Test Reliability: A Practical Exemplification of Ordinary Language Philosophy

Robert H. Ennis
University of Illinois, Urbana-Champaign

Although *reliability* has received a greater amount of scholarly psychometric attention than any other assessment topic, its interpretation and use by the world of test consumers has been neglected. In this essay, I shall argue that there is a problem here: that the word "reliability" does not mean to the test-consuming public what it means to test specialists, resulting in avoidable confusion and likely mistakes. The test-consuming public to whom I refer consists of parents, school board members, foundation personnel, legislators, governors, presidents, and prime ministers.

The problem has become more serious as testing has assumed greater importance in our society. Most states are developing — or have developed — statewide accountability testing programs. The results are reported in the newspapers, real estate values are affected, and parents and school board members sometimes make crucial choices and exert pressures on the basis of the tests. Principals are exhorting teachers to "get those scores up." But more is coming: In the United States, for example, national accountability testing is being promoted by the National Governors' Association through its National Educational Goals Panel (NEGP), by Congress through its National Council on Educational Standards and Testing (NCEST), and by President Clinton's *Goals 2000*. The National Assessment of Educational Progress (NAEP) is moving to a much-expanded testing program, accompanied by the reporting of results by individual states. Lay people are deeply involved in this testing avalanche and frequently make judgments about the quality of the tests, the "reliability" of a test almost always being mentioned in discussions about test quality.

In this essay, I try to: (1) exhibit a significant discrepancy between ordinary usage and the psychometric meaning of "reliability" — and its unfortunate consequences, (2) show that the discrepancy extends beyond "reliability" to the associated terms "true score" and "error of measurement," (3) show that it is exacerbated by the desires for objectivity, high numbers, and ready availability, as well as by common procedures for estimating reliability, (4) examine some defenses of existing psychometric usage of the term "reliability," (5) propose alternative terminology, and (6) provide a practical example of what ordinary language philosophy was often trying to do, namely eliminate confusion and its harmful consequences.

The Discrepancy in Usage

According to my *Webster's New Collegiate Dictionary*, "reliable" means "suitable or fit to be relied on; trustworthy." Other dictionaries that I have consulted convey the same idea, my *Oxford Universal English Dictionary* supplementing these interpretations with: "in which reliance or confidence may be put; safe, sure." In my experience with people who have not been exposed to the psychometric definition of "reliable," these similar definitions fit their usage of the term. This can be encapsulated as "trustworthy," which implies "accurate" as well as "consistent,"

but certainly not only "consistent." This definition also fits the usage of many people who have been exposed to the psychometric definition, but for whom the exposure did not take.

This common meaning of "reliability" when applied to tests, is very similar to what many people call "test validity," which roughly means, in a given situation, the extent to which a test measures what it is supposed to measure. Leonard Feldt and Robert Brennan, authors of the chapter "Reliability" in the third and most recent edition of the authoritative *Educational Measurement,* admit this when they say, "In everyday conversation, 'reliability' refers to a concept much closer to the measurement concept of *validity*."[1]

The psychometric meaning of "reliability," most simply put, is "consistency."[2] More completely, Feldt and Brennan say "Quantification of the consistency and inconsistency in examinee performance constitutes the essence of reliability analysis."[3] Thus, for a test to be highly reliable it does not matter whether it measures what it is supposed to measure, or is claimed to measure, in the given situation. All that matters is that it be consistent.

A compass that reads 180 degrees off of the correct magnetic heading (indicating north when headed south, east when headed west) would then be perfectly reliable in this psychometric sense. It would, for example, consistently indicate east when headed west. If a pilot flying in the clouds over Denver then believed such a compass to be trustworthy (the common sense of "reliable") he or she might steer inadvertently westward into a mountain, thinking that the airplane was headed eastward away from the mountains. This kind of mistaken setting can occur with the type of directional gyrocompass commonly found in small airplanes. Such a compass, if so set, would not be a reliable compass in the ordinary sense of "reliable," even though it would be thoroughly consistent, and thus reliable in the psychometric sense.

A bathroom scale that consistently reads ten pounds under weight would not be considered a reliable scale in the common meaning of "reliable." It would be a thoroughly reliable scale in the psychometric sense of the term.

Similar examples are provided by reliability specialists Feldt and Brennan, using the contexts of weather reporting and medical tests. Employing what they term the "everyday conversation" sense of "reliable" they note, "Weather reports…are thought to be unreliable if they are frequently contradicted by prevailing conditions a day or two later. Medical tests are said to be unreliable if they often give false cues about the condition of a patient." But in the psychometric sense of "reliable" these predictions and tests could be quite reliable, if several meteorologists agreed (inter-rater consistency) on the predictions "day in and day out," and if the medical tests consistently yield (test-retest stability) "often-erroneous conclusions about patients."[4]

Perhaps even more significant, given the current emphasis on accountability testing promoted by the power establishment, is evidence of the use of the term "reliability" among its membership. For example, *The Wall Street Journal* used the word "reliable" in the "trustworthy" sense in an article about product studies. A

10.47925/1999.242

page-one headline states, "Studies galore support products and positions, but are they reliable?"[5] The associated article exhibits many examples of studies that are not deemed reliable by the author, but which consistently appear to show that the sponsors' products are more desirable than their competitors. These studies are thus reliable in the psychometric meaning of the term, since they are consistent, but they are not trustworthy, as the author repeatedly points out, and are judged not to be reliable by that author.

This discrepancy is known to the psychometric community and to most members of the educational research community. Some have even expressed regret about it. For example, Feldt and Brennan say, "It is somewhat unfortunate that this word ["reliability"] was adopted originally for the phenomenon under consideration."[6] I am here agreeing with them, but, in addition, urging that something be done about it.

A Problem?

Is this discrepancy a problem? Yes, it is, because in my experience the public (including the decision-makers) is almost always inclined to interpret test consistency, when labeled "reliability," as test trustworthiness. This happens even if experts warn them (which they often do not do) of the discrepancy. Linguistic habits are too inflexible, especially when attention is directed elsewhere. The standard meaning of "reliable" is too well-entrenched in the broad public mind.

It even happens to most of us in the educationist community as well. Ask yourself (without thinking very hard about the question): "Is it more important to have a reliable test than a consistent test?" My impulse is still to say "Yes," even though I have worked with the psychometric interpretation of "reliability" for forty years. I am not alone. I often hear colleagues, sometimes even psychometricians, using the word "reliable" to mean trustworthy, even in testing contexts.

The discrepancy means that a test battery that is represented as a set of educational achievement tests, and that has high "reliability" indices will be regarded by many people as a trustworthy battery of educational achievement tests, even if it fails to test for essential components of educational achievement. Current psychometric usage will move people to ignore the absence of essential components in testing programs used to judge the success of an education system. Put more generally, current psychometric usage pressures us to accept as valid, tests that in the circumstances do not measure what they are supposed to measure in the situation.

Exacerbating Factors

This discrepancy problem is exacerbated by the fact that consistency indices generally provide the highest evaluative numbers that we get about tests, often running higher than .9, compared to a frequent range of roughly .2 to .4 for correlations with readily-obtainable validity criteria, and roughly .5 to .7 for correlations with tests of a similar character (often given as evidence of convergent validity).

Numbers are considered desirable by many members of the public — and by many in education as well. They are supposedly objective (their subjectivity being hidden in the assumptions behind test-making and test-scoring decisions). The

10.47925/1999.242

combination of reliability estimates being the highest numbers (motivating test makers to emphasize them), their being numbers (desired by many), their being readily available, and their being attached to a high-sounding term ("reliability"), makes the problem more serious. As a result of these factors, reliability has often become a more powerful operative criterion in test development than validity. In my experience working with a national testing organization, this was the case.

The problem is further exacerbated by the use of internal consistency indices (such as Cronbach's alpha and the Kuder-Richardson formulas) to estimate reliability. These indices essentially indicate the degree of item intercorrelation, that is, the degree to which the items are consistently doing the same thing. One reason for the use of internal consistency indices is their ease of computation from a single administration of a test. We do not have to worry about giving the same test twice to the same population (test-retest consistency), or giving different forms to the same population (interform consistency), both of which are administratively difficult. Nor do we have to make judgment-demanding (and sometimes dubious) assignments of items to half tests that are then correlated.

Another reason for the use of these indices is that they can be pushed higher by the judicious selection and replacement of items, and by lengthening the test. The higher the indices are, the better the test looks. As a result, test makers often feel that there are definite steps that can be taken to improve a test: eliminate items that do not correlate well with others in the same test, and make the test longer.

But this selection of items to raise the internal consistency index comes at a cost, or more accurately, several costs. One is that the index no longer tells us the extent to which the test is consistent in its results from administration to administration, or form to form, which is the aspect of consistency about which I want to know for critical thinking tests, a special interest of mine. This particular cost, it might be argued, can be overcome by assuming that the item-administration set is a representative sample of a universe including items from other forms and from other item-administrations. This assumption is convenient, though speculative and often dubious.

A more severe cost comes from selecting items that will correlate highly with one another, a process that is in the interest of the person trying to secure high internal-consistency indices. But it is also a process that promotes unidimensionality. A unidimensional test does violence to many complex constructs, including, for example, most conceptions of critical thinking. The pressure exists to eliminate multiple dimensions of critical thinking, or to eliminate critical thinking altogether in a test, because critical thinking is multidimensional. I have seen this pressure operating in my experiences working with professional psychometricians.

A stratified approach to internal consistency is a possible way around this problem. It secures an internal consistency index for each dimension or part. But the process is rarely pursued, presumably because the resulting indices are lower than we would get with a unidimensional test, since the number of items for each dimension is likely to be small in the world of limited time for testing. Internal consistency coefficients are lowered by reducing the number of items.

10.47925/1999.242

If everyone thought of the process as securing consistency indices, rather than reliability indices, people in the educational community would feel less pressure to get those "reliability" numbers up. They would then feel free to focus more on whether the test is a trustworthy measure (in a given situation) of what it is supposed to measure.

"True Score" and "Error of Measurement"

The same problem exists for the psychometric terms "true score" and "error of measurement." All three terms are defined in terms of each other, with consistency (or lack of it) being the basis for all. But because these terms are presented to the public much less frequently than "reliability," the problem is not so severe for "true score" and "error of measurement."

In the section on reliability in a recent version of *Standards for Educational and Psychological Testing* "reliability" is defined in terms of error of measurement: "Reliability refers to the degree to which test scores are free from errors of measurement."[7] Furthermore, "error of measurement" is defined in terms of true score: the error of measurement equals the obtained score minus the true score. These definitions seem to be in accord with the ordinary notion of reliability, trustworthiness.

But no. True score is not what we might expect from the common meaning of the words, according to which a person's true score on a test of X would be a perfectly accurate score reflecting the person's competence on X, or degree of possession of the trait X. Rather "true score" in psychometric language can be viewed as the average score that the person would get if the person took the same test over and over again — without getting tired. There are definitional differences within the psychometric community, but in all cases true score is a consistency concept. It is not concerned with the extent to which the person has the quality supposedly being tested.

From this it follows that error of measurement is also a consistency concept, since it is defined in terms of the obtained score and the true score. In the psychometric sense, error of measurement is simply inconsistency of measurement. This conflicts with what we might expect from the standard meaning of the words "error of measurement," in accord with which a measurement is in error to the extent that it fails to reflect accurately what is being measured. According to psychometric usage, the type of bathroom scale mentioned earlier, the one that continually gave a reading ten pounds low, would not be exhibiting error of measurement.

As a result, attempts to define "reliability" to the public in terms of error of measurement and true score will reinforce their mistaken assumption that experts are talking about trustworthiness when they talk about reliability. So these definitions and labels will further interfere with the public's understanding of what is going on in testing.

Alternative Terminology

In view of the difficulties attendant upon the psychometric interpretations of "reliability," "true score," and "error of measurement," I urge that we in the

10.47925/1999.242

educational research community assign different words to the psychometric concepts to which these terms are presently assigned. It seems clear that "reliability" should be replaced by "consistency," since the latter is such a perfect fit for the concept. We would then have test-retest consistency, split-half consistency, inter-rater consistency, and internal consistency as topics of concern.

A possible alternative to "true score" would be "consistent score." For "error of measurement" we could substitute "inconsistency of measurement."

Possible Reasons for Not Changing Terminology

One reason that some will have for not changing terminology is that people in the field have developed technical linguistic habits that are embedded in their minds. Another is that the literature is all expressed in the current psychometric language. Literature reviews, computerized library searches, and understanding the literature would be more difficult. But these are surely minor considerations when there are serious threats to the education of tens of millions of students that can result from the equivocation inherent in the psychometric use of these terms. That is, the neglect of important things on batteries of influential educational tests for tens of millions of students will be hidden behind the word "reliable," as well as "true score" and "error of measurement." A few thousand psychometricians can handle these inconveniences, given the complexities that they have already demonstrated they can handle.

Another possible reason for not changing is the belief that word meanings are arbitrary; that is, that we should have full freedom to use words any way we choose, particularly if they are technical terms. According to this view, it does not really matter what you mean by a term, just so long as you tell others how you are using it, a position expressed in Lewis Carroll's *Through the Looking Glass*: "When I use a word, it means just what I choose it to mean — neither more nor less."

In an ideal world where all people would communicate their word meanings effectively, and where all people are flexible enough to remember and interpret others' words as they are intended, this doctrine might be workable. But the public is not that flexible, and psychometricians are not that effective in communicating their meanings anyway. The total-flexibility doctrine has limited application. It does not apply here.

Summary and Comment

In this essay I have exhibited discrepancies between the ordinary meanings and the psychometric meanings of the terms "reliability," "true score," and "error of measurement." The discrepancies are known in the field of psychometrics, and the discrepancy for at least "reliability," the most serious one, is also known among most educational researchers.

I urge the consciousness-raising of the educational research community with respect to the public's (and possibly our own) real understanding of these terms (and hope that you will join me in this consciousness-raising effort), and argue that continued use of these terms as they are now used in psychometrics will injure students because of inevitable misinterpretations. In particular, people will be

10.47925/1999.242

inclined to think that tests having some type of consistency are therefore trustworthy — even though the test might be a very untrustworthy measure of the construct, or whatever it is claimed to measure, in the situation. Schools, students, teachers, and administrators can then be rewarded for the wrong things and pressured to neglect important things. The problem is particularly important now that testing is assuming such a significant role in our educational system — with lay people making decisions on the basis of tests, and judging the tests in part on the basis of their alleged reliability.

A second unfortunate type of consequence (for which ease of calculation and administration, and eagerness to secure high numbers are also responsible) is the pressure to shape tests so that they are unidimensional. This results in omitting important aspects of multidimensional concepts like critical thinking, or omitting multidimensional concepts from our testing programs altogether. Multidimensionality is an enemy of internal consistency, though not of authenticity or importance.

An obvious replacement for the term, "reliability," is the term "consistency." Possible replacements for the terms "true score" and "error of measurement" are "consistent score" and "inconsistency of measurement." These replacements would eliminate the false and misleading advertising inherent in the terms "reliability," "true score," and "error of measurement." Some will think that all this is a mere verbal dispute, mere semantics. They are in part correct. These are verbal and semantic matters. But the "mere" is inappropriate. Language is a powerful implement. It can communicate, or it can obfuscate and cause damage. Which is it to be in this case?

In presenting this essay, I hope in addition to have successfully exemplified the sort of thing that ultimately many ordinary language philosophers of the fifties, sixties, and seventies were trying to do: eliminate confusion and consequent problems by focusing on the ordinary meaning of crucial terms. If you think that I have, in this essay, done this, I shall be pleased. If you think that it is a worthwhile thing to do, I shall be even more pleased.

HELPFUL SUGGESTIONS HAVE BEEN MADE by Terry Ackerman, Kathleen Black, Michelle Commeyras, Helen Ennis, Sean Ennis, Peter Griffin, Bruce Haynes, Laird Heal, Clarence Karier, Mark Larson, Ken Nafziger, and Joseph Shively.

---

1. Leonard Feldt and Robert Brennan, "Reliability," in *Educational Measurement*, 3d. ed., ed. Robert L. Linn (Washington, D.C.: American Council on Education; Macmillan, 1989), 106.

2. See, for example, Lee J. Cronbach, *Essentials of Psychological Testing*, 2d. ed. (New York: Harper, 1960), 126.

3. Feldt and Brennan, "Reliability," 105.

4. Ibid, 106.

5. Cynthia Crossen, "Studies Galore Support Products and Positions, but are they Reliable?" *The Wall Street Journal*, 1 November 1991, A1.

6. Feldt and Brennan, "Reliability," 106.

7. Committee to Develop Standards for Educational and Psychological Testing, *Standards for Educational and Psychological Testing* (Washington, D.C.: American Psychological Association, 1985), 19.

10.47925/1999.242