Trustworthiness and Consistency

Stephen P. Norris
University of Alberta

Robert Ennis urges that the psychometric community replace its central technical terms, "reliability," "true score," and "error of measurement," with terms from ordinary usage that capture the same meanings but do not have the same misleading connotations. The recommended new terms are "consistency," "consistent score," and "inconsistency of measurement," respectively. He argues that current psychometric linguistic practice distorts the public's consumption of test results, because the public takes reliability, true score, and error of measurement to be about the trustworthiness and accuracy of test results. According to psychometricians, reliability, true score, and error are about consistency, not accuracy. There is thus the risk that the public might infer from published characteristics of tests that they are accurate when they perhaps are merely consistent. Ennis argues that the problem is exacerbated by the fact that linguistic practices are so inflexible to the adoption of technical meanings for ordinary words that even psychometricians fall frequently into the trap of using their technical terms with ordinary meanings, for example, of using "reliable" when they mean "trustworthy." Furthermore, he points to several other possible negative consequences for education of such linguistic inflexibility.

I have discussed these issues with Bob Ennis over many years. I am sympathetic with his concerns and share many of his conclusions. His paper has given me the opportunity to think about the issues once more, having removed myself from them for a long time. From that more distanced perspective, I wish to raise a number of questions about the paper that are intended to push his thoughts further:

1. Does existing psychometric terminology stifle debate about the quality of tests, that is, "[pressure] us to accept as valid, tests that in the circumstances do not measure what they are supposed to measure?"
2. Does the process of attempting to secure high internal consistency "[promote] unidimensionality" in tests?
3. If the suggested modifications to terminology were adopted, would there be "less pressure to get [the consistency] numbers up" as there currently is to get the reliability numbers up, and should there be?
4. Is not consistency one of the best indicators of trustworthiness that exists?

Stifling of Debate

A test that produces consistent performance from examinees is not necessarily a trustworthy measure of what it is advertised to measure. As Ennis correctly points out, a compass that points consistently 180 degrees from the correct heading cannot be trusted to lead one homewards (unless one knows of the bug), any more than a test on which examinees consistently perform but use only rote recall can be trusted to indicate mathematical reasoning ability. However, does the current psychometric

practice of using "reliability" to mean "consistency" move people to ignore problems with measures that are otherwise consistent? This question points to an empirical issue in two dimensions: first, whether people do ignore other essential components of tests when they are advertised as reliable; and, second, whether it is psychometricians' linguistic usage that causes any such ignoring. Ennis has provided no adequate evidence to support either of these conclusions, though he has suggested some intuitive considerations that might support a suspicion they are true.

My intuitions are that it would be difficult indeed to gain acceptance from the educational community, and from other consumers of tests, for measures that met only the criterion of reliability. For instance, it would be very difficult for a reliable measure that contained as items only simple addition problems to gain acceptance as a test of high school algebra; it would be difficult for a reliable measure that asked examinees to provide only their names, addresses, and dates of birth to gain acceptance as a biology test; it would be difficult for a reliable measure whose tasks involved listing the 50 states of the union and their respective capitals to gain acceptance as a critical thinking test. Even less extreme deviations from quality tests would also draw criticism, in my judgment. For example, it would be difficult for a reliable measure that poorly represented the content of a history course to gain acceptance as a test of achievement in that course. Why do I believe it would be difficult for such tests to gain acceptance? Test consumers use criteria that enable them to judge between tests that appear at least approximately right and those that appear dreadfully wrong. They also have criteria, such as one covering representativeness of content, that allow them to distinguish among tests that have more nearly equal levels of quality. They are not able to be fooled so easily as to be swayed by a claim of reliability when a test appears awry according to their criteria of quality tests. Claims of reliability will not ward off challenges when tests have other features that appear suspicious. These are my intuitive judgments, supported by the same amount of evidence as Ennis's. Clearly, we need evidence concerning these matters. Will a claim for reliability ward off challenges to a test when nothing appears awry? I believe, as does Ennis, that perhaps it will. However, if a test is reliable, and if nothing else appears awry, then the test has a number of qualities other than consistency in its favor already.

In general, I do not view the critical discourse about testing to be as barren as Ennis appears to view it. In the educational contexts that I have witnessed, when high-stakes test results are under consideration, there are always critics knowledgeable of testing. The critics reside in the teaching force, in the highly competent personnel of education district offices, in departments of education, and in universities. These individuals know that reliability is only one criterion of test quality used by the testing field, and therefore typically look for more. I simply am not as convinced as Ennis of the influence of one word, or of the influence of altering it to some other word, on the critical discourse surrounding test results.

Unidimensionality

Ennis claims that the process of creating high internal consistency in tests, a process driven by a desire to increase lower-bound estimates of consistency between

10.47925/1999.249

administrations of parallel forms, is also a process that promotes unidimensionality in tests. Unidimensionality is a characteristic of tests that measure a single psychological dimension. Most constructs of educational significance contain many dimensions, so unidimensional tests are inappropriate for many educational purposes.

It is a common error in thinking about tests to conflate the properties of dimensionality and internal consistency. The error arises because unidimensionality implies high internal consistency, but high internal consistency does not imply unidimensionality. It is not only possible, but easy, to obtain high internal consistency in tests that are multidimensional.[1] Therefore, striving for high reliability, as psychometricians use the term, is not an impediment to developing high quality tests of complex educational constructs, including critical thinking. Ennis's worries should be allayed on this matter.

Eased Pressure for Consistency

Ennis believes that if everyone thought of calls for what is now called "reliability" as calls to secure consistency, then there would be less pressure to get high numbers on this test characteristic. It is not clear to me that this would be so, or that it should be so. First, I cannot imagine people being satisfied with tests that give inconsistent results. How would the inconsistent results be interpreted if what people thought they were measuring were traits that endured over given periods of time? Second, why should individuals be satisfied with tests with lower consistency when consistency is necessary for trustworthiness, as Ennis and I once argued?[2] Perhaps he has changed his mind on this conclusion, but I hope that he has not. How can I trust a test to be telling me anything if, allowing for a reasonable range of inconsistency, subsequent administrations are as likely to fall above or below that range as within it?

Trustworthiness and Consistency

I wish finally to explore the idea that consistency in tests, including internal consistency, is far more closely tied to the trustworthiness of those tests than Ennis acknowledges. I agree that a consistent test need not be a trustworthy measure of what it is intended to measure. Nevertheless, once internal consistency has been established, we are in a position to explore other consistencies related to the trustworthiness of the test. For example, suppose we desired to explore the question of whether differences in format, say between constructed-response and multiple-choice formats, affect what critical thinking tests measure. To examine in a robust manner such a question it is important to begin with tests in each format that individually display high internal consistency. An important constraint in such a study of critical thinking tests is that their internal consistencies tend to be low by comparison to other psychological tests. However, supposing an adequate beginning point is established, then it is possible to explore consistency on other levels, for instance, to explore whether performance on multiple-choice tests is more consistently related to performance on other multiple-choice tests than to performance on constructed-response tests. Knowledge of consistency on such matters is largely what trustworthiness of tests is based upon, because it is such consistency

10.47925/1999.249

that allows us dependably to infer what performances on tests mean. Without internal consistency, which is at the basis of all else, meanings assigned to performances on tests can be little more than arbitrary.

I THANK Mark Gierl for suggestions helpful in preparing this response.

1. Samuel B. Green, Robert W. Lissitz, and Stanlen A. Mulaik, "Limitations of Coefficient Alpha as an Index of Test Unidimensionality" *Educational and Psychological Measurement* 37 (1977): 827-38.

2. Stephen P. Norris and Robert H. Ennis, *Evaluating Critical Thinking* (Pacific Grove, Calif.: Midwest, 1989), 44.

10.47925/1999.249