R&D Connections
March 2007

Is Test Score Reliability Necessary?

Michael E. Walker

"My car is not very reliable." "That television station reports the news reliably." "That meteorologist gives reliable weather reports." In everyday usage, when we say that something is reliable, we mean that it is trustworthy or truthful or dependable. A reliable car rarely breaks down. Reliable news is objective and unbiased. The weather report is reliable if it is accurate.

In the standardized testing arena, however, the definition of reliability is somewhat narrower, although it may at first glance appear broader. Simply put, reliability equals consistency. Does the object in question produce similar responses in comparable situations? My old pickup truck consistently broke down if I drove it more than 100 miles on the highway. According to the narrow definition, it was highly reliable because its behavior was predictable. The news programs on two different television stations may both be highly reliable in their messaging, although they may give completely contradictory accounts of the same event. Finally, a meteorologist who consistently predicts rain on sunny days and sun on rainy days would always be wrong, of course, but could still be considered reliable.

Do the above examples mean that reliability in the narrow sense is a worthless concept? Not at all. Consistency is a virtue. If we know what to expect, we know how to react. For example, I learned never to take trips more than 90 miles in my truck. Likewise, we know how to interpret the news we receive from different sources. The misguided meteorologist is perhaps the most useful: If we simply do the opposite of whatever the weather report would suggest, we will never get caught in the rain without an umbrella.

By contrast, what if something is inconsistent or unreliable? In this situation we will not know what to expect and therefore cannot take the proper actions. Sometimes, if we know something is unreliable, we can discount it. How much faith, for example, would we put in the reports of a television station that delivers some actual news intermingled with random or fabricated news reports? At other times, we do not know that something is unreliable and so act as if it is. In these situations, at best the outcome is annoying (e.g., we may be caught in the rain); but at worst the outcome can be very dangerous (e.g., breaking down in the midst of heavy traffic). The same arguments apply to standardized testing.

Error in Measurement

Central to the concept of reliability is the notion of observational error. By the term *observation*, I mean not just perception but any measurement or recording of a characteristic of an object that is separate from the thing being measured. We may weigh vegetables, listen to a heartbeat, or make a plaster cast of a baby's foot. Each of these observations, no matter how precise we try to make it, is still subject to error. The plaster cast may include slight imperfections not actually present on the baby's foot (and may miss some features); the stethoscope may not detect all sounds of the heartbeat, and movement may add extraneous sounds; and if we weigh the vegetables a second time, even with the same scale, we will get a slightly different answer.

Most people today would agree that no observation or measurement is perfect, although they might agree selectively (e.g., when they are recorded on radar as speeding). Thus, if we were

confronted with several measurements of the same object, all slightly different, we would probably not be surprised. Moreover, we might consider that all of the measurements were equally correct. This was not always the case, as the following story will illustrate.

Locating a Star

In 1572 a nova appeared. Thirteen different astronomers, located in different parts of the world, recorded the position of the nova in the night sky. If we took any two of these astronomers, by knowing their exact location and the directions in which they recorded the nova, we could plot the location of the star by estimating where the astronomers' lines of sight crossed each other. The problem is that we would get different estimates of the location of the star, depending upon which pair of astronomers we picked. For this particular set of data, some locations would even be completely impossible (placing the star inside the Earth, for example).

One scientist of the day used the observations from 12 pairs of these astronomers (out of the 78 possible pairs) to compute the position of the star in the sky. He took as correct only the data that agreed with his preconceived notion of the universe, discarding as wrong the majority of the data that did not agree (Meli, 2004). This was the generally preferred method at the time for dealing with seemingly contradictory information.

Galileo Galilei (1632/2001) changed all of that. He provided us with arguably the earliest and most complete exposition of a theory of observational error. In refuting this scientist, Galileo began by asserting that all of the astronomers' observations had error, but that those errors were more likely to be small rather than large. He said that measurement had the same chance of being too high as too low. With this in mind, he argued that contradictory values should not be discarded but corrected. By making very small corrections, Galileo could make the observations consistent with one another, thereby leading to an estimated placement of the star in the sky.

Galileo's ideas may not seem very brilliant from today's perspective, but they revolutionized scientific inquiry at the time and also shaped the way we currently do many things. Among other things, Galileo's work laid the foundation for the psychometric development that would take place in later centuries, culminating in the development of classical test theory nearly 300 years later.

Classical Test Theory

Classical test theory (CTT; see Gulliksen, 1950; Lord & Novick, 1968; Nunnally & Bernstein, 1994) is the collective name given to several statistical advancements in testing that took place in the first half of the twentieth century. It concerns itself mostly with the measurement of human ability, although the theory can be applied in other areas as well. At the heart of CTT is Galileo's notion of error: Every observed test score is made up of a true score (reflecting a person's actual ability) plus an error score.

There are many reasons a person may score higher or lower on a test on any given day: situational variables, the particular set of items (test questions) that appear on the test, and a number of other factors. We say that these things, and anything else except for the person's true ability, contribute to the error score. We assume that the error can be positive or negative with equal likelihood, and that it is completely random or unpredictable. Thus, if we observe only one score for an individual, we have no reason not to take it as an unbiased estimate of the person's true ability. If we have two or more observations for the same individual, we would expect the error scores to offset each other to some degree. This means that we can improve our estimate of the person's true score by testing the person multiple times and averaging the results.[1] We may even think of the true score as being the

[1] Especially with ability testing, we generally use different but equivalent versions of the test every time we test an individual. This prevents all sorts of problems, such as the examinee memorizing the answers and responding the same way every time, or the fact that any given version of the test cannot cover every possible aspect of the subject matter being tested.

average of an infinite number of observed test scores for a given individual (Suen, 1990).[2]

It is important to be very specific about what I mean by true score. Let us take weight as an example. The true score could be thought of as an object's weight in some absolute sense. However, this absolute weight is a theoretical construct that we can never actually measure, because every scale has its own built-in bias (Sutcliffe, 1965). Nevertheless, we might estimate the true weight of an object by averaging several measurements using this slightly biased scale. We may legitimately consider this average to be the true score in the context of CTT, but it is not necessarily the correct answer. Although an exact definition of true score is not essential to understand the discussion that follows, the distinction is nevertheless a good one to bear in mind.

Quantifying Consistency

We can imagine that different measurement instruments would differ in their precision or in how little error they introduce into the measurement. We need some index to quantify the measurement consistency. One such index is known as the error variance. This is conceptualized as the squared error scores, averaged across several observations or measurements. We can also think of it as the average squared difference between each measurement and the corresponding true score. Because the error variance is represented in squared measurement units, it is customary to take its square root and to use that instead. This quantity is known as the *standard error of measurement* (SEM), or sometimes just the standard error.

The SEM is an index of the typical error we could expect with any given observation. If we are weighing something, and we know that the SEM is two pounds, then upon repeated measurements, we would expect our observed weight to differ from our true weight by more than two pounds only about one-third of the time. Of course, whether we consider this error to be acceptable or not will depend upon how heavy the things are we are weighing. An SEM of two pounds is probably negligible if we are weighing dump trucks, but not if we are weighing kittens.

This is one reason that a standardized index of measurement consistency is useful. The *reliability coefficient* provides such an index.[3] The reliability coefficient relates the amount of error variance in the scores of a set of objects or people to the total amount of variability among those scores.

Recall that an observed score is made up of a true score plus error. Ideally, we want all of the observed differences among scores to represent actual differences in ability (i.e., the true score). We want none of the observed differences to be due to measurement error.[4] The reliability coefficient tells us how close we are to this ideal. It reflects the proportion of observed differences in a set of scores that is attributable to true score differences. Reliability, then, can be seen as an index of what proportion of total variability in observed scores is due to true differences among the people or objects in the group. Reliability can range in value from zero to one. A measurement

[2] This simple definition immediately raises several issues. For example, it is quite likely that the act of measuring in fact changes the object being measured, so that repeated measurement is impossible (Feldt & Brennan, 1989). More generally, we may ask what conditions must remain the same for us to be able to say that we have repeatedly observed a single phenomenon, as opposed to our saying that we have observed several different phenomena (Brennan, 2001). This particular issue is not a trivial one; however, a satisfactory answer has not been reached in more than 100 years. It is certain that we could not reach one here. For our purposes, we can accept the definition at face value without thinking too hard about the particulars.

[3] There are many other indices related to the precision of measurement that are used in different contexts. For example, for auditory signals, we may use the decibel or signal-to-noise ratio. In the testing arena, the SEM and reliability are the two most often used indices. Kane (1996) showed the relationships among these and other indices of precision.

[4] There are several methods for estimating test reliability, which yield somewhat different results. The difference among the results reflects the difference in what constitutes error for each method. Error in one context may not be considered error in another.

procedure with a reliability of one gives the true score every time. A procedure with a reliability of zero never does; it just gives random answers.

Reliability can be said to reflect the proportion of variability in observed scores that is *not* attributable to error. Going back to our example, the observed weight of dump trucks might vary quite a bit across a set of trucks (perhaps on the order of a few hundred pounds), so that an error variance of two pounds would indicate a highly reliable measurement procedure. By comparison, an error variance of two pounds would make up the majority of the variance of observed kitten weights. Thus, the reliability of the kitten weights would be very low.

Reliability in Context

Reliability is not an inherent property of a measurement instrument but of a set of scores produced by that measurement instrument for a specified set of objects or people. The last example illustrates that the same measurement procedure can yield reliable scores for one group and unreliable scores for the other. The reliability depends greatly upon how much the scores vary across the members in the group.

We also have to be precise about how we define the measurement procedure. Do the scales automatically record the weight? Then we only need to worry about the scale error. What if a human records the weight? Then we need to take into account any errors in how the person reads the weight. Perhaps two different people operate the scales. One person waits for the scale needle to stabilize before recording the weight, while the other estimates the weight while the needle is still fluctuating. We must consider differences in operators as measurement error in this context. All aspects of the measurement procedure that may vary from one occasion to the next (except, of course, the object of measurement) should be classified as error for the purposes of estimating reliability (see Brennan, 2005). This is one reason there is such emphasis on standardizing the administration procedures for large scale ability tests.

We need to consider, too, what constitutes the outcome of interest. A certification test may have scores that range from 100 to 200, for example. People who score above 170 are considered certified. In this case, the pass-fail distinction constitutes the measurement. The reliability of interest concerns how consistently the test places people in the certified or uncertified categories. It is quite possible that the test scores (100 to 200) will demonstrate only a fairly moderate overall reliability, which might lead some people to doubt the test's usefulness. Yet even in this case, the reliability of classification could be quite high. To avoid misinterpretation, it is important to focus on the reliability that is directly relevant to how the scores will be used.

How the scores will be used will also to a great extent dictate what level of reliability is required. If we only need a rough estimate of ability, for example, we can tolerate a certain level of imprecision. Thus, we do not need a test with extremely high reliability. When scores will be used to help make high stakes decisions or fine gradations of ability, we need to maintain high precision. College entrance and licensure exams, for example, tend to have high reliability. How do we achieve higher reliability? Recall that we can always improve our estimate of a person's ability by testing the person multiple times. We can think of each item on a test as a miniature test with a certain amount of error. By increasing the number of items, we can decrease the amount of error. In general, we can increase reliability by adding more items to the test. That is why tests used to help make high stakes decisions tend to be fairly long.

One source of possible confusion with respect to reliability involves constructed-response items. More and more today, tests of human ability include short answer or essay items. These items hold out the promise of gaining more information about the candidate than what can be gleaned from multiple-choice items alone. Further, these constructed-response items more closely resemble what teachers teach in the classroom than do multiple-choice items. The major difficulty with constructed-response items is that people for the most part still need to score them. Because humans are fallible, the scores they assign to essays and other constructed-response items contain error. Compare this to scores for multiple-choice items, which we expect to

contain little or no error at all.[5] In this situation, we can talk about two distinct reliabilities. First, we have the reliability of the human rater's score in estimating an examinee's true score on an essay. Then we have the reliability of the essay score as an estimate of the examinee's true ability. These two reliabilities are distinct but related. If the rater does not score an essay reliably, the essay score cannot estimate the examinee's ability reliably. Thus, essay score reliability depends upon rater score reliability.

We can increase the rater score reliability by having more raters score an examinee's essay, until (at least theoretically) the rater reliability is perfect. At this point, the rater score would perfectly reflect the examinee's true score on that essay. Unfortunately, the essay score would still not perfectly reflect the examinee's ability. If we want to increase the essay score reliability, just as with multiple-choice tests, we need to increase the number of essays. Increasing the number of essays will raise test score reliability faster than will increasing the number of raters for each essay. A three-essay test wherein each essay is scored by a single rater is much more reliable than a single-essay test scored by three raters. Bear in mind that constructed-response items generally take longer to answer than multiple-choice items. So, in general, a highly reliable constructed-response test will have to be longer (in terms of testing time) than a multiple-choice test with comparable reliability (Wainer & Thissen, 1993).

What Are We Measuring?

Whenever we use a test score to make any kind of decision, we need to ensure that our interpretation and actions based on the test score are appropriate (Messick, 1989). In most cases, we need a test score with a certain level of reliability to be able to make any claim at all about the examinee. However, reliability alone does not ensure that our decision will be appropriate. We could have a highly reliable test of music theory, but we would probably not want to use it to select surgeons. Instead, we would want to use some test of medical knowledge. The music theory test is reliable, but it is measuring the wrong thing if we want capable surgeons.

When we talk about what a test measures, we are referring to the test's validity. Any claim based on a test score has validity issues at its center. Validity evidence comes from various sources. Does the test cover the correct content? Is the test score related to relevant outcome measures? Does the test accurately differentiate novices and masters in the field? We accumulate validity evidence by showing that the test score is related to things it should be related to, and that it is not related to things it should not be related to.

Reliability places an upper limit on validity. We can think of the reliability of a test score as the degree to which the test score can predict the true score on whatever the test is measuring. The test score should predict this true score better than it predicts anything else. To say that a test is highly reliable is to say that it predicts its true score well. If that true score turns out to represent exactly the ability in which we are interested, our test will demonstrate high validity as well.

If the true score is slightly different from but related to the ability we wish to measure, the validity evidence for the test will be weaker, although the test can still serve its purpose. As in the earlier example, if the true score for the test is not very related to what we wish to measure, then even if the test score is highly reliable it will not be useful for our intended purpose. By contrast, to say that a test score is unreliable is to say that the score is not a good estimate of the true score on whatever the test is measuring. In this case, it doesn't matter what use we plan for the test scores: We can expect that we will be unsuccessful. Thus, test score reliability is crucial whenever any kind of action will be based on that score.[6]

A Sad But True Story

In the 2004 summer Olympics in Athens, Greece, U.S. Olympian Matt Emmons had a commanding lead in the 50-meter three-position rifle final. He

[5] Just the scoring process is error-free. If we scored the same multiple-choice item several times, we would expect to get the same answer every time. On the other hand, when we talk about the multiple-choice score as an estimate of ability, it still contains error.

[6] There are unusual exceptions, of course. One could argue that military drafts and state lotteries need to be completely unreliable to be valid.

had been shooting consistently during the entire competition and stood poised to claim his second Olympic gold medal. For his last shot, he did not need a perfect shot through the bull's-eye of the target to win the gold medal. He only needed to hit the target somewhere toward the middle part.

When his turn came, he was naturally nervous. "On that shot," he said, "I was just worrying about calming myself down and just breaking a good shot." He aimed and shot. Afterwards, he was fairly confident that he had hit the target close enough to the middle to win. "When I shot the shot, everything felt fine," Emmons would later say (Associated Press, 2004). And for good reason. Unfortunately, that was only part of the story.

The shot had indeed hit the target near enough to the bull's-eye to guarantee a win, yet he failed even to place in the competition. What happened? Shooting from Lane 2, Emmons had fired at the target in Lane 3. As a result, the judges awarded him a score of zero for the round. So, although his shot was a good one, he was aiming at the wrong thing. In measurement terms, we might say that Emmons' shot was very reliable, but it had zero validity.

Lesson Learned

In testing, as in the Olympics example, consistency plays a key role in success. If we cannot measure what we want to measure consistently, then we have little hope of using those measurements for anything worthwhile. That is why testing organizations place such a large emphasis on maintaining high reliability of test scores. Perfect reliability is impossible (except in the most trivial of cases), but fortunately not necessary. Just as in any situation requiring measurement, how precise or reliable the score needs to be depends upon the use to which it will be put.

Again, as in the example, high reliability alone cannot ensure useful measurement. Responding to the question that formed the title of this article, *Is Test Score Reliability Necessary?* the answer is yes. We need to make sure that our measurement instrument measures consistently and is not just producing random numbers. However, this condition alone is not sufficient. We also need to make sure that the instrument is measuring what we think it is measuring. We need to make sure that it is aiming at the correct target.

References

Associated Press. (2004, August 22). *Emmons shoots wrong target, loses gold medal.* Retrieved from MSNBC.com on May 5, 2006, at: http://www.msnbc.msn.com/id/5785670

Brennan, R. L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement*, *38*(4), 295–317.

Brennan, R. L. (2005). *Generalizability theory*. New York: Springer.

Feldt, L. S., & Brennan, R. L. (1989). In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: American Council on Education / Macmillan.

Galilei, G. (2001). *Dialogue concerning the two chief world systems* (S. Drake, Trans.). New York: Modern Library. (Original work published 1632).

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Kane, M. (1996). The precision of measurements. *Applied Measurement in Education, 9*(4), 355–379.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Meli, D. B. (2004). The role of numerical tables in Galileo and Mersenne. *Perspectives on Science*, *12*(2), 164–190.

Messick, W. (1989). *Validity*. In R. L. Linn (ed.). *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education/Macmillan.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Sutcliffe, J. P. (1965). A probability model for errors of classification. I. General considerations. *Psychometrika*, *30*(1), 73–96.

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*(2), 103–118.

R&D Connections is published by

ETS Research & Development
Educational Testing Service
Rosedale Road, 19-T
Princeton, NJ 08541-0001

Send comments about this publication to the above address or via the Web at:

http://www.ets.org/research/contact.html