepaper-September, 2013

ASSESSMENT  SYSTEMS

6053 Hudson Road, Suite 345
St. Paul, MN 55125 USA

To begin a discussion of reliability and validity, let us first pose the most fundamental question in psychometrics: **Why are we testing people?** Why are we going through an extensive and expensive process to develop examinations, inventories, surveys, and other forms of assessment? The answer is that the assessments provide information, in the form of test scores and subscores, that can be used for practical purposes to the benefit of individuals, organizations, and society. Moreover, that information is of higher quality for a particular purpose than information available from alternative sources. For example, a standardized test can provide better information about school students than parent or teacher ratings. A pre-employment test can provide better information about specific job skills than an interview or a resume, and therefore be used to make better hiring decisions.

So, exams are constructed in order to draw conclusions about examinees based on their performance. The next question would be, just how supported are various conclusions and inferences we are making? What evidence do we have that a given standardized test can provide better information about school students than parent or teacher ratings? This is the central question that defines the most important criterion for evaluating an assessment process: **validity**. Validity, from a broad perspective, refers to the evidence we have to support a given use or interpretation of test scores. The importance of validity is so widely recognized that it typically finds its way into laws and regulations regarding assessment (Koretz, 2008).

Test score **reliability** is a component of validity. Reliability indicates the degree to which a person's test scores are stable – or reproducible – and free from measurement error. If test scores are not reliable, they cannot be valid since they will not provide a good estimate of the ability or trait that the test intends to measure. Reliability is therefore a necessary but not sufficient condition for validity.


Reliability

Reliability refers to the accuracy or repeatability of the test scores. There is no universally accepted way to define and evaluate the concept; classical test theory provides several indices, and item response theory drops the idea of a single index and reconceptualizes it as a conditional standard error of measurement. However, an extremely common way of evaluating reliability is the internal consistency index, called **KR-20** or **α (alpha)**.

The KR-20 index ranges from 0.0 (test scores are comprised only of random error) to 1.0 (test scores have no measurement error). Of course, because human behavior is generally not perfectly reproducible, perfect reliability is not possible; typically, a reliability of 0.90 or higher is desired for high-stakes certification exams. The relevant standard for a test depends on its stakes. A test for medical doctors might require reliability of 0.95 or greater. A test

for florists or a personality self-assessment might suffice with 0.80.

Reliability depends on several factors, including the stability of the construct, length of the test, and the quality of the test items. Reliability will be higher if the trait/ability is more stable (mood is inherently difficult to measure repeatedly), the test has more items (observations of performance), and better items. A test sponsor typically has little control over the nature of the construct – if you need to measure knowledge of algebra, well, that's what we have to measure, and there's no way around that.

 However, a test sponsor can obviously specify how many items they want, and how to develop those items.

Reliability and Test Length

There must be enough questions on an exam to obtain reliable scores and adequately span the content covered by the examination. Lengthy exams are typically required for examinations with an extremely wide range of content that must

be covered, such as an educational exam covering multiple subjects. That said, too many questions can make for an exhausting exam and do not necessarily provide a significant increase in reliability.

Certification exams often utilize 100 to 200 items (Raymond, undated), though some use substantially more if necessary to cover more content. Table 1 (from the Raymond article) describes the interaction of test length with score reliability and consistency of pass/fail decisions for a certification examination in the medical field.

The latter two columns of Table 1 present the percent of examinees for which the pass/fail decision would change if the test was shortened from the full length of 200 items. For example, reducing the test to 160 would result in 0.5% of examinees changing from a pass result to a fail result.

Table 1: Reliability and decision consistency

Items	Reliability	Pass to Fail	Fail to Pass
200	0.926		
180	0.920	0.3%	0.7%
160	0.914	0.5%	0.9%
140	0.903	0.8%	1.3%
120	0.888	1.1%	1.3%
100	0.864	1.1%	2.3%

This table demonstrates that 200 (or more) items utilized by many certification examinations are not necessary to produce reliable scores. The reliability of the test was 0.926 with 200 items, and eliminating 40 items reduced the reliability to only 0.914. Less than 1.5% of examinees would have different results; a similar or greater percentage could be expected to change by random effects if taking two forms of the tests (i.e. how examinees feel on a given day, specific content of test questions, etc.). This means that scores produced by a 160-item test have nearly the equivalent accuracy of a 200 item test. A 100-item test produces slightly less accuracy, but still has a reliability greater than 0.86, indicating adequate reliability for the stakes of many certification examinations.

Validity



Messick (1989) defines validity as an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of measurement. This definition suggests that the concept of validity contains a number of important characteristics to review or propositions to test and that validity can be described in a number of ways.

The modern concept of validity (AERA, APA, & NCME, 1999) is multi-faceted and refers to the meaningfulness, usefulness, and appropriateness of inferences made from test scores. Validity is conventionally defined as the extent to which a test measures what it purports to measure, and test validation is the process of gathering evidence to support the inferences made by test scores. Validation is an ongoing process which—incidentally—makes it difficult to know when one has reached a sufficient amount of validity evidence to interpret test scores appropriately.

First of all, **validity is not an inherent characteristic of a test.** It is the reasonableness of using the test score **for a particular purpose or for a particular inference.** It is not correct to say a test or measurement procedure is valid or invalid. It is more reasonable to ask, “Is this a valid use of test scores or is this a valid interpretation of the test scores?” Test score validity evidence should always be reviewed in relation to how test scores are used and interpreted.

Secondly, **validity cannot be adequately summarized by a single numerical index like a reliability coefficient or a standard error of measurement.** A validity coefficient may be reported as a descriptor of the strength of relationship between other suitable and important measurements. However, it is only one of many pieces of empirical evidence that should be reviewed and reported by test score users. Validity for a particular test score use is supported

through an accumulation of empirical, theoretical, statistical, and conceptual evidence that makes sense for the test scores.

Thirdly, **there can be many aspects of validity dependent on the intended use and intended inferences to be made from test scores.** Scores obtained from a measurement procedure can be valid for certain uses and inferences and not valid for other uses and inferences. Ultimately, an inference about probable job performance based on test scores is usually the kind of inference desired in test score interpretation in today's test usage marketplace. This can take the form of making an inference about a person's competency measured by a tested area.



Validity Evidence in Certification Testing

A **job analysis** study provides the vehicle for defining the important job knowledge, skills, and abilities (KSA) that will later be translated into content on a certification exam. During a job analysis, important job KSAs are obtained by directly analyzing job performance of highly competent job incumbents or surveying subject-matter experts regarding important aspects of successful job performance. The job analysis generally serves as a fundamental source of evidence supporting the validity of scores for certification exams. After important job KSAs are established, subject-matter experts write test items to assess them. The end result is the development of an **item bank** from which exam forms can be constructed.

The quality of the item bank also supports test validity. There should be evidence that each item in the bank actually measures the content that it is supposed to measure; in order to assess this, data must be gathered from samples of test-takers. After items are written, they are generally **pilot tested** by administering them to a sample of examinees in a low-stakes context—one in which examinees' responses to the test items do not factor into any decisions regarding competency. After pilot test data is obtained, a **psychometric analysis** of

the test and test items can be performed. This analysis will yield statistics that indicate the degree to which the items measure the intended test content. Items that appear to be weak indicators of the test content generally are removed from the item bank or flagged for **item review** so they can be reviewed by subject-matter experts for correctness and clarity.

Standard setting also is a critical source of evidence supporting the validity of certification (i.e. pass/fail) decisions made based on test scores. Standard setting is a process by which a mastery test score (or cutscore) is established; a mastery test score is the point on the score scale that differentiates between examinees that are and are not deemed competent to perform the job.

In order to be valid, the cutscore cannot be arbitrarily defined. Two examples of arbitrary methods are the quota (setting the cut score to produce a certain percentage of passing scores) and the flat cutscore (such as 70% on all tests). Both of these approaches ignore the content and difficulty of the test.

Instead, the cutscore must be based on one of several well-researched criterion-referenced methods from the psychometric literature. There are two types of criterion-referenced standard-setting procedures (Cizek, 2006): *examinee-centered* and *test-centered*. The Contrasting Groups method is one example of a defensible examinee-centered standard-setting approach. This method compares the scores of candidates previously defined as Pass or Fail.

Obviously, this has the drawback that a separate method already exists for classification. Moreover, examinee-centered approaches such as this require data from examinees, but many testing programs wish to set the cutscore before publishing the test and delivering it to any examinees. Therefore, test-centered methods are more commonly used in credentialing.

The most frequently used test-centered method is the Modified Angoff Method (Angoff, 1971) which requires a committee

of subject matter experts (SMEs). These SMEs begin by discussing and establishing the concept of a minimally competent candidate (MCC). An MCC is a person who should barely pass the exam and earn the certification, but is not an expert. The SMEs then proceed through each item, providing an estimate of the percentage of MCCs that should answer each item correctly. A rating of 100 means that the item is expected to be so easy that all MCCs will answer correctly, while a rating of only 40 would indicate a very difficult question. An average item might have a rating of 70. After ratings are completed, the results are analyzed for inter-rater reliability. Items that show large disparity in ratings are discussed among the SMEs, and SMEs are given an option to change their initial rating after the discussion. The average of the final ratings is calculated, and this calculation serves as the initial cut score recommendation. Taking into account uncertainty in the SMEs ratings, the standard error (SE) of the SME's ratings can be used in conjunction with the mean Angoff rating to determine a range of possible cut scores (e.g., mean \pm 1 SE). The cutscore recommendations are discussed in light of examinee results (when available), and a final recommendation is confirmed as the cutscore for the exam form



Summary

In conclusion, reliability and validity are two essential aspects in evaluating an assessment process, be it an examination of knowledge, a psychological inventory, a customer survey, or an aptitude test. Validity is an overarching, fundamental issue that drives at the heart of the reason for the assessment in the first place: the use of test scores. Reliability is an aspect of validity, as it is a necessary but not sufficient condition. Developing a test that produces reliable scores and valid interpretations is not an easy task, and progressively higher stakes indicate a progressively greater need for a professional psychometrician.