Tests in Search of Constructs versus Constructs in Search of Tests:

Clinical Test Selection for Research Purposes

R.W. Kamphaus

Georgia State University

Abstract

Clinical tests are abundant, making selection of a measure for use in a research program seem deceptively simple. Since I believe that this assumption is not tenable I will offer three conceptual principles that researchers may wish to consider in the process of test selection. First, I caution researchers to carefully specify the constructs to be assessed and then begin the search for tests that measure said constructs. I began thinking about this first premise in the 1980s when I was evaluating the validity of inferences made based on the then popular draw a person, house, tree, or related drawing tests used in clinical practice with young children. I discovered that these tests were considered to assess a variety of constructs ranging from intelligence, to personality, to visual-motor coordination. I read about good reliability of the test scores and found correlations between some draw a person and intelligence measures in the .50 range. In a review article I noted, however, that although this is a statistically significant correlation it is nevertheless consistently lower than the typical criterion-related validity coefficient between intelligence tests of about .70 or higher. It then struck me that these measures indeed had some psychometric evidence to support them but that this same evidence base was poor when compared to measures designed to assess the intelligence, personality, and visual-motor skill constructs apriori. I then concluded that drawing tests were indeed tests but relatively poor measures of important developmental, characterological, or cognitive constructs - they were

indeed decent tests that were nevertheless still in search of a construct to measure. I will say more about this consideration soon. Second, I think that it is best to determine if the tests and their score inferences meet the basic criteria set forth in the Standards for Educational and Psychological Testing of the AERA, APA, and NCME. And third, I counsel researchers to avoid developing customized measures as this process requires a lengthy research program of its own.

Construct Specification

Scientific progress is linked to better definition and measurement of constructs or, as is worth repeating Edgar Doll's opinion from 1953 (p. 60), "*The problem of definition is an embarrassment to all of science."* The process of construct definition, which is frequently not emphasized, must precede that of development or selection of a measurement tool or device. Later, in the 1960s, Oscar Buros identified the typical considerations used by professionals to select tests, which did not include construct definition.

> At present, no matter how poor a test may be, if it is nicely packaged and if it promises to do all sorts of things which no test can do, the test will find many gullible buyers. When we initiated critical test reviewing (1938) we had no idea how difficult it would be to discourage the use of poorly constructed tests of unknown validity. Even the better informed test users who finally become convinced that a widely used test has no validity after all are likely to rush to use a new instrument which promises far more than any good test can possibly deliver… Highly trained psychologists appear to be as gullible as the less well trained school counselor. It pays to know only a little about testing; furthermore, it is much more fun for everyone concerned the examiner, examinee, and the examiner's employer. O. K. Buros. (1961). Tests in Print: A Comprehensive Bibliography of Tests

for Use in Education, Psychology and Industry. Highland Park, New Jersey: Gryphon Press.

Consequently, a first step that I recommend in construct specification is to determine the extent to which one seeks to assess "latent traits," versus skills, competencies, or other types of constructs. "Latent" traits are labeled as such because they cannot be observed or, as a physician might say, palpated. Examples of latent traits include anxiety, intelligence, spatial ability, and vision – yes, vision. Although it sounds implausible, "vision" cannot be "seen," and is only inferred from test score results. Consider, for example, the ubiquitous Snellen Chart used to screen for vision problems as cited by John Carroll (1993). The Snellen Chart does not "look" like a vision test. In fact, to a four or five year old child it is not a test of vision, but rather is a test of English language letter recognition skills, that is, a variant of a reading test. As it turns out, the content, or "content validity," of a test of a latent trait is not very important as exemplified by the Snellen Chart. As stated in the peer developed guidelines provided in the *Standards for Educational and Psychological Testing*, the determination of whether the Snellen Chart, or any other assessment is a good measure of its intended latent trait is based on the logical network of evidence gathered to support its test score inferences (AERA, APA, NCME, 1999). If frequency of test use is any indication, the Snellen Chart is considered by Opthalmologists and optometrists to be a good screening measure of the latent trait of general visual acuity.

The distinction between test content and validity for the assessment of a latent trait is a subtle and challenging one. It defies the typical human intuition that the content of a test item should be crucial for determining the validity of test score inferences. It is as challenging as, and

not unlike, the counterintuitive, but well supported principle of negative reinforcement, as any instructor of an introduction to psychology or educational psychology course can attest.

There are, however, some testing domains where content validity is crucial; as is the case for the assessment of academic achievement domains (Lissitz & Samuelsen, 2007). In fact, Lissitz and Samuelsen have elevated content validity to be one of three key internal test components that constitute evidence of construct validity, the other two being latent process and reliability. R. L. Thorndike contrasted academic achievement testing with latent trait assessment by referring to the former as "domain mastery" assessment. Content validity is crucial in domain mastery achievement testing as explained by Thorndike (1982).

> The fruitfulness of the orientation in terms of domain mastery depends on the possibility of defining a domain clearly and incisively, so that the range of performances that lie within the domain can be fully specified and agreed on. It should then be possible to sample tasks from that domain in such a way that the complete domain is adequately represented and inferences about completeness of mastery of the domain are a reasonable possibility. The approach really applies only to aspects of school achievement. (p. 2)

Further distinctions between latent trait and domain mastery assessment exist in the assumptions generated from one's performance on a specific test. Changes in domain mastery test performance are presumed to be affected by teaching and schooling, whereas latent trait test performance is presumed to be less amenable to mild or short-term factors. In contrast, latent traits are often, although not always, presumed to be stable over time, particularly short periods of time.

If a researcher seeks to measure domain mastery then the selection of test content should be, as Thorndike suggests, "incisive." I prefer the term "authoritative" to incisive in that I think

definition of the test content should be informed by recognized curricular domain experts and/or the curricula guidelines or standards produced by respected learned societies. The blueprint for the KeyMath-3 Diagnostic Assessment (DA) for example, measures the following elementary and middle school mathematics skill areas based on U.S. national standards.

- Numeration
- Algebra
- Geometry
- Measurement
- Data Analysis and Probability
- Mental Computation and Estimation
- Addition and Subtraction
- Multiplication and Division
- Foundations of Problem Solving
- Applied Problem Solving

[These subtests are grouped into three areas: Basic Concepts (conceptual knowledge), Operations, (computational skills), and Applications (problem solving)].

The Key Math-3's content blueprint reflects the content and process standards described in the National Council of Teachers of Mathematics's (NCTM's) *Principles and Standards for School Mathematics* (NCTM, 2000). The existence of this content blueprint means that scores derived from the KeyMath subtests and composites may be defended as estimates of mathematics domain mastery.

Gronlund (2003) suggested that academic achievement measures may be further separated into two types; those seeking to assess student progress, or formative assessments, and

those that assess knowledge and skills acquired, or summative assessments (Gronlund, 2003). Sample achievement assessments used by special educators and psychologists that may be formative include: curriculum based measures and multilevel survey achievement test batteries (e.g. ITBS). Summative assessments may include individually administered diagnostic tests or the same survey multilevel achievement tests.

Curriculum based measurements of academic achievement are typically of the formative variety in that they are intended to be brief, and designed to determine whether or not a child is making adequate academic progress (Jiban & Deno, 2007). Clinical, or individually administered comprehensive and screening measures of academic achievement, may be either summative or formative. Screening measures, such as those that assess only a few constructs such as spelling, word reading, and mathematics achievement and use only a few items per subtest to do so, for example, 50 to 70 items or less, are more likely to be useful for formative interpretation. This is so because the individual subtests do not have strong evidence of content validity to support inferences regarding summative knowledge or skills acquired. Clinical measures with detailed content blueprints for their subtests, as is sometimes revealed by options for extensive error analyses may be used for summative assessment purposes, such as may be claimed for the KeyMath-2 DA cited earlier. In summary, researchers are advised to study test content blueprints to determine whether or not the test's scores are most useful for summative (strong content blueprint), formative, or both assessment purposes.

Meeting Standards of Practice

As the legal consultants have been telling the American Psychological Association for many years now, peer developed guidelines or standards, such as those provided in the *Standards for Educational and Psychological Testing* developed jointly by AERA, APA, and NCME, and

published in 1999, quickly became adopted as legal standards of practice. As such, they cannot be ignored. Speaking more nobly, however, these standards were developed by our most august measurement peers, making them important guideposts for test selection, use, and development. They derive much of their influence from the fact that core aspects of the test standards have withstood the test of time. For example, the elucidation of basic principles of construct validty traces its roots to the seminal 1955 work of Cronbach and Meehl.

It is lamentable, however, to discover how few Ph.D. students in special or general education, the various fields of psychology, or related disciplines have ever been required to learn the standards as part of a course. This practice stands in contrast to ethical and other standards, or federal regulations, which are commonly required as course readings. Consequently, many scientists do not provide evidence of understanding the most basic principles in their grant proposals or publications. The inclusion of the phrase, "valid test," intimating that validity is a property of a test per se, is a common indicator of a lack of knowledge of the generally well-reasoned test standards. As noted in the standards for educational and psychological testing, and I quote,

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests…The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretation of test scores required by proposed uses that are evaluated, not the test itself. (p. 9)

Furthermore, the test standards provide a useful framework for test development that may guide the process of, and present a convincing rationale for, the developmental steps and

procedures proposed for a project. For example, the test standards define the following indicators as sources of validity evidence:

- Evidence based on test content
- Evidence based on response processes
- Evidence based on internal structure
- Evidence based on relations to other variables
- Evidence based on consequences of testing

Many measurement scientists have lamented a continuing lack of measurement training among social science researchers. It is not clear that most doctoral education and social science programs require an introduction to measurement science or measurement theory course, in addition to the standard statistical coursework in ANOVA, HLM, regression, etc. I know of only one journal that requires adherence to the test standards in its publication guidelines, and that is *Educational and Psychological Measurement.*

Intricacies of Test Development

A strong program of test validation is characterized by the prominent role construct validity takes in each stage of instrument development as noted by Benson in her 1998 article. In many ways Benson presaged the revision of the test standards by providing a three-stage framework to guide instrument development, using a strong program of test validation drawn from writings of Loevinger in 1957 and Nunnally in 1967 among others. Her three stages include: a substantive component, a structural component, and an external component. The substantive component is where the theoretical domain of the construct is specified and then operationally defined in terms of the observed variables, for example, measurable behaviors that represent the construct. The structural component involves relating the items to the structure of

the construct by determining to what extent the observed variables relate to one another and to the construct. Finally, the external component begins to give meaning to the test scores by determining whether or not the measures of a given construct relate in expected ways to measures of other constructs.

Whether using this test development model or another, the process is lengthy, expensive, and arduous, ad this is my third point. Without referring to the specific steps and studies involved in detail, speaking from personal experience, it has taken me and my co-authors 7 years to create a behavior rating scale, and another 4 years revising it. Similarly, we spent 4 years creating a clinical test of intelligence, and a behavioral and emotional screener that we conceptualized in 1986 was finally published in 2007. Having served as the research program director for several well regarded tests as well, I can attest to the hundreds of thousands of dollars, or in some cases millions of dollars, necessary to conduct the many studies required to aspire to the strictures of the test standards.

Furthermore, as is the case with all technological change, the test development bar has been raised considerably over the years to include more studies of differential item functioning or DIF analyses, studies of inter-item and inter-test dependence, cross cultural and linguistic equivalence, and creation of sophisticated software programs for test scoring and interpretation. Unfortunately, many published and popular tests used by researchers do not include all of these analyses in their technical manuals.

Biases Laid Bare

Given time constraints, and in light of my three considerations, I would like to share my biases for the criteria that I tend to apply when reviewing the instrumentation section of IES or other grant proposals. I often ask myself the following questions:

1) If a researcher uses the term "valid test" in her or his narrative, how is it possible for the researcher to have a sufficient understanding of relevant measurement concepts and the test standards?

2) Why did the researchers select a specific test when there are at least a half dozen newer measures of the same construct available, all with more detailed validity evidence in their manuals?

3) Why did the researcher propose to use a certain clinical assessment as a summative measure when the assessment purpose expressed in the narrative calls for a formative measure?

4) Why do some of the test outcome variables, vocabulary knowledge, for example, seem unrelated to the constructs under study, for example, social science knowledge?

5) If the researchers propose to use only portions or selected items from a parent measure, do they present evidence that this adaptation does not affect construct measurement?

I must admit that my most commonly inadequately answered question has to do with qualifications of the researcher and consultant team. I ask whether or not the researchers have published articles in measurement science journals, thus invoking peer review as a measure of test development expertise. The journal outlets that I look for are *Journal of Educational Measurement, Psychometrika, Multivariate Behavioral Research,* and *Educational and Psychological Measurement* among others. I also find it curious when a proposal involves test development as its centerpiece, but the measurement scientist is only participating as a consultant or a distant member of the research team. This disjuncture is often evident in the narrative where the expertise of the measurement scientist is not reflected in the various sections of the proposal, but is reflected in the research plan.

I realize that my points are basic to many of you and to this group of researchers I apologize. I often find, however, that a lack of grasp of the basics, as is often the case in the arts, sports, or other endeavors, may differentiate successful from unsuccessful research or research proposals. I propose that addressing issues of construct definition and specification of underlying assumptions about the construct, adherence to standards of test development, selection, and usage practices, and choosing to develop new measures only when necessary and when adequate expertise can be accessed, represent but a few of the basic considerations in applied measurement in the conduct of education science. In this regard I will end with a quote from author Jim Rohn:

> Success is neither magical nor mysterious. Success is the natural consequence of consistently applying the basic fundamentals.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice, 17,* 10-22.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor analytic studies.* New York: Cambridge University Press.

Gronlund, N.E. (2003). *Assessment of student achievement (7th Ed.)*. Boston, MA: Allyn & Bacon.

Kamphaus, R. W., & Pleiss, K. (1991). Draw-a-person techniques: Tests in search of a construct. *Journal of School Psychology*, *29*, 395-401.

Jiban, C.L., & Deno, S.L. (2007). Using math and reading *curriculum-based measurements* to predict state mathematics test performance: Are simple one-minute measures technically adequate*?* *Assessment for Effective Intervention*, 32, 78-89.

Lissitz, R.W., & Samuelsen, K. (2007). A suggested change in terminoloty and emphasis regarding validity and education. *Educational Researcher, 36*, 437-448.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Spearman, C. (1927). *The abilities of man.* New York: Macmillan.

Thorndike, R.L. (1982). *Applied psychometrics.* Boston, MA: Houghton Mifflin.