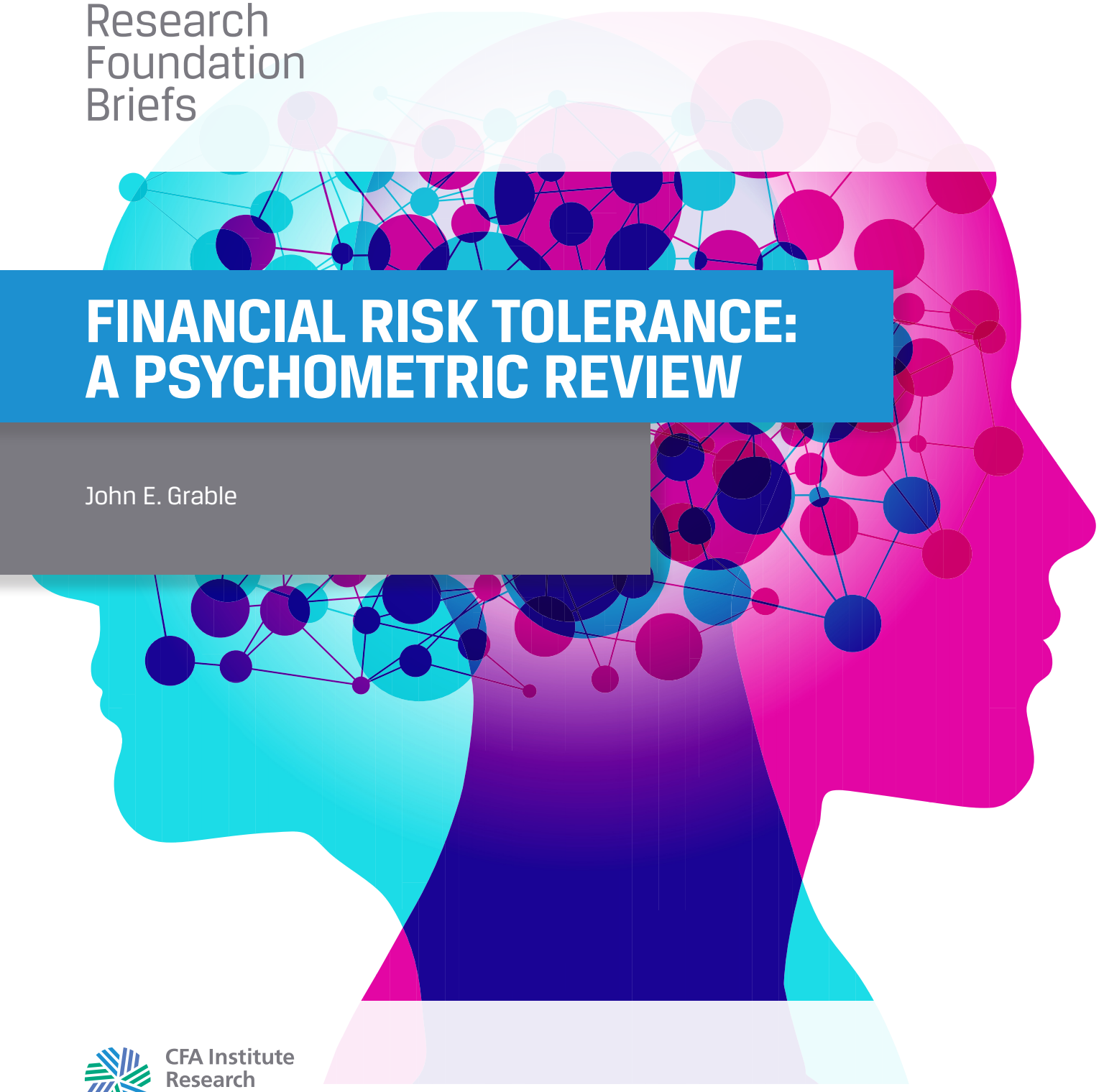
Research
Foundation
Briefs



FINANCIAL RISK TOLERANCE: A PSYCHOMETRIC REVIEW

John E. Grable

CFA Institute
Research
Foundation

FINANCIAL RISK TOLERANCE: A PSYCHOMETRIC REVIEW

John E. Grable

CFA Institute
Research
Foundation

Statement of Purpose

The CFA Institute Research Foundation is a not-for-profit organization established to promote the development and dissemination of relevant research for investment practitioners worldwide.

ISBN 978-1-944960-19-3

CONTENTS

FINANCIAL RISK TOLERANCE: A PSYCHOMETRIC REVIEW

John E. Grable
Professor of Financial Planning
Financial Planning Performance Lab
University of Georgia

This content provides financial analysts, investment professionals, and financial planners with a review of how financial risk-tolerance tests can and should be evaluated. It begins by clarifying terms related to risk taking and is followed by a broad overview of two important measurement terms: validity and reliability. It concludes with examples for practice.

INTRODUCTION AND SUMMARY

The concept of risk, and the specific evaluation of risk attitudes and risk taking, has a long and colorful history. Bernstein (1996) wrote the seminal review of the history of risk, pointing out that the concept of risk being related to outcome probabilities goes back more than 800 years. The first major breakthrough in thinking about risk, however, occurred in 1738 when Daniel Bernoulli used his knowledge of probabilities to uncover an important relationship between wealth and risk taking. He concluded that individuals prefer to take less risk and that they demand greater potential returns to engage in risky activities. Bernoulli's work laid the foundation for the development of expected utility theory and modern portfolio management principles. However, challenges to assumptions imbedded in the standard utility function began to emerge shortly after World War II. The notion that individuals, when dealing with financial decisions, always make rational choices across scenarios could not be fully supported empirically.

The first systematic risk attitude measurements were developed in the late 1950s. Kogan and Wallach (1964), for example, created the choice dilemma questionnaire, which remained a standard paradigm for the next 30 years. Their assessment tool was based on asking respondents to indicate the lowest probability of success required to undertake a risky choice in 12 scenarios dealing with a multitude of contexts. The data demonstrated that choice dilemmas did not do a consistently good job at explaining or predicting an individual's behavior (Kamalanabhan, Sunder, and Vasanthi 2000), particularly in the domain of investment and financial planning.

As behavioral economics and behavior finance gained traction as fields of study, researchers and investment professionals justifiably began to question both traditional models of economic behavior and the tools used to evaluate client attitudes. A general skepticism regarding existing frameworks led to the publication of a handful of validated financial risk-tolerance assessment instruments in the 1980s (MacCrimmon and Wehrung 1984; The American College 1994). Since that time, dozens (if not hundreds) of tools have emerged to evaluate an individual's willingness to engage in a financial behavior in which at least one outcome is both unknown and potentially negative. Nearly all of these instruments have been designed by practitioners and firms. Unfortunately, few risk-tolerance assessment tests have been created using recognized test theory principles.

This review provides financial analysts, investment professionals, and financial planners with an examination of how financial risk-tolerance tests can and should be evaluated. The review begins by clarifying terms related to risk taking. A broad overview of two important measurement terms, validity and reliability, follows. The review concludes with examples for practice.

DEFINITION OF TERMS

Financial risk tolerance is a ubiquitous phrase commonly used among financial advisers. When used broadly, financial risk tolerance is sometimes used as a catchall for many risk-related concepts. It is important to note, however, that financial risk tolerance has a very specific meaning. Cordell (2001) stated that financial risk tolerance is the maximum degree of uncertainty someone is willing to accept when making a financial decision that entails the possibility of a loss. This statement matches well with the International Organization for Standardization's (2006) definition that financial risk tolerance is the extent to which someone is willing to experience a less favorable outcome in the pursuit of an outcome with more favorable attributes. When framed this way, financial risk tolerance is distinct from concepts such as risk preference, risk perception, risk capacity, risk need, or risk composure. Each of these concepts is an essential input into the development of a person's risk profile; however, these terms are not interchangeable. **Exhibit 1** provides a brief summary of common risk terms. These definitions follow the nomenclature provided by Nobre and Grable (2015), who culled the literature for definitional frameworks.

EXHIBIT 1. RISK TERMS AND DEFINITIONS

Risk Term	Definition
Risk aversion	The inverse of risk tolerance.
Risk capacity	An objective evaluation of an individual's financial ability to withstand a financial loss.
Risk composure	An individual's propensity to behave in a consistent manner; sometimes called risk appetite (Carr 2014).
Risk need	The amount of risk an individual needs to take to reach a financial objective; typically based on a predetermined required rate of return.
Risk perception	A subjective evaluation, based on a cognitive appraisal, of the riskiness of a decision outcome.
Risk preference	An individual's general feeling that one situation is better than another.
Risk profile	An amalgamation of factors that help shape an individual's risk-taking behavior.
Risk tolerance	The willingness to engage in a risky behavior in which possible outcomes can be negative.

The following discussion summarizes issues related to the evaluation of risk tolerance specifically, and risk assessment generally.

THE SCIENCE OF PSYCHOMETRICS AND THE EVALUATION OF FINANCIAL RISK TOLERANCE

Psychometrics is a field of study that combines concepts from psychology and statistics into tools and techniques to improve psychological measurement. When psychologists, test developers, and test evaluators think about behavior, they tend to distinguish between intellectual (cognitive) and emotional (affective) pursuits. Some tests are designed to measure cognitive ability. Examples include US college-entrance examinations such as the SAT and ACT. Other tests focus on evaluating affective domains of behavior, such as personality characteristics and attitudes. Risk tolerance falls within this latter category. Generally, it is easier to measure cognitive, rather than affective, aspects of human behavior.

The application of scientific principles to the study of psychological states is a relatively new development. The origins of the field go back to the mid-1800s when researchers began to investigate intelligence from an evolutionary perspective. Since the 1930s, psychometrics has evolved dramatically. The field now encompasses the measurement and evaluation of personality, beliefs, achievement, and attitudes.

Initially, psychometricians were united around concepts embedded in what is now called classical test theory (CTT). As the field has matured, an approach known as item response theory (IRT) has been proposed as an alternative method of test evaluation. IRT had its start with tests measuring cognitive characteristics, but the principles have also been applied to tests measuring personality characteristics. The basic idea underlying IRT is that not all questions on a test measure a given characteristic to the same extent. For instance, on a cognitive test, not all questions are equally difficult. Some questions are answered correctly by everyone taking the test, whereas other questions are so hard that they can be answered correctly by only the most proficient test takers. Likewise, on personality tests, not all questions tap the measured characteristic equally. Thus, test questions can be assigned different weights based on how well they can differentiate between test takers with different levels of ability (or some other characteristic being measured). Different versions of IRT exist, based on the number of aspects of the test situation a researcher considers when developing and scoring the test. One-parameter models take into account only the test questions' difficulty. Two-parameter models take into consideration question difficulty, as well as the test taker's ability, whereas three-parameter models consider question difficulty, the test taker's ability, and the fact that guessing occurs on tests.

In a test created based on CTT, everyone takes the same test. In contrast, on a test developed using IRT, it is not necessary that all test takers be administered all the same questions. The questions individuals get asked depend on their skill level on a cognitive test (or level of a characteristic on an affective domain), as estimated by a set of questions administered at the beginning of the test. IRT proponents claim that this method allows for tailoring of test items to each test taker's ability level (adaptive testing). Another advantage of IRT is that different versions of a particular test can be equated more precisely for difficulty level. The chief disadvantage associated with IRT is that the statistical assumptions needed to use it correctly are more difficult to meet than with CTT, and typically a much larger sample size is needed to develop the test (relative to the sample size required in CTT).

Proponents of CTT and advocates of IRT do not agree about which method is superior for a given purpose. However, because nearly all financial risk-tolerance measures have been developed using CTT, the discussion in the remainder of this review is restricted to the CTT method of test design. Readers interested in a nonmathematical primer on IRT models should review DeMars (2010).

BASIC NOTIONS UNDERLYING CLASSICAL TEST THEORY

CTT is based on the notion that the score an individual obtains on a test is composed of two parts: a true score and measurement error. This relationship is expressed in the following formula:

Observed score = True score + Measurement error.

This true score represents an individual's correct score without contamination by any factors unrelated to the construct being assessed. However, it is impossible to totally avoid such contamination, and therefore, any given administration of a test is merely an estimate of this true score because each administration is tainted to at least some degree by measurement error. The true score can never be observed, only approximated.

If a test were to be administered thousands of times to the same individual, the scores that person obtained on each administration would vary, with the resultant scores typically being distributed in the form of the normal (bell-shaped) curve. The fewer the errors, the narrower the spread. Barring any change in the characteristic being measured, and no practice effects (not really possible), the observed differences in these scores on the same test would be caused by the amount of error in each score. Some scores would have little error, whereas other scores would have much error. Some scores would underestimate the person's true status on the characteristic of interest, whereas other scores would overestimate it. The average score across the multiple administrations would most closely approximate the true score because it counterbalances errors of underestimation and overestimation. Because it is not possible to obtain a true score, classical test theorists rely on observed scores to determine the quality of a measurement and to estimate the range within which the true score is likely to be.

The most important takeaway is that an observed score will begin to match the theoretical true score as measurement error decreases. All other things being equal, the less measurement error in the observed score, the better the test and the narrower the range in which the true score falls. Researchers will want to have as narrow a range as possible. In other words, the better a test is in practice, the less measurement error it will have and, therefore, the more precise estimates can be about the range in which the true score is likely to fall. Although overly simplified here, it follows that the quality of a test or other measurement tool, such as a risk-tolerance questionnaire, developed using CTT principles can be evaluated using two psychometric concepts: validity and reliability.

VALIDITY

Validity refers to the extent to which a measurement tool measures the attribute it was designed to evaluate. As noted by Roszkowski (2011), a test can be valid for one purpose yet invalid for another, which is particularly true for those who use risk-tolerance questionnaires. Some questionnaires are designed to measure an individual's willingness to engage in a risky financial behavior, whereas other tests are developed to gauge an individual's risk preference, risk perception, or risk capacity. Some evaluation scales are developed to provide a comprehensive measure of someone's risk profile. Thus, it is important to understand the purpose of an instrument and its intended audience before concluding that it is valid.

Generally, validity is measured using a combination of techniques. When an instrument is first developed, and as long as psychometric procedures are used, the test developer brings together several subject matter experts to identify preexisting questions and/or to write new questions. The use of experts to recommend and screen questions is the primary way *content validity* is ensured.

It is worth noting at this point that content validity shapes many of the outcomes associated with the use of a risk-tolerance questionnaire. An adviser who hopes to obtain a comprehensive risk profile for a client will likely be disappointed if he or she uses a questionnaire that was designed to measure the client's willingness to take risk in a specific context. Alternatively, an adviser who needs a specific evaluation score for a client's risk perceptions will find that a comprehensive risk-profiling questionnaire will provide an invalid output.

Construct validity is an important aspect of test development. Something that cannot easily be observed is considered a *construct*. Risk tolerance is a construct. A risk-tolerance test will have construct validity if the items that make up the test are actually related to the construct. A risk-tolerance questionnaire that includes questions related to a client's time horizon, cash flow needs, or economic expectations will have low construct validity. Why? These items, although important to know in their own right, are only tangentially associated with a client's willingness to engage in a risky financial behavior in which a loss is possible. A factor, such as a client's investment time horizon, may be a critical input into portfolio management decisions, but it is not theoretically associated with the construct of risk tolerance.

A subtype of construct validity is called *convergent* (*divergent*) *validity*. Scores on a test of a particular construct should be correlated with other tests of that same or a similar construct (convergent validity) but be unrelated (or related to a lesser degree) to scores from tests of dissimilar constructs (divergent validity). Convergent validity can be demonstrated by correlating scores from a newly developed test with scores

from an established scale that is known to measure something closely associated with what the new test measures. For example, a reasonable assumption is that scores from a financial risk-tolerance questionnaire should be positively correlated, to some extent, with scores derived from a scale measuring sensation seeking (e.g., people who like to gamble are also likely to be more willing to take financial risks). Divergent validity exists when the scores from a test can be shown to be unrelated to scores from a test measuring a totally unrelated construct (e.g., scores from a test of financial risk tolerance should not be highly correlated with scores from a test intended to measure interest in gardening).

Of particular importance is the concept of *criterion-related validity*. This type of validity requires that the assessment instrument be positively correlated with a criterion, such as actual behavior. Two forms of criterion-related validity can be specified. *Concurrent validity* is assessed when test takers are asked about their behavior at about the same time that the test is being taken. Evidence of *predictive validity* is collected when a test is administered prior to the measurement of a behavior. Imagine, for example, that a financial adviser knows how a group of potential clients have allocated their assets among stocks, bonds, and cash. The adviser should expect a risk-tolerance questionnaire score to be logically consistent with each client's asset allocation framework. That is, in terms of concurrent validity, the evidence should show that clients with a high risk-tolerance score hold a significant percentage of their portfolios in equities. Predictive validity would be present if it turns out that prospective clients with low risk-tolerance scores sold equity holdings in the future during a market correction.

Criterion-related validity is typically measured with a correlation coefficient. Saad, Carter, Rothenberg, and Israelson (1999) recommended that the following correlation guidelines be used when evaluating criterion-related validity:

Above 0.35: Useful

0.21 to 0.35: Some usefulness

0.11 to 0.20: Acceptable in some circumstances

Below 0.11: Problematic

As an example, assume that a financial adviser uses a risk-tolerance questionnaire with 100 clients. If the adviser were to correlate risk scores with the ratio of equities-to-fixed-income securities or with future behavior (e.g., using 1 = sold stock in correction and 0 = held stock during correction), these coefficient guidelines could be used to determine the criterion-related validity of the questionnaire. For those familiar with statistical norms, these correlation coefficients may seem low. In terms of validity

assessment, however, the size of the coefficients is acceptable. When explaining human behavior (such as investing), one variable (risk score) is unlikely to explain a significantly large percentage of the variance in the behavior. As such, modest correlation coefficients are to be expected and should not be dismissed.

When evaluating the validity of a risk-tolerance instrument, calculating sensitivity and specificity estimates is sometimes helpful. Imagine a risk-tolerance test that is designed to categorize clients into one of two categories: high or low risk tolerance. As **Exhibit 2** shows, four outcomes are possible: (1) true positive, (2) false positive, (3) false negative, and (4) true negative.

EXHIBIT 2. OUTCOMES ASSOCIATED WITH TEST ADMINISTRATION

	Actual High Risk Tolerance	Actual Low Risk Tolerance
High Risk Tolerance Prediction	True positive (TP)	False positive (FP)
Low Risk Tolerance Prediction	False negative (FN)	True negative (TN)

Sensitivity refers to how well a test correctly identifies the presence of an attribute. Sensitivity is calculated by dividing the number of true positives by the number of individuals with the attribute:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{TN}).$$

Specificity is the proportion of test takers without the attribute. It can be calculated by dividing the number of true negatives by the number of individuals without the attribute:

$$\text{Specificity} = \text{TN}/(\text{FN} + \text{TN}).$$

A test's *accuracy* is then the proportion of cases that are true to the total number of cases:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN}).$$

Data from Exhibit 2 can also be used to predict the validity of a test at the individual level. Generally, a financial adviser will not know a prospective client's real risk tolerance. As such, the adviser will use a risk-tolerance questionnaire score to predict the client's attitude and behavior. A positive predictive value (PPV) refers to the probability that an individual possesses an attribute. A negative predictive value (NPV) describes

the probability that the individual does not have the attribute. PPV and NPV can be calculated as follows:

$$PPV = TP/(TP + FP)$$

$$NPV = TN/(FN + TN).$$

For example, assume a financial risk-tolerance questionnaire was administered to 100 clients. After a market correction, the actual behavior of clients was assessed to see who held, added to, or reduced their equity holdings. **Exhibit 3** shows the results from the analysis.

EXHIBIT 3. TEST PREDICTIONS AND OUTCOMES

	Held or Added to Equities Holdings	Reduced Equity Holdings
High Risk Tolerance Prediction	40 TP	10 FP
Low Risk Tolerance Prediction	20 FN	30 TN

Each of the validity indicators can be calculated from data in Exhibit 3, as follows:

$$\text{Sensitivity} = 40/(40 + 20) = 67\%$$

$$\text{Specificity} = 30/(10 + 30) = 75\%$$

$$\text{Accuracy} = (40 + 30)/(40 + 10 + 20 + 30) = 70\%$$

$$PPV = 40/(40 + 10) = 80\%$$

$$NPV = 30/(20 + 30) = 60\%.$$

The risk-tolerance test has an overall accuracy level of 70%, with a slightly higher level of specificity. This degree of specificity means that the test does a somewhat better job of predicting the behavior of those with a low risk tolerance. In terms of predictive power, the PPV indicates that there is an 80% chance that those with a high risk-tolerance score will hold or increase their position in equities during a market correct or decrease their equity holdings if they have a low risk tolerance.

RELIABILITY

All risk-tolerance assessments contain some *measurement error*. The extent to which measurement error influences the calculation of a final score is important. A test's reliability helps answer the question of how much margin of error is provided by an

assessment instrument.[1] Tests with high measurement error result in low reliability estimates. *Reliability*, within the context of financial risk-tolerance assessment, is an important concept because financial advisers almost always want to measure their clients' risk attitudes consistently.

Whereas validity indicates the extent to which an assessment tool measures what it purports to measure, reliability denotes how repeatable the score from an assessment tool is in practice. Consider again the traditional CTT formula:

Observed score = True score + Measurement error.

A reliability estimate indicates how much of the observed score is distorted by measurement error: the higher the reliability, the lower the measurement error (and vice versa). The higher the reliability, the greater the confidence that the observed score is closer to the true score.

Determinants of Reliability

What leads to measurement error? Random events certainly contribute to errors. The exam taker's mood or health situation can influence outcomes. Distractions in the room where an individual is taking an exam or environmental factors, such as an overly cold or hot testing room, can increase measurement error. For these reasons, nearly all standardized tests (e.g., SAT, Graduate Management Admission Test, Graduate Record Examinations, securities licensing exams) are administered in tightly controlled environments to minimize environmental factors that can increase measurement error.

The primary source of measurement error, however, comes from poorly designed tests with ambiguous wording, which explains the linkage between validity and reliability. Essentially, a valid test is generally a reliable test; however, a reliable test may not be valid. The first statement is self-evident. If the questions used to make up an assessment are badly worded, inconsistent, or confusing—signs of low validity—the resulting test will be less reliable. The second statement is more nuanced. A test may be very reliable, in that it consistently measures something in a repeatable manner, but it may not actually measure what it is intended to assess. For instance, some risk-tolerance tools used by financial advisers are thought to measure an individual's willingness to take risk when, in fact, the tests measure something entirely different, such as a client's investment time horizon, spending preferences, or some other personal characteristics.

[1] Within the psychometric community, reliability is the ratio of a test's true score to the observed score, based on the calculated score from the test.

Consider the following questions. Each question represents a typical item in what some financial advisory firms call an investment questionnaire, or more broadly, a financial risk assessment.

1. I plan to begin taking money from my portfolio in:

    a. 1 year or less
    b. 1 to 2 years
    c. 3 to 5 years
    d. 6 to 10 years
    e. 10 years or more

2. When you withdraw money from investments, you usually spend the distribution over what time period?

    a. 1 year or less
    b. 1 to 2 years
    c. 3 to 5 years
    d. 6 to 10 years
    e. 10 years or more

3. During the global financial crisis of 2007–2009, stocks lost 57% of their value from top to bottom. If you owned stocks that lost this amount in just a few months, you would:

    a. buy more of the investment.
    b. hold the investment and do nothing.
    c. sell a portion of the investment.
    d. sell all of the investment.

4. Which investment do you *prefer*?

    a. One with little or no fluctuation in value
    b. One with some fluctuation in value
    c. One with moderate fluctuations in value
    d. One with large fluctuations in value

5. You would invest in a stock or mutual fund based on a conversation with a coworker, friend, or family member.

    a. Strongly agree
    b. Agree
    c. Neither agree or disagree
    d. Disagree
    e. Strongly disagree

Which of these questions is an appropriate item to include is a financial risk-tolerance questionnaire? When answering, remember that risk tolerance is defined as an individual's willingness to take risk when a possible outcome is negative. It turns out that only one of the questions works reasonably well in this context. Roszkowski, Davey, and Grable (2005) noted that "mixing questions about more than one construct in a single brief questionnaire will almost invariably lead to an inaccurate assessment of all the constructs because none can be measured adequately due to the brevity of the questionnaire" (p. 68). Specifically, with these examples,

- The first question is intended to measure a client's investment time horizon.
- The second question is designed to evaluate a client's spending behavior.
- The third question—the best of the five examples—is intended to predict future behavior. Even though this is the best of the five questions, note that little academic evidence exists to suggest that people are particularly good at forecasting their future actions.
- The fourth question clearly measure's a client's risk preference, not tolerance.
- The fifth question is "double barreled," meaning that it requires a client to make one choice based on two conditions. Stocks and mutual funds are not exactly the same, which could cause confusion if a client's choice would change if stocks and mutual funds were presented separately. In addition, a client's answer might be different if the question asked about receiving information from one source, such as a family member only.

The sample questions illustrate how intertwined the concepts of validity and reliability are in practice. It is possible that these questions, when included in a questionnaire, might result in a high reliability estimate, even though the validity of the questions as risk-tolerance items is rather mixed. In other situations, these questions may work well. For example, if a financial adviser wanted to design a test to determine an asset allocation framework, these questions might be appropriate.

It is important to understand the primary determinant of a test's reliability: the number of questions. Generally, shorter tests have lower reliability coefficients. Thus, financial advisers should be skeptical of claims that one, two, or three questions can be used to adequately measure a person's risk tolerance. The *Spearman–Brown prophecy formula*[2]

[2] The Spearman–Brown prophecy formula was conceptualized in the mid-20th century. The formula allows a test user to estimate the reliability coefficient of a test when the number of assessment items is either increased or decreased. A practical example of the formula's use can be found in Beckman, Ghosh, Cook, Erwin, and Mandrekar (2004).

can be used to determine how many questions are needed to obtain a given reliability coefficient. Roszkowski and his coauthors (2005) used the Spearman–Brown prophecy formula to conclude that a 15-item risk-tolerance scale with a reliability estimate of 0.71 would need an additional 10 questions to achieve a reliability estimate equal to 0.80. This finding highlights a potential problem with the development and use of financial risk-tolerance tests. Ideally, a test should have a very high reliability estimate; however, this may require a large pool of questions. Unfortunately, clients cannot be expected to answer a battery of risk queries without becoming fatigued and bored, which helps explain why shorter assessments are preferred. When evaluating a test, a financial adviser should use professional judgment in balancing the number of items with an acceptable level of reliability. But be aware that for two tests of the same length, the one with the highest reliability coefficient is likely to provide the most consistent and repeatable outputs.

Reliability Scores

Reliability estimates are more difficult to calculate but easier to interpret compared with validity estimates. Within CTT, reliability is measured with a correlation coefficient. Correlations can theoretically range from –1.0 to +1.0. A test with a reliability of +1.0 is said to be perfectly reliable; that is, the same outcome is obtained when the test is given repeatedly. In practice, obtaining a reliability estimate of 1.0 is very rare (nearly impossible). In contrast, a published test would likely never exhibit a negative reliability coefficient. A negative coefficient would indicate that the test is seriously flawed.

Although a financial adviser can generally conduct a validity check on items in a risk-tolerance assessment, estimating a reliability coefficient directly is difficult. Instead, test users tend to rely on reported estimates from a test's authors or from an independent evaluation. The general rule is that whenever a test score is used to make judgments about an individual, the test's reliability should be relatively high. Nunnally (1967) provided guidelines on the acceptability of reliability estimates. The US Department of Labor revised Nunnally's original guidelines as follows (Saad et al. 1999):

Excellent = 0.90 or higher

Good = 0.80 to 0.89

Adequate = 0.70 to 0.79

Questionable = 0.69 or below

In practice, financial advisers should use risk-tolerance questionnaires with a reported reliability estimate of at least 0.70. Using a test with an undocumented reliability estimate

or one with a lower reliability score will increase the probability that the obtained score varies too much from the theoretical true score. In other words, the likelihood that the observed score is accurate diminishes as the reliability estimate falls because there is too much "noise" (measurement error) to pick up the "signal" (true score) when reliability is low.

Reliability Measurement Approaches

Although it is possible for financial advisory firms to obtain reliability coefficients based on data obtained from their own clients, nearly all advisers instead rely on the estimates provided by the test publisher. A number of approaches to determining reliability are available to researchers. The most widely used method of reliability estimation is the *internal consistency* measurement. Reliability based on internal consistency is premised on the notion that the reliability of a test can be obtained by looking at the number of items in a test, their variances, and their covariances. A basic internal consistency approach involves taking a longer instrument, splitting it in half, and comparing the correlation of the two measures. This approach is called *split-half reliability*. Adjustments need to be made for test length to obtain an accurate estimate of reliability.

The problem with the split-half approach is that the size of the reliability coefficient depends on which items go into each half. An alternative internal consistency approach involves calculating *Cronbach's alpha* (α), which can be estimated as follows (see Cronbach 1951):

$$\alpha = (N \times \overline{C})/(\overline{V} + (N-1) \times \overline{C}),$$

where N is the number of items in the test, $\overline{C}$ is the average inter-item covariance among the items, and $\overline{V}$ is the average variance. Cronbach's alpha is conceptualized as the average of all possible split-half estimates of reliability. The value of Cronbach's alpha goes up as the number of items and their covariance increase. In other words, all else being equal, tests in which the items inter-correlate and the number of questions is large will produce high Cronbach's alphas.

Cronbach's alpha is used whenever the items in a test are coded as a continuous variable. An analogous test, the *Kuder–Richardson formula 20,* can be used to derive a similar reliability estimate when a test is composed of dichotomous items (Kuder and Richardson 1937). There is active debate within the psychometric community regarding the usefulness of reliability estimates based on Cronbach's alpha and the Kuder–Richardson formula 20; however, the general consensus is that tests that report reliability coefficients above 0.70 likely provide consistent and repeatable outputs for use in practice. Some statisticians warn that Cronbach's alpha assumes a true score

equivalence (tau equivalent) model, which requires that a single latent trait (factor) underlies the scale, that the items forming it have equal variances, and that the covariances between these items must be the same (Tavakol and Dennick 2011). If these conditions are not met, Cronbach's alpha will underestimate reliability.

Other techniques used to measure reliability include test–retest reliability and inter-rater reliability. The *test–retest reliability* procedure involves administering a test to a group of participants, allowing a short interval to pass, and retesting the same group. The correlation between the first and second test provides an estimate of test reliability. Threats to this procedure include changes in environmental factors that can increase measurement error and *recall bias*, which is related to a test taker's ability to recall what he or she answered previously. *Inter-rater reliability* is used in situations where judges assign a score to a person or an object and the similarity among their ratings is assessed. For instance, assume that two investment professionals are rating the risk tolerance of the same group of clients based on an interview. Reliability would be high if they rated the clients similarly. A number of statistical measures have traditionally been used to measure inter-rater reliability, including traditional correlation coefficients (Pearson and Spearman), Cohen's kappa, and Kendall's coefficient (Gwet 2014).

VALIDITY AND RELIABILITY IN PRACTICE: THE STANDARD ERROR OF MEASUREMENT

Imagine that a financial adviser decides to do the maximum possible to assess and evaluate her clients' financial risk tolerance. After searching the marketplace, she chooses a risk assessment tool that provides an output based on a 0 to 100 scale, with higher scores representing a greater tolerance for financial risk. This financial adviser is so systematic that she decides to administer the test to clients on a yearly basis. After a few years, she starts to notice something that is both intriguing and potentially worrisome. Looking at one particular client, as an example, she notices that the risk-tolerance scores fluctuate up and down year by year. Over a five-year period, the scores are 75, 71, 82, 69, and 75. The financial adviser is concerned because although she knows that client scores can fluctuate, she did not expect such wide swings from year to year.

Should the financial adviser in this case be concerned? Fortunately, she can estimate whether the variation in scores exhibited by her client is within a reasonable *margin of error*. If she has used a valid assessment instrument (i.e., the questions are appropriate

and the intent matches her purposes) and the average score on the test is 75, she can apply a standard error of measurement procedure to identify the range of scores where her client's true risk-tolerance score is located. To do so, she will need two pieces of data: (1) the reliability estimate for the test and (2) the standard deviation of the test based on a normed sample. Both of these data points generally come from the test's developers.

After completing her research, the financial adviser determined that the test's reliability is 0.65 with a sample score standard deviation of 10 points. With these data, she can determine whether the observed scores fall within the margin of error by applying the following analytic process:

Subtract the reliability coefficient from 1.0:

$$1.0 - 0.65 = 0.35.$$

Calculate the square root of the estimate:

$$\sqrt{0.35} = 0.5916.$$

Multiply the calculated square root by the test's standard deviation to estimate the *standard error of measurement* (SE_m):

$$0.5916 \times 10 = 5.9161, \text{ or } 6.0 \text{ rounded.}$$

Estimate a 95% *confidence interval* by multiplying the SE_m by 1.96 (this is the approximate *Z*-score associated with 95% coverage within a normal distribution):

$$6.0 \times 1.96 = 11.76, \text{ rounded to } 12.0.$$

The confidence interval can then be used to answer the financial adviser's primary question. She can use the average test score of 75 as the baseline, and add and subtract the confidence interval from the baseline. In other words, she can use the SE_m to identify that the client's true score falls plus or minus 12 points from the baseline or between 63 and 87, and that the variation in the client's scores, based on the test reliability and standard deviation, is reasonable.

However, SE_m may not be sufficient to directly answer the adviser's larger worry: Is a 12-point variation in risk-tolerance scores acceptable when the scores are used as an input to designing a portfolio recommendation or when drafting other financial planning recommendations? (When estimating repeatability, some psychometricians multiply the SE_m by a much higher [2.77] *Z*-score to obtain a 99% confidence interval;

this procedure provides an estimate of the expected variability—*reliability limits of agreement*—that one client may exhibit.) Although the answer depends on each client's situation, it may seem somewhat intuitive that this margin of error is a bit wide. It would be more beneficial to obtain an estimate in which the band of error is narrower. One way to move toward this outcome is to use a risk-tolerance assessment instrument with a higher reliability coefficient.

If, for example, the financial adviser could find a test with a similar standard deviation of normed scores and a reliability coefficient equal to 0.80, the confidence interval would fall to 9.0. The resulting narrower range in scores would give her greater confidence that the client's true score was somewhere between 66 and 84. The key takeaway is that the higher a test's reliability, the smaller the SE_m.

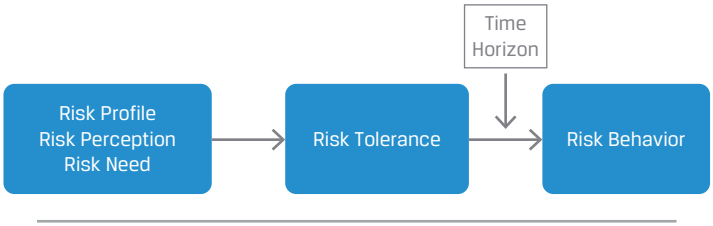
PRACTICAL GUIDELINES FOR PRACTITIONERS

The incorporation of financial risk-tolerance test scores into investment plans is a topic of interest to nearly all financial advisers. In the context of portfolio management, risk tolerance, as defined in this review, can be conceptualized in one of two ways. The first way is to view risk tolerance as a single input into a client's overall risk profile. Klement (2016) argued that risk tolerance is just one of several factors that comprise a client's risk profile. Other factors include age, investment objectives, time horizon, experience, and risk capacity. The combination of these factors, as evaluated by a financial adviser, determines the appropriate asset allocation mix. Carr (2014), for example, showed that a client's risk perception and risk need, in addition to risk tolerance, were the most important characteristics shaping an individual's risk profile.

The second way risk tolerance can be conceptualized is as a primary determinant of portfolio decisions. This matches the notion of financial risk tolerance as singularly more important within the context of investment management decisions than other factors, such as client financial knowledge. Nobre and Grable (2015), for example, noted that an individual's willingness to take financial risk is influenced by his or her risk perception, risk need, and risk profile—which they defined as being composed of risk capacity, risk preference, and risk composure. When viewed this way, a client may be willing to take risks when presented with one financial decision but be unwilling to take risks in another situation. This is true even if a client is both fiscally and emotionally primed to engage in either behavior. A client's perception that the risk in one scenario is lower (or higher) than in another may shape his or her willingness to take a risk. As shown in **Figure 1**, risk tolerance, from this perspective, acts as a mediator

between a client's risk profile, risk perception, risk need, and engagement in a risky behavior. In this model, a client's time horizon serves to either enhance or reduce the influence of risk tolerance on behavior. For instance, someone with a long investment time horizon could reasonably be more aggressive than a similar person with a short investment time horizon.

FIGURE 1. FINANCIAL RISK TOLERANCE AS MEDIATOR



SUMMARY

Regardless of how one views financial risk tolerance within the investment planning process, several practical guidelines are worth noting when assessing and evaluating a client's willingness to take risk. First, the tool used to measure and evaluate a client's risk tolerance should be valid. Validity in this context means that the items reflect actual risk tolerance, not concepts related to time horizon, spending plans, or risk capacity. Second, the assessment tools should exhibit strong reliability. At a minimum, the questionnaire should have a Cronbach's alpha of at least 0.70. Third, and perhaps most important, the resulting score from a test should be used as a starting point in the investment planning process. As noted by Klement (2016), the derived financial risk-tolerance score (as well as the client's overall risk profile estimate) should form the foundation for ongoing discussions between the adviser and client. A valid and reliable financial risk-tolerance test is not only an essential investment planning tool but also an important data point that can be used to better understand a client's beliefs and behaviors.

REFERENCES

The American College. 1994. *Survey of Financial Risk Tolerance: User's Guide*. Bryn Mawr, PA: The American College.

Beckman, T.J., A.K. Ghosh, D.A. Cook, P.J. Erwin, and J.N. Mandrekar. 2004. "How Reliable Are Assessments of Clinical Teaching?" *Journal of General Internal Medicine*, vol. 19, no. 9: 971–977.

Bernstein, P.L. 1996. *Against the Gods: The Remarkable Story of Risk*. New York: Wiley.

Carr, N. 2014. "Reassessing the Assessment: Exploring the Factors That Contribute to Comprehensive Financial Risk Evaluation." Unpublished doctoral dissertation, Kansas State University.

Cordell, D.M. 2001. "RiskPACK: How to Evaluate Risk Tolerance." *Journal of Financial Planning,* vol. 14, no. 6: 36–40.

Cronbach, L.J. 1951. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, vol. 16, no. 3: 297–334.

DeMars, C. 2010. *Item Response Theory*. New York: Oxford University Press.

Gwet, K.L. 2014. *The Handbook of Inter-Rater Reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics.

International Organization for Standardization. 2006. *Personal Financial Planning 22222:2005*. Geneva: ISO.

Kamalanabhan, T.J., D.L. Sunder, and M. Vasanthi. 2000. "An Evaluation of the Choice Dilemma Questionnaire as a Measure of Risk-Taking Propensity." *Social Behavior and Personality: An International Journal*, vol. 28, no. 2: 149–156.

Klement, J. 2016. *Investor Risk Profiling: An Overview*. Charlottesville, VA: CFA Institute Research Foundation.

Kogan, N., and M.A. Wallach. 1964. *Risk Taking: A Study in Cognition and Personality*. New York: Holt Rinehart & Winston.

Kuder, G.F., and M.W. Richardson 1937. "The Theory of the Estimation of Test Reliability." *Psychometrika*, vol. 2, no. 3: 151–160.

MacCrimmon, K.R., and D.A. Wehrung. 1984. "The Risk In-Basket." *Journal of Business*, vol. 57, no. 3: 367–387.

Nobre, L.H.N., and J.E. Grable. 2015. "The Role of Risk Profiles and Risk Tolerance in Shaping Client Investment Decisions." *Journal of Financial Service Professionals,* vol. 69, no. 3: 18–21.

Nunnally, J. 1967. *Psychometric Theory*. New York: McGraw-Hill.

Roszkowski, M.J. 2011. "Issues to Consider When Evaluating 'Tests.'" In *Financial Planning and Counseling Scales*. Edited by John E. Grable, Kristy L. Archuleta, and R. Roudi Nazarinia. New York: Springer.

Roszkowski, M.J., G. Davey, and J.E. Grable. 2005. "Insights from Psychology and Psychometrics on Measuring Risk Tolerance." *Journal of Financial Planning,* vol. 18, no. 4: 66–77.

Saad, S., G.W. Carter, M. Rothenberg, and E. Israelson. 1999. "Testing and Assessment: An Employer's Guide to Good Practices." Washington, DC: US Department of Labor Employment and Training Administration.

Tavakol, M., and R. Dennick. 2011. "Making Sense of Cronbach's Alpha." *International Journal of Medical Education,* vol. 2, no. 1: 53–55.

The author is indebted to Mike Roszkowski for his comments and his review of an earlier version of this manuscript.

Named Endowments

The CFA Institute Research Foundation acknowledges with sincere gratitude the generous contributions of the Named Endowment participants listed below.

Gifts of at least US$100,000 qualify donors for membership in the Named Endowment category, which recognizes in perpetuity the commitment toward unbiased, practitioner-oriented, relevant research that these firms and individuals have expressed through their generous support of the CFA Institute Research Foundation.

Ameritech
Anonymous
Robert D. Arnott
Theodore R. Aronson, CFA
Asahi Mutual Life Insurance Company
Batterymarch Financial Management
Boston Company
Boston Partners Asset Management, L.P.
Gary P. Brinson, CFA
Brinson Partners, Inc.
Capital Group International, Inc.
Concord Capital Management
Dai-Ichi Life Insurance Company
Daiwa Securities
Mr. and Mrs. Jeffrey Diermeier
Gifford Fong Associates
Investment Counsel Association of America, Inc.
Jacobs Levy Equity Management
John A. Gunn, CFA
John B. Neff, CFA
Jon L. Hagler Foundation
Long-Term Credit Bank of Japan, Ltd.
Lynch, Jones & Ryan, LLC
Meiji Mutual Life Insurance Company
Miller Anderson & Sherrerd, LLP
Nikko Securities Co., Ltd.
Nippon Life Insurance Company of Japan
Nomura Securities Co., Ltd.
Payden & Rygel
Provident National Bank
Frank K. Reilly, CFA
Salomon Brothers
Sassoon Holdings Pte. Ltd.
Scudder Stevens & Clark
Security Analysts Association of Japan
Shaw Data Securities, Inc.
Sit Investment Associates, Inc.
Standish, Ayer & Wood, Inc.
State Farm Insurance Company
Sumitomo Life America, Inc.
T. Rowe Price Associates, Inc.
Templeton Investment Counsel Inc.
Frank Trainer, CFA
Travelers Insurance Co.
USF&G Companies
Yamaichi Securities Co., Ltd.

Senior Research Fellows

Financial Services Analyst Association

The CFA Institute Research Foundation Board of Trustees 2016–2017

Officers and Directors

Research Foundation Review Board

CFA Institute
Research
Foundation

Available online at **www.cfapubs.org**