Working Within the Limits: Thoughts on Stein and Heikkinen

Theo Dawson

As a woman on an assessment mission, I echo many of the concerns raised by Stein and Heikkinen. It is important for those of us who use assessments to ensure that they (1) measure what we say they measure, (2) measure it reliably enough to justify claimed distinctions between persons, and (3) are used responsibly. It is relatively easy for testing experts to create assessments that are adequately reliable for individual assessment, and although it is more difficult to show that these tests measure the construct of interest, there are reasonable methods for showing that an assessment meets this standard. However, it is more difficult to ensure that assessments are used responsibly.

Few consumers of tests are aware of their inherent limitations. Even the best tests, those that are highly reliable and measure what they are supposed to measure, provide only a limited amount of information. This is true of all measures. The more we hone in on a measureable dimension—the greater our precision becomes—the narrower the construct becomes. Time, weight, height, and distance are all extremely narrow constructs. This means that they provide a very specific piece of information extremely well. When we use a ruler, we can have great confidence in the measurement we make, down to very small lengths (depending on the ruler, of course). No one doubts the great advantages of this kind of precision. But we can't learn anything else about the measured object. Our measurement cannot tell us what the object is, how it is shaped, its color, its use, its weight, how it feels, or how attractive it is. We only know how long it is. To provide an accurate account of the thing that was measured, we need to know many more things about it, and we need to construct a narrative that brings these things together in a meaningful way.

A really good psychological measure is similar. The LAS, for example, is designed to go to the heart of development, stripping away everything that does not contribute to the pure developmental "height" of a given performance. Without knowledge of many other things—such as the ways of thinking that are generally associated with this "height" in a particular domain, the specific ideas that are associated with this particular performance, information from other performances on other measures, qualitative observations, and good clinical judgment—we cannot construct a terribly useful narrative.

And this brings me to my final point: Formal measures, no matter how many great ones we design, should always be employed by knowledgeable mentors, clinicians, teachers, or coaches as a single item of information about learners that may or may not provide useful insights into their needs. Consider this relatively simple example: a given 2-year-old may be tall for his age, but if he is somewhat under weight for his age, the latter measure may seem more important. However, if he has a broken arm, neither measure may loom large—at least until the bone is set. Once the arm is safely in a cast, all three pieces of information—weight, height, and broken arm—may contribute to a clinical diagnosis that would have been difficult to make without any one of them.

It is my hope that the integral community will choose to adopt high standards for measurement, then put measurement in its place—alongside good clinical judgment, reflective life experience, qualitative observations, and honest feedback from trusted others.

Appendix: Reliability – Some Basics

There is a great deal of confusion in the assessment community about the interpretation of statistical reliability. This confusion results in part from the different ways in which researchers and test developers approach the issue. Researchers learn how to design research instruments, which they use to study population trends or compare groups. They evaluate the quality of their instruments with statistics. One of the statistics used is Cronbach's Alpha, an indicator of statistical reliability that ranges from 0 to 1. Researchers are taught that Alphas above .77 or so are acceptable for their instruments, because this level of reliability ensures that their instrument is measuring real differences between people.

Test developers use a special branch of statistics called psychometrics to build assessments. Assessments are designed to evaluate individuals. Like researchers, test developers are concerned about reliability, but for somewhat different reasons. From a psychometric point of view, it is not enough to know that an assessment measures real differences between people. Psychometricians need to be confident that the score awarded to an individual is a good estimate of that particular individual's true score. Because of this, psychometricians set higher standards for reliability than those set by researchers.

The table below will help to clarify why it is important for assessments to have higher reliabilities than research instruments. It shows the relationship between statistical reliability and the number of distinct levels (strata) a test can be said to have. For example, an assessment with a reliability of .80, has 3 strata, whereas an assessment with a reliability of .94 has 5.

Reliability	Strata
.70	2
.80	3
.90	4
.94	5
.96	7
.97	8
.98	9

Strata have direct implications for the confidence we can have in a specific person's score on a given assessment, because they tell us something about the range within which a person's true score would fall, given a particular score. Imagine that you have taken a test with a scoring range of 0 to 500 and a reliability of .94. The number of strata into which this assessment can be divided is 5, which means that each strata equals about 100 points on the 500 point scale. If your score on this test is 350, your true score is likely to fall within the range of 300 to 400.[1]

Statistical reliability is only one of the ways in which assessments should be evaluated. Test developers should also ask how well an assessment measures what it is intended to measure. And those who use an assessment should ask whether or not what it measures is relevant or important.

Resources

Guilford J. P. (1965). *Fundamental statistics in psychology and education*. 4th Ed. New York: McGraw-Hill.

Kubiszyn T., Borich G. (1993). *Educational testing and measurement*. New York: Harper Collins.

Wright B. D. (1996). Reliability and separation. *Rasch Measurement Transactions*, 9, 472.

Theo Dawson *is the founder of Developmental Testing Systems and creator of the Lectical Assessment System. Her dissertation demonstrated the power and utility of a novel methodology that makes it possible to describe conceptual development in any domain of knowledge without the expense of conducting longitudinal research.*
Email: Theo@devtestservice.com

[1] This range will be wider at the top and bottom of the scoring range and a bit narrower in the middle of the range.