Chapter 2
Issues to Consider When Evaluating "Tests"

Michael J. Roszkowski and Scott Spreat

Proper measurement of various client characteristics is an essential component in the financial planning process. In this chapter, we will present an overview of how to go about evaluating the quality of a "test" that one may be using or considering, including the meaning of important underlying concepts and the techniques that form the basis for assessing quality. We will present the reader with the tools for making this evaluation in terms of commonly accepted criteria, as reported in *Standards for Educational and Psychological Testing,* a document that is produced jointly by the American Educational Research Association (AERA), the National Council on Measurement in Education (NCME), and the American Psychological Association (APA) (American Psychological Association, American Educational Research Association, and the National Council on Measurement in Education (Joint Committee), 1985). Conceptually similar guidelines for test development and usage are published by the International Test Commission (ITC), a multinational association of test developers, users, and the agencies charged with oversight of proper test use (see http://www.intestcom.org/).

What Is a Test?

A commonly accepted definition of measurement in the social sciences and business, first formulated by Stevens (1946, p. 667), is that it is "the assignment of numerals to objects or events according to some rule." Strictly speaking, a test is a procedure that allows for the evaluation of the correctness of something or how much of a given quality it reflects. Broadly defined, however, a "test" can be any systematic procedure for obtaining information about a person, object, or situation. Thus, a test can be any scale meant to gauge some quality, and this term can include instruments such as questionnaires, inventories, surveys, schedules, and checklists in addition to the (dreaded) assessment devices that first come to mind when one hears

M.J. Roszkowski (✉)
Institutional Research, La Salle University, 1900 West Olney Avenue, Philadelphia, PA 19141, USA
e-mail: roszkows@lasalle.edu

J.E. Grable et al. (eds.), *Financial Planning and Counseling Scales*,
DOI 10.1007/978-1-4419-6908-8_2,

the word "test." *As used here, a test refers to any procedure that collects information in a uniform manner and applies some systematic procedure to the scoring of the collected data*.

Advantage of Tests. In contrast to informal interviews, tests are generally a bit more standardized. One can standardize a test by using the same instructions, same questions, and same response possibilities. The standardization is the quality of tests that creates consistency in collecting the data and scoring and allows for the scores to be meaningfully compared across different administrations of the instrument.

Level of Measurement. The type of information available on a test is determined by the level of measurement of the question: (a) nominal, (b) ordinal, (c) interval, and (d) ratio. Nominal variables permit classifications only (e.g., sex: male versus female); there is no hierarchy of any sort to the answers. Ordinal variables allow one to rank objects as to whether one object has more (or less) of a given characteristic, but we can't tell how much more of that quality one thing has over the other (e.g., gold, silver, and bronze medals for an Olympic running event, without consideration of the time differences). On interval-level variables, the differences between the ranked objects are assumed to be equal (e.g., Fahrenheit temperature: 70 versus 80 is the same as 80 versus 90, namely 10°), and so one can consider size of differences when comparing groups or events. The most informative are variables on the ratio scale, where not only are the differences between any two objects of the same magnitude, but the scale has an absolute zero point (e.g., temperature measured on a Kelvin scale). Scales of measurement are important because the type of statistical tests one can perform on the test data is a function of the level of measurement (as well as other factors), although the same techniques apply to interval as to ratio scales (the latter are quite rare in the behavioral sciences).

Interpretation of Scores. Test scores are yardsticks. The ruler can be created on the basis of a criterion-reference or a norm-reference. Criterion-referenced tests describe performance against some pre-set absolute standard (proficiency or mastery). Norm-referenced tests, on the other hand, provide information in relative terms, giving you the person's position in some known distribution of individuals. In other words, the score has no meaning unless it is benchmarked against how other people like the ones being tested perform on this test. In general, the closer the match between the characteristics of the reference group and the person you are assessing, the more confidence one can have that the test will produce meaningful comparative scores. It is therefore necessary for the test user to compare the reference group with the individuals being tested in terms of demographic characteristics such as sex, age, ethnicity, cultural background, education, and occupation. For example, a test on knowledge of investment concepts requiring command of the English language would not be an appropriate measure of such knowledge for a newly arrived immigrant since he or she would be benchmarked against native speakers.

Published or Unpublished. Tests are either "published" or "unpublished." Published means that the instrument is available commercially from a company that charges for its use. In contrast, generally unpublished tests are not sold. Typically, they are found in journal articles, books, and other types of reports. However, their use too can be restricted based on the copyright laws, in which

case it is prudent for the test user to obtain written permission to administer the test, unless a blanket permission has been given by the copyright holder. At the very least, the test user has an ethical obligation to give proper credit to the test author in whatever written documents result from the test's use (for guidelines, see http://www.apa.org/science/programs/testing/find-tests.aspx#). Although the information that will allow one to evaluate the quality of a published test is more readily available (typically in test manuals), the principles for evaluating the quality of a test remain the same for both published and unpublished instruments.

Many published tests, including ones meant for use by businesses, have been evaluated by measurement professionals for the Buros Institute of Mental Measurements, and these reviews can provide an additional source for the decision about whether to use the particular test (see http://buros.unl.edu/buros/jsp/search.jsp). Given the choice of a published test with known (good) properties and an unpublished one without any information about its psychometric quality, one would be wise to select the former, unless there are extenuating circumstances. One such factor may be that there is no published counterpart to the test one is seeking.

Is It a Good Test?

When using a test, the typical person is concerned about whether the test is a "good" one. However, a test can be good in one respect and yet bad in another, and good for one purpose and bad for another. To professionals who design tests, the quality is determined in terms of the concepts of "reliability" and "validity." A good test is one that is both reliable and valid. Without these two qualities, it cannot provide the information one seeks to learn by using it. Stated succinctly, a test is reliable if it measures something consistently, and it is valid if it actually measures what it claims to assess. It may help to think of reliability as *precision* and validity as *accuracy*. With a reliable test, a person would be expected to obtain a similar score if he or she were to take the test again, provided that there were no changes in his or her circumstances that would be expected to produce a change. The measurement would be precise because the score did not change. However, even if the score were to be identical on the two occasions, it does not necessarily mean that the test is valid. The test could be measuring some characteristic other than the intended one, even if it measures it consistently. So, if the test is not capturing the information you are seeking, it cannot be accurate. A test that is not reliable cannot be valid. However, one that is reliable may not be valid for its intended purpose. In other words, reliability is a necessary but not sufficient requirement for having validity.

It is incumbent on the test user to ensure that the test has adequate reliability and validity for the planned use. Although it is necessary that there be supporting evidence of both reliability and validity, the latter is clearly the "bottom line" when it comes to test selection. If a test is published, then the test distributor should make available such evidence and allow for an independent scrutiny and verification of any claims. Most test users expect that a test publisher will provide a manual that

instructs them on how to administer, score, and interpret the test results. Less recognized is the need for the publisher to make available the supporting technical background information necessary for making an informed decision as to whether the test is suitable for one's intended purpose. The latter can be either a section of a general manual or a separate document. Often, with unpublished tests, this evidence may not exist at all or it may not be readily available or open to independent scrutiny (Hinkin, 1995; Hogan & Angello, 2004). For example, an analysis of the American Psychological Association's *Directory of Unpublished Experimental Measures* revealed that only 55% reported any sort of validity information (Hogan & Angello, 2004). When using tests without this information, one is essentially sailing into uncharted waters and should proceed with caution.

Reliability

The old carpentry adage of "measure twice, cut once" comes to mind. A test is reliable to the extent that an individual gets the same score or retains the same general rank on repeated administrations of the test. A more professional statement of the concept is that reliability refers to the repeatability of measurement or, more specifically, the extent to which test scores are a function of systematic sources of variance rather than error variance (Thorndike & Hagen, 1961). A test score comprises the test taker's true level of the characteristic that the test intends to measure, plus or minus some error component. In other words, the observed test score does not reflect precisely the degree to which the person possesses that characteristic because this error element accompanies any test score. Any given score may overestimate or underestimate an individual's true level. However, tests that have acceptable levels of reliability tend to have lower amounts of error than do tests with less acceptable levels of reliability. Therefore, a reliable test yields higher quality information to the user than does a less reliable test because it is more precise.

What Produces This Error? It could be due to various random events. On tests of ability, people frequently attribute things to "luck" when guessing on an answer. Because luck is a random event that can be either good or bad, the actual obtained score will fluctuate above or below the true ability score. Other factors leading to error include misinterpretation of the instructions, one's mood, distractions, and various other idiosyncratic conditions under which the testing occurred that particular day. A major reason for the error is unclear wording of the questions (DeVellis, 2003). For instance, it is known that statements containing "double negatives" often confuse people (e.g., having to agree or disagree with the statement: "I am not incompetent in money management"). Also problematic are "double barreled" items, which contain two-part statements, such as "I am good at saving money and making investments." If one requires a respondent to agree or disagree with this statement, it is a frustrating task for the test taker. One can be good at saving and bad in investing, so only a person who is good or bad at both tasks will be able to correctly answer it. Often, the basis for the agreement/disagreement is one part of

the question and it is impossible to tell with which part of the statement the person is agreeing or disagreeing. That part may not be the same one if the test is given again, hence the inconsistency.

What Is an Acceptable Level of Reliability? Reliability is typically expressed as a correlation coefficient, which is called a "reliability coefficient" when used for this purpose, although there are other statistics with which to quantify reliability. Typically reported are Pearson correlations and Spearman correlations; the difference between them is that the latter is used with ordinal-level data, whereas the former is most appropriate with interval- or ratio-level data. The values of both types of correlation coefficients may range from –1.0 to +1.0, and reliability coefficients vary along this same dimension. A reliability coefficient of +1.0 indicates perfect reliability; the test yields exactly the same rankings on repeated administrations. Conversely, a reliability coefficient of 0 indicates that the test scores obtained on repeated administrations are entirely unrelated (and therefore perfectly unreliable). Negative reliability coefficients are possible but exceedingly rare and would suggest a major problem with a test. For all practical intents and purposes, reliability varies on a dimension from 0 to 1.0, with higher values being suggestive of greater test reliability.

The literature regarding test and scale construction suggests that an acceptable level of reliability is a function of the intended use of the test results. If a test is to be used to make decisions about an individual, it is important for that test to be highly reliable. This need for higher levels of reliability goes up as the risk associated with a poor decision based on the test increases. For example, if a stress test was being administered to determine the need for open heart surgery, one would certainly hope that the stress test was highly reliable. The same logic pertains in the investment world, even though life and death might not be in the balance. If one were attempting to assess the risk tolerance of an investor in order to develop an investment plan, the financial advisor would certainly want to know that the risk-tolerance scale yielded highly reliable data before taking action based on the results of the test. The consequences of developing such an investment plan based on unreliable risk-tolerance data could be economically catastrophic to both the investor and the advisor.

Nunnally (1967) recommended that when a test or scale is used to make decisions about individuals, the reliability coefficients should be at least 0.90. It has been pointed out by others that while it is possible to get such reliability for tests measuring intellectual skills and knowledge (in the professional jargon, so-called cognitive domains), it is much more difficult to achieve this level of precision with tests assessing personality and feelings (known to testing professionals as "affective" variables). Consequently, others are somewhat less conservative, suggesting that a reliability coefficient of 0.80 is acceptable for a test or scale that will be used for making decisions about an individual (Batjelsmit, 1977). The US Department of Labor (Saad, Carter, Rothenberg, & Israelson, 1999) suggests the following interpretations: 0.90 or higher = excellent, 0.80 to 0.89 = good, 0.70 to 0.79 = adequate, and 0.69 and below = may have limited applicability. Obviously, both the test taker and the researcher are safer with higher levels of

reliability, and for that reason readers should follow Nunnally's advice whenever possible. It should be noted that when test or scale data from a single individual are supplemented by other forms of information regarding the characteristic that is being measured by the scale, these stringent reliability requirements can be relaxed somewhat.

Lower levels of reliability are acceptable when a test or scale is to be used for research purposes or to describe the traits of groups of individuals. The risks associated with a single flawed piece of datum are minimized because the random errors of measurement tend to cancel each other out. Overestimates on some people are balanced and therefore canceled out by underestimates on other people. When data are aggregated for research purposes, the main risk associated with lower reliability is that it will become harder to detect "truly" significant differences via statistical analysis. (The failure to detect "truly" significant differences is sometimes called a Type 2 error.)

Approaches to Assessing Reliability

Reliability can be estimated from multiple administrations of a test to the same group, and it can also be estimated from one administration of a test to a single group. Each approach will be discussed below.

Test-Retest Reliability. Test-retest reliability is conceptually derived from the basic notion of reliability. That is, if a test is administered two times to the same group of individuals, it should generally yield the same results. In order to estimate test-retest reliability, a test or scale is administered twice within a relatively short time period (typically two weeks) to the same group of individuals. The scores obtained from test administration #1 are correlated with those of administration #2 to yield a reliability index that will typically range from 0 to 1.0. Because it measures stability over time, this reliability estimate is sometimes called a "coefficient of stability" or "temporal stability."

It should be noted that while test-retest methodology is an entirely acceptable means with which to estimate reliability, it does contain threats that may confound results. For example, if the same test is administered twice, an individual may remember certain items, and this will give the same answer for that reason. Further, if the time period between test administrations is more than a couple of days, the test taker's performance may be influenced by learning or development. External events entirely unrelated to the test but that occur between the two administrations can impact the assessment of reliability. Suppose one were assessing the reliability of a new risk-tolerance scale, and there was a major stock market crash immediately after the first test administration. One would have to assume that this event would affect the results obtained on the second test administration and, in turn, render the reliability estimate questionable. While test-retest reliability is an acceptable form of reliability, it can be negatively affected by memory for items or external events unrelated to the test.

Another word of caution must be given here. It is necessary to distinguish between "absolute stability" and "relative stability" (Caspi, Roberts, & Shiner, 2005). Absolute stability deals with consistency of the actual scores when measured across occasions. It addresses the question: "Did the individual score the same or differently each time?" Relative stability involves the consistency of an individual's rank order within a group when tested multiple times. In this case, the question asked is: "Did the individual maintain her or his position in a group?" The absolute level of the characteristic measured can change over time, but the rank-ordering of individuals can remain the same; so a change in absolute stability does not preclude rank-order stability. For example, Roszkowski and Cordell (2009) studied the temporal stability of financial risk tolerance in a sample of students enrolled in an undergraduate financial planning program, finding that relative stability was around 0.65 after a period of about a year and a half. In terms of absolute stability, however, there was an average increase of about eight points.

The traditional Pearson product moment and Spearman rank-order correlation procedures that are taught in all introductory statistics courses have a bit of a limitation when used as reliability coefficients. While both of them are sensitive to changes in relative performance, they are insensitive to a uniform change that affects all test takers. Thus, if each test taker scored exactly five points more on the second administration of some test, the reliability would appear to be perfect (1.0), even though no one had the same exact score. While this is generally not a significant threat, some psychometricians recommend the use of the intraclass correlation for the estimation of absolute reliability coefficients because it is sensitive to those situations in which all test takers improve or decline by the same amount, and this will reduce the reliability estimate of the instrument accordingly.

Internal Consistency Reliability. Because of the above-described threats to estimating reliability via test-retest methodology, as well as the cost of a second test administration, an alternate procedure was sought that would enable one to estimate reliability from a single test administration. It was reasoned that a sufficiently lengthy scale might be divided in half, and the two resultant scales (for example, odd-numbered items and even-numbered items) might be correlated. This approach to assessing reliability is called "split-half reliability." It is meant to correct for the threats associated with the test-retest method. It should be noted that reliability is partially a function of the length of a scale (Stanley, 1971). Dividing the scale in half thus reduces the scale length and, in turn, the reliability of the instrument. When using split-half reliability, it is therefore necessary to correct for the test length via the Spearman–Brown formula in order to obtain an estimate that pertains to the full length of the test.

Because the split-half reliability may depend on which items are placed into which half, psychometricians have also developed mathematical formulas to determine the hypothetical "average" reliability that would have resulted had all possible different combinations of items been explored in such split halves. One such formula is known as the "Kuder–Richardson 20," but it is only appropriate for nominal-level items that are scored dichotomously, such as yes-no or correct-incorrect. Cronbach's "alpha" (Cronbach, 1951) is a more versatile and sophisticated variant

of this approach, which can be used for items that are scored with any number of continuous-answer options. Cronbach's alpha too is the average correlation for all possible split halves of a given test (not just the odds versus evens), corrected for scale length. There is some debate among measurement specialists about whether the variable has to be on an interval scale of measurement to use Cronbach's alpha or whether it can be legitimately applied to ordinal-level data as well.

Internal consistency reliability is the most frequently reported estimate of reliability, probably because it requires only one sample for its derivation, but it has been argued that internal consistency approaches to estimating reliability can be misleading and not entirely useful. Cattell (1986) contended that it is possible for a test to yield an excellent test-retest reliability estimate and yet be much less impressive in terms of an estimate based on internal consistency. This assertion is true, but any scale that yielded such results would be in violation of basic guidelines for test construction. Specifically, the test or scale would have to have been constructed of unrelated items. An example of such a scale might be a two-item scale consisting of the items: (1) What is your cholesterol level? and (2) How much did you invest with Bernie Madoff? While test-retest reliability is likely to be good on this two-item scale, the independence (unrelatedness) of the two items would suggest a likely poor internal consistency. One must wonder why anyone would attempt to build a scale of such unrelated items.

However, combining unrelated items does sometimes make sense under some circumstances, as when an "index" is developed. It is necessary to differentiate between a "scale" and an "index." Although frequently the two terms are used interchangeably, conceptually there is a distinction between the two. The similarity is that both are composite measures, consisting of a sum of some sort of multiple items. However, a scale consists of the sum of *related* items attempting to measure a *unidimensional* (*single*) construct. In contrast, an index is a summary number derived from combining a set of possibly *unrelated* variables to measure a *multidimensional* construct. Good *scale* construction generally strives for items that are modestly related to each other but strongly related to the total scale score, and internal consistency reliability is a relevant basis for estimating reliability of scales. However, in concert with Cattell's (1986) reservations, internal consistency may not be an appropriate standard for the estimation of reliability of an index because it may comprise unrelated items, each of which measures a *different* attribute or dimension.

One example of an index is a "life events' scale that measure how much stress one is experiencing, based on a count of recent traumatic events occurring in one's life, such as the death of a spouse, a job loss, moving, etc. Although the events may be related in some instances (e.g., my wife died; so I sold our house and moved to another city to be closer to the kids, and I got a new job there), but it is unreasonable to expect that these events necessarily have to typically be connected to each other. Even though the individual components may not be highly correlated, the combination does produce meaningful information, such a predicting stress-related illness. Another example is socioeconomic status, which is derived from consideration of one's occupation, income, wealth, education, and residence (admittedly, here, there is some correlation between the items).

Inter-rater Reliability. On certain types of tests, a rater has to determine subjectively the best option among those offered to describe someone on the items constituting the test (in this case, a rating scale). For example, one's job performance may be appraised by a supervisor, subordinates, peers, and clients. Differences in ratings among the raters will produce variations in test scores. For scales that require a subjective judgment by raters, inter-rater reliability thus becomes an issue. Inter-rater reliability indicates the degree of agreement between different judges evaluating the same thing. Typically, inter-rater reliability coefficients are smaller than test-retest or internal consistency reliability coefficients. One has to be especially careful in how the questions are phrased in ratings scales to be used by multiple informants so as to reduce any chance that the raters will interpret things differently. Although some differences are due to the perspectives of the different classes of raters, frequently the reliability can be improved if raters are trained on how to go about this task (Woehr & Huffcutt, 1994).

In addition to the Pearson correlation and the Spearman correlation, the procedures often employed to determine this type of reliability include Cohen's kappa, Kendall's coefficient of concordance, and the intraclass correlation. If the things to be rated cannot be ranked because they constitute discrete categories (nominal level measurement), then Cohen's kappa is appropriate. Kendall's procedure indicates the extent of agreement among judges when they have to rank the people or objects being rated (ordinal level measurement). When the data are in a form in which the scores indicate equal differences (interval level measurement), the intraclass coefficient may be used to check on the inter-rater reliability. There are six versions of the intraclass coefficient, and the results may differ depending on which version was used in the calculation.

Standard Error of Measurement

As noted earlier, an individual′s performance on a test or a scale is a function of his or her skill plus or minus chance factors. Sometimes these chance events will inflate a score and deflate it other times. Chance, of course, is really measurement error, and measurement error detracts from the reliability of a test or scale. It is possible to estimate the extent to which an individual's score would be expected to fluctuate as a function of such random events (unreliability) via a statistic called the Standard Error of Measurement (SEM).

If one were to administer the FinaMetrica Risk-Tolerance scale an infinite number of times to an individual, we would expect the scores for that individual to vary somewhat. In fact, we would expect these obtained scores to take on the form of the bell curve (a normal distribution). Most of the time, the individual would score in the middle of the range of scores, with more divergent scores being predictably less frequent. The standard deviation calculated from that distribution is the SEM. Of course, we can't administer a test an infinite number of times; so one cannot directly calculate the SEM. It must be estimated using information about the reliability of

the test and the standard deviation of the distribution of scores of people who took the test.

The utility of the standard error of measurement is that it permits the user to establish confidence intervals around a given test score, based on the known properties of the normal curve. It is commonly known that 95% of all scores fall within ± 1.96 standard errors of measurement of the mean. This fact can be used to establish confidence intervals surrounding obtained scores. Consider the FinaMetrica measure of financial risk tolerance (www.riskprofiling.com), which has an SEM of 3.16. Multiplying 3.16 by 1.96, we get 6.2. Thus, we can be 95% confident that with an obtained score of 50, the true score will be within plus or minus 6.2 points of the observed score. That is, the true (error free) score will be in a range somewhere between about 43.8 and 56.2. Reschly (1987) referred to this range as the "zone of uncertainty." We can be 95% confident that the individual's true score lies within that range, but we can't be certain where. If we wanted to be 99% confident, the zone of course would be much wider. Some test makers suggest reporting confidence ranges rather than individual scores because individual scores imply a level of precision that is beyond the capabilities of most psychological instruments.

Length of Test and Reliability

Reliability is at least partially a function of the number of items within a scale (Stanley, 1971). A longer scale, in general, will be more reliable than a shorter scale. The logic behind this is similar to, but somewhat more complicated than, the logic that underlies the sampling strategies used in polling. In polling, the margin of error (which is a measure of reliability) varies primarily by the number of persons being polled. The margin declines as the number of persons polled goes up (to a point of diminishing returns). Similarly, by increasing the number of items used to measure a given construct, one increases the likelihood of getting a precise measure. Precision is increased by using longer scales.

So predictable is this relationship between scale length and scale reliability that a mathematical formula, the Spearman–Brown prophecy, can predict the degree to which reliability can be increased by simply adding items of equal or similar reliabilities to the test. For example, Spreat and Connelly (1996) used the Spearman–Brown prophecy formula to illustrate how an insufficiently reliable psychological test could be made sufficiently reliable by simply adding more items. They reported that quadrupling the length of the Motivation Assessment Scale would increase the reliability sufficiently to support independent decision making based on the results of that scale. In other words, an unreliable scale could be made into a reliable one.

Achieving the correct balance in the number of items can be quite tricky. Too few, and one faces the prospect of low reliability; too many, and there is the threat of fatiguing the test taker. It could be a good test from a psychometric perspective, but

impractical because few people would want to take it given the time commitment required. There is a family of models called "item response theory" (IRT) that is used to develop and to score ability and achievement tests, such as the Certified Public Accountant (CPA) exam. Tests developed with IRT models can be shorter when administering through adaptive testing, where the questions are tailored to the test taker's ability level. All examinees are first given questions of medium difficulty. If an examinee does well on the items of intermediate difficulty, she or he will then be presented with more difficult questions. If he or she performs poorly at the intermediate level, easier questions are administered. This tailoring is possible because on the basis of pre-testing, questions are classified on three dimensions: (1) difficulty (how hard or easy is the question), (2) discrimination (the degree to which the question differentiates between people who know the material and those who do not), and (3) guessing (the probability that the question can be answered correctly without knowledge of the material).

In the CPA exam for instance, there are three different multiple choice versions of the test, called "testlets." Depending on how well they score on the first "testlet," which has a medium level of difficulty, the candidates next take either another testlet of medium difficulty (if did poorly) or a more difficult testlet (if did well). The third testlet is again of either medium or high difficulty, and the decision as to which version will be administered depends on the person's performance on the two previous testlets. Thus, two candidates can answer the same number of questions correctly, but get different scores because the questions on their respective exams were weighted differently on the basis of these three characteristics. IRT allows one to estimate an individual's proficiency on the latent trait (i.e., construct) that the test is designed to measure as well as the standard error of measurement at that level of proficiency. This is a departure from the notion of standard error measurement found in what is now known as the "classical psychometric model," where the standard error is the same across the different test scores. In IRT, the standard error of estimate can be a continuous function that may vary over the possible range of scores. The drawback to developing tests using IRT is that it requires huge samples.

Validity

From a validity perspective, one first needs to compare the recommended use of the test with one's intended purpose. Some authorities argue that it is preferable to speak about the validity of test *scores* rather than the test *itself* because validity is a property of the test scores in a particular context. The same test can be valid for one purpose and invalid for another. Likewise, it may be valid for one type of individual and not for another. Consequently, proof of validity for different applications needs to be produced. The evidence for validity can be gathered using three strategies: content validation, construct validation, and criterion-related validation.

Approaches to Assessing Validity

Content Validity. Content validity provides evidence that the scope of the test is sufficient, so that it covers comprehensively the attribute it intends to measure. For example, a test meant to measure the competency to be a CPA would be judged on whether it assesses all the important characteristics that make for a successful accountant. Generally, subject matter experts determine if a test has content validity by comparing the items in the test with the known universe of such items and checking for item clarity. For example, practicing CPAs and staff at the AICPA review the questions on the CPA exam. According to Rudner (1994), the questions to ask oneself when evaluating a test for content validity include the following:

1. Was a rational approach used to make sure that the items provide sufficient coverage of the desired attributes to be tested?
2. Is there a match between what one wants to test and the items on the test?
3. What were the qualifications of the panel of experts that examined the adequacy?

Unfortunately, there does not exist a generally accepted quantitative index of content validity; it is matter of judgment by experts. It is important to understand that content validity is not the same thing as what is called "face validity." Face validity does not involve any type of scientific procedure for validation. If someone states that a test has face validity, all it means is that the test appears (on the face of it) to be valid. While this is not a proof of validity, it sometimes helps if the potential test user tries out the test herself or himself in order to see whether it "feels right" since it has been shown that having face validity may increase a test taker's acceptance of the results. Hinkin (1995) found that only about 17% of the unpublished scales mentioned in business journal articles discussed the content validity of the measures. He also concluded that for many of these unpublished tests, the item development process leaves a lot to be desired. Many of the scales had low reliability (and hence validity) because of poorly designed items and/or too few items.

Construct Validity. In order to understand construct validity, one first has to know what is meant by a construct. A construct is an abstract theoretical concept, such as "intelligence," for instance. An IQ test should measure intelligence and not something else, such as amount of education. The process for establishing construct validity involves the accumulation of a variety of evidence to show that the test captures the intended concept. The more methods and samples used for construct validation, the greater the confidence one can have that the test is valid from a construct validity perspective. Published tests generally deal with construct validity issues. However, in Hinkin's (1995) review of the use of unpublished scales in business journal articles, fewer than a quarter of the studies mentioned evidence for construct validity of their scales.

A popular method of construct validation is to show that the items on the test are measuring the same thing, as determined by means of a statistical technique called "factor analysis." A factor is a group of items that are measuring the same underlying characteristic. The higher an item "loads" on the factor, the stronger the

relationship between the item and the underlying construct that the factor represents in the "real world." One aims to show that all items on the test load on the factor that the test intends to measure.

Another proof is when one can show that the groups known to differ on the construct score in accordance with these expectations. Further supporting evidence can come from studies reporting that people low on a particular construct on a pre-test measure improve on the post-test measure following an intervention meant to raise the scores.

Another technique for construct validation involves a demonstration that the test has a strong relationship with other constructs known to be related to it (e.g., intelligence with academic achievement) and a weak(er) relationship with constructs not expected to correlate with it as strongly (e.g., intelligence with income). The terms "convergent validity" are sometimes used to denote the high correlations with similar constructs and "divergent validity" for the low correlations with unrelated constructs. Studies using what is called a "multitrait-multimethod" approach to examine convergent and discriminant validity provide especially strong evidence of construct validity because multiple methods are used to assess a given characteristic (trait). It has been found that some of the correlation between two measures can be inflated when the same method is used to collect the data (technical name: "method-specific variance") on both measures. The multitrait-multimethod analysis allows the researcher to determine the role that method-specific variance plays in the correlation.

Criterion-Related Validity. To household researchers and financial advisors, perhaps the most persuasive evidence is criterion-related validity. Criterion-related validation involves the demonstration of a relationship between test results and some relevant criterion measure. Evidence for criterion-related validity is most often presented in terms of correlation coefficients, which in this type of context are called "validity coefficients." It involves correlating the test (called a predictor) with the criterion. For example, a valid risk-tolerance scale should show a correlation with actual investing behavior. In other words, investors who scored high on the risk-tolerance test would be expected to hold riskier assets than investors who scored low on the test. When the criterion is available at approximately the same time as the test is given, showing such a relationship is called evidence of "concurrent validity." If the criterion is obtained in the future, it is called "predictive validity." For instance, if one gave a risk-tolerance test to a group of investors and compared the resultant scores with the riskiness of their investment portfolios at that particular point in time, it would constitute a concurrent validation strategy. On the other hand, if the test was administered to a group of people before they held any investments and their risk-tolerance test scores were then compared with their investment behaviors once they started investing, then finding a correlation between the riskiness of their portfolios and the test score would be evidence for predictive validity.

What Is an Acceptable Level of Validity? Validity coefficients can range from 0 (no relationship) to 1 (a perfect relationship). The higher the validity coefficient, the more useful is the test. Saad et al. (1999) offered the following guidelines for evaluating the magnitude of a validity coefficient:

above 0.35 = very beneficial
0.21–0.35 = likely to be useful
0.11–0.20 = depends on circumstances
below 0.11 = unlikely to be useful

Some readers may find it perplexing that such low correlations are considered proof of validity, when the requirement for reliability coefficients is so much higher. Generally, validity coefficients for any given test are going to be lower than the reliability coefficients, and this should be expected for a number of reasons. For example, the validity coefficient for a risk-tolerance test relating the test score to investing behavior should not be expected to be very high because there are numerous other factors that influence an investor's choice of investment vehicles besides the person's level of risk tolerance. It has been reported that people have perceptions and notions about the riskiness of investment decisions that may not be based on risk defined strictly by variance in returns (Weber & Milliman, 1997; Weber, Blais, & Betz, 2002; Weber, Shafir, & Blais, 2004). Moreover, portfolio allocation is subject to factors unrelated to risk tolerance, such as holding investments that were gifts and inheritances as well as inertia in making changes to investments they bought themselves (Yao, Gutter, & Hanna, 2005).

Also, in many instances, such as when the criterion is another test, the other test has some unreliability, and this attenuates the size of the correlation. If the reliability of the criterion is low, even a valid test can produce a low-validity coefficient due to the problem with the criterion. The size of the maximum possible correlation between any two tests is limited by their respective reliabilities. All other things being equal, when the reliability of the criterion is good, the validity coefficient will be higher than when the reliability of the criterion is poor. The mathematical formula to calculate the maximum possible correlation between a predictor and a criterion involves (1) multiplying the two reliability coefficients and next (2) taking the square root of the product. For example, if our predictor test has a reliability coefficient of 0.80 and the criterion has a reliability of 0.90, the largest possible validity coefficient is the square root of 0.8 times 0.7, which is about 0.85 (rounded). In contrast, if the predictor has a reliability of 0.80 and the criterion only has a reliability of 0.50, then the maximum possible validity coefficient shrinks to approximately 0.63. As you can see, this example illustrates how the validity coefficient depends on the precision of the criterion as well as the precision of the predictor.

Therefore, when evaluating a test, consider not only the size of the reported validity coefficients but also the nature of the criterion. Also take into consideration the characteristics of the samples used to derive these coefficients. If the individuals who will be taking the test given by you are different from the sample used in the validation studies, the level of validity may not be the same for these test takers.

Sensitivity and Specificity. No discussion of criterion-related validity would be complete unless the reader was made aware of the concepts of sensitivity and specificity. Obviously, the results of a test can be either right or wrong. However, in order to be able to determine the correctness of the test, one needs some way to tell what the reality is. The method for assessing the real state of events is called the

"gold standard." In medicine, pathologic findings from an invasive procedure such as surgery, tissue biopsy, or autopsy can provide the "gold standard," but in financial advising and household financial assessment, it may be necessary for the gold standard to be another test. Typically this will be a test that is known to be a reliable and valid measure of the characteristic of interest.

When the results of the test are compared with the actual situation (as determined by the gold standard), there are the four possible outcomes as illustrated in Table 2.1. Note that in addition to having verbal descriptors, the four cells in the table are labeled with the letters A, B, C, and D. This specification will allow for an easy way to illustrate the formulas for calculating sensitivity and specificity.

Table 2.1 Test Outcomes

	Condition present	Condition absent	
Test positive	True positive: A	False positive: B	(A + B)
Test negative	False negative: C	True negative: D	(C + D)
	(A + C)	(B + D)	

The four possible outcomes are as follows:

- **True positive**: when test indicates that the characteristic is present and it really is.
- **False positive**: when test indicates that the characteristic is present and it really is not.
- **False negative**: when test indicates the characteristic is not present and it really is present.
- **True negative**: when test indicates that the characteristic is not present and it really is not present.

"Sensitivity" shows the degree to which a particular test correctly identifies the presence of the characteristic. It is the proportion of people known to have the characteristic by some other means (A + C) who are correctly identified by the test (A). Stated differently, sensitivity is calculated by dividing the number of true positives by the number of individuals who have the characteristic. Expressed as a formula, it is

$$\text{Sensitivity} = \text{A}/(\text{A} + \text{C}) = \text{True Positive}/(\text{True Positive} + \text{False Negative}).$$

In contrast, "specificity" is the proportion of people without the characteristic (B + D) who have a negative test result (D). One determines specificity by dividing the number of true negatives by the number of people with the characteristic. The equation is

$$\text{Specificity} = \text{D}/(\text{C} + \text{D}) = \text{True Negative}/\text{False Negative} + \text{True Negative}$$

A test's "accuracy" is a function of both its sensitivity and its specificity. It is defined as the proportion of the cases that the test correctly identified (called "hits") as either having or not having the characteristic (A + D) from the base of all cases (A + B + C + D). Shown as an equation, it is

$$\text{Accuracy} = (A + D)/(A + B + C + D) = (\text{True Positive} + \text{True Negative})/(\text{True Positive} + \text{False Positive} + \text{False Negative} + \text{True Negative})$$

A test can be good in terms of sensitivity, yet bad in terms of specificity, and vice versa. The perfect test would be one in which both sensitivity and specificity are 100%, meaning that it would only classify people as either true positive (A) or true negative (D).

Predictive Value. Sensitivity and specificity are useful concepts for evaluating test validity, but when a new individual is tested, the person's real status is unknown at that point. Consequently, it may be more helpful for the decision maker to determine the likelihood of the characteristic being present given the test results. To help with this decision, one needs to consider the "positive predictive value" (PPV) and the "negative predictive value" (NPV) of the diagnostic test on the basis of the data shown in the 2 × 2 table used to calculate sensitivity and specificity. The PPV answers the question: "Given that this test is positive, what is the probability that this person actually possesses the characteristic?" Likewise, the NPV provides an answer to corresponding question when the results are negative: "What is the probability that the person does NOT possess the characteristic?" The ideal test would have both a PPV and an NPV of 100%. In calculating specificity and sensitivity, we were working with the columns in the table. For the PPV and NPV calculations, one needs to focus on the rows. The two equations are: PPV = A/(A + B) and NPV=D/(C + D).

Example. To help the reader understand these concepts, we present a concrete example with some numbers and a context. Suppose that a test was developed that could determine whether a client would terminate his/her business relationship with a financial advisor in the face of a financial downturn. Assume that it was tried out on 310 clients, and the results of the trial are shown in Table 2.2.

Table 2.2 Test Outcome Example

	Did terminate (positive)	Did not terminate (negative)	
Test indicates that will terminate (positive)	25 (A)	60 (B)	85 (A + B)
Test indicates that will not terminate (negative)	25 (C)	200 (D)	225 (C + D)
	50 (A + C)	260 (B + D)	310

For this hypothetical case, the indices of the various facets of criterion-related validity are as follows:

- Sensitivity = A/(A + C) =25/(25 + 25)= 25/50 = 50%
- Specificity = D/(B + D) = 200/(60 + 200) = 200/260 = 77%
- Accuracy = (A + D)/(A + B + C + D) = (25 + 200)/(25 + 60 + 25 + 200) = 225/310 = 73%
- Positive Predictive Value (PPV) = A/(A + B) = 25/(25 + 60) =25/85 = 29%
- Negative Predictive Value (NPV) = D/(C + D) =200/(25 + 200) = 200/225 = 89%

Sometimes, one will see reports of overall "hit rates" for a diagnostic test and nothing else. The overall hit rate is the same thing as the Accuracy. But the overall accuracy index can mask marked differences between Sensitivity and Specificity and thus could be a misleading indicator of test validity. The Accuracy of the test in our hypothetical example was 73%, but there was a distinct difference in its Sensitivity (50%) and its Specificity (77%). The test is clearly better at identifying who WILL NOT terminate (i.e., continue to be a client) than identifying who WILL terminate. Likewise, in concert with this, the NPV (89%) was higher than the PPV (29%). In other words, there is less than a 1 in 3 chance (29%) that the client will terminate if the test indicates he or she will do so, but close to 9 in 10 (89%) chance that the client will continue if the test indicates that he or she will not terminate.

Cut Scores. The above example deals with a test that has only two possible outcomes (positive or negative). In the case of many tests, particularly psychological ones, the results are continuous rather than discrete in nature. Therefore, one has to establish a "cut score," and where it is set will impact the specificity versus sensitivity results. The cut point decision can be made using a number of criteria, but a popular technique is to explore what happens to specificity and sensitivity when different test scores are the basis for making the cut between positive (condition present) and negative (condition absent) cases. The accuracy of any given value for the cut-point can be analyzed by the probability of a true positive (sensitivity) and the probability of a true negative (specificity). To determine the optimal value, the cut points are plotted on "Sensitivity" versus "1-Specificty" over all possible cut points. The resultant line is known as the Receiver Operating Characteristics (ROC) curve.

One method for summarizing the information from ROC is "Youden's *J*" statistic (Youden, 1950), which captures the performance of a diagnostic test with single number: $J = [(\text{sensitivity} + \text{specificity}) - 1]$. Had the cut scores in our hypothetical example been based on a certain cut point from a nondiscrete test result, $J = [(0.5 + 0.77) - 1] = 0.27$. The values of the Youden index can range from 0 to 1. A value of $J = 0$ means that there is total overlap between the positive cases and the negative cases, whereas $J = 1$ indicates that complete separation was obtained between the positive and the negative cases. The optimal cut-point based on *J* (not without its

critics) is the value that maximizes *J* (Greiner, Pfeiffer, & Smith, 2000). Again, the one number may mask things. Different types of misclassification errors may carry different associated costs, depending on the particular context. Whether to maximize specificity or sensitivity can be context dependent.

Concluding Remarks

Tests are helpful tools when carefully developed and validated. Proper test development is a laborious process. If a test already exists for what you want to measure, see if you can apply it to your situation rather than attempt to develop a "homegrown" instrument. When searching for a test, focus on the instruments with the most comprehensive documentation, and from these, select the one with the best reliability and validity. You are more likely to find validity information for a published test than for similar unpublished test. Realize, however, that even a good test will not provide totally error-free information. If you want to use a test for a purpose other than for what is was intended, it is prudent to gather evidence that the test is also valid for this new purpose.

References

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.

Batjelsmit, J. (1977). Reliability and validity of psychometric measures. In R. Andrulis (Ed.), *Adult assessment*. New York: Thomas, 29–44.

Caspi, A., Roberts, B. W., & Shiner, R. L. (2005). Personality development: Stability and change. *Annual Review of Psychology*, *56*, 453–484.

Cattell, R. B. (1986). The 16PF personality structure and Dr. Eysenck. *Journal of Social Behavior and Personality, 1*, 153–160.

Cronbach, L. (1951). Coefficient alpha and the internal consistency of tests. *Psychometrika, 16*(3), 297–334.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, CA: Sage.

Greiner, M., Pfeiffer, D., & Smith, R. D. (2000). Principles and practical application of the receiver operating characteristic analysis for diagnostic tests. *Preventive Veterinary Medicine, 45*, 23–41.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management, 21*, 967–988.

Hogan, T. P., & Agnello, J. (2004). An empirical study of reporting practices concerning measurement validity. *Educational and Psychological Measurement, 64*, 802–812.

Nunnally, J. (1967). *Psychometric theory*. New York: McGraw Hill.

Reschly, D. (1987). Learning characteristics of mildly handicapped students: Implications for classification, placement, and programming. In M. C. Wang, M. C. Reynolds, & H. J. Walberg (Eds.), *The handbook of special education: Research and practice* (pp. 35–58). Oxford: Pergamon Press.

Roszkowski, M. J., & Cordell, D. M. (2009). A longitudinal perspective on financial risk tolerance: Rank-order and mean level stability. *International Journal of Behavioural Accounting and Finance, 1,* 111–134.

Rudner, L. M. (1994). Questions to ask when evaluating tests. *Practical Assessment, Research & Evaluation*, 4(2). Retrieved February 18, 2010 from http://PAREonline.net/getvn.asp?v=4&n=2

Saad, S., Carter, G. W., Rothenberg, M., & Israelson, E. (1999). *Testing and assessment: an employer's guide to good practices.* Washington, DC: U.S. Department of Labor Employment and Training Administration.

Spreat, S., & Connelly, L. (1996). A reliability analysis of the Motivation Assessment Scale. *American Journal on Mental Retardation, 100*(5), 528–532.

Stanley, J. (1971). Reliability. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, DC: American Council on Education.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 667–680.

Thorndike, R., & Hagen, E. (1961). *Measurement and evaluation in psychology and education*. New York: John Wiley and Sons.

Weber, E. U., Blais, A-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*, 263–290.

Weber, E. U., & Milliman, R. A. (1997). Perceived risk attitudes: Relating perception to risky choice. *Management Science, 43*, 123–144.

Weber, E. U., Shafir, S., & Blais, A-R. (2004). Predicting risk-sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111*, 430–445.

Woehr, D. J., & Huffcutt, A. L. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational & Organizational Psychology*, *4*, 189–216.

Yao, R., Gutter, M.S., & Hanna, S. D. (2005). The financial risk tolerance of Blacks, Hispanics and Whites. *Financial Counseling and Planning*, *16*, 51–62.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*, 32–35.

http://www.springer.com/978-1-4419-6907-1

Financial Planning and Counseling Scales
Grable, J.E.; Archuleta, K.; Nazarinia Roy, R. (Eds.)
2011, XVIII, 647 p., Hardcover
ISBN: 978-1-4419-6907-1