25. *EXPLAINING VALIDITY AND RELIABILITY*

"Validity" and "reliability" are ubiquitous terms in social science measurement. They are prominent in the APA "Standards" (1985) and earn chapters in test theory texts. Yet they remain ambiguous and confusing in both theory and practice.

QUALITATIVE AND QUANTITATIVE ASPECTS OF VALIDITY

Validity has both qualitative and quantitative aspects. The qualitative aspects are conceptual. The quantitative aspects are numerical.

Fundamental to validity is the concept of a variable. A variable is intended to be a undimensional manifestation of one clear idea. It is the embodiment of an intention that is defined by the items written to implement the idea. Successful realization of a variable results from its concrete representation in items. This representation is then used to collect data from which the coherence and utility of the idea and its items is determined.

QUALITATIVE VALIDITY

The conceptual plan of a variable is its qualitative validity. Qualitative validity refers to the abstract idea of a variable and its content and illustrations that transform the abstraction from an idea to manifestation by items.

The qualitative aspects of a variable are its "content" and "construct" validity. These two forms of validity express the meaning of the variable. They explain the organization and construction of items and their use in eliciting manifestations of the variable. The items carry the intent of the variable and bring to concrete realization the abstract intention.

Content validity is demonstrated by items planned and written to bring the variable to life. But successful item writing must follow a plan. The plan requires thinking of the variable as a line with direction (an arrow) and arranging items according to their difficulty along this line. Such a plan follows a single dimension. The plan requires empirical confirmation. A dialogue results between the tasks of writing items according to the plan and gathering empirical evidence. Variable development requires this dialogue first to construct a useful variable and then successively to substantiate it. A variable by definition is undimensional. Its items operationally define the variable. This activity addresses the "content" validity question.

When such a plan and its resultant items are a useful representation of a variable, the items operationally define the variable and the variable is manifest by responses to the items. If the structure of the variable is supported by the item calibrations and if the characteristics of persons can be substantiated by their placement on the variable, we have construct validity.

QUANTITATIVE VALIDITY

The quantitative aspects of validity are numerical and statistical. They are content-free and devoid of text. They apply to any variable whatever its intention. They have no meaning by themselves, but are useful tools for crafting a variable and defining its numerical properties.

THE DIALECTIC BETWEEN QUALITATIVE AND QUANTITATIVE

The qualitative and quantitative interact in a dialogue between the idea of a variable and its quantitative manifestation. The idea provokes a plan and the writing of items. The calibration of items and measurement of persons provides the empirical realization. The process involves successive cycles modifying the plan of the variable and the items until perturbations between the two reduce to a point where further refinement is either unnecessary or not possible.

We might wish this dialectic to reach perfect agreement. But should we ever reach that level, the problem would become trivial or disappear altogether. Indeed, we do not expect these two aspects to fully reconcile. That would be the death of discovery. It would mark the end of improvement to the variable.

Instead, the result of this interaction produces something more than is possible by considering only the plan or only the results. Actually, we know there will and must always be a sense in which these two aspects of validity are never brought together. They can never achieve a perfect union, because it is precisely their *interaction* that engenders our motivation to press on.

INTEGRITY AND UTILITY VALIDITY

The two "validities," qualitative and quantitative, are often applied to test practice to determine concomitant and predictive utility. The Standards call these "criterion" and "predictive" validity.

Their designations as validity, however, are not apt because they do not bear on the integrity of the test. They only bear on the test's utility. It is useful to differentiate between integrity and utility.

By *integrity* we mean the sine qua non of the variable, the demonstration through construction of the variable intended, illustrated by items (qualitative validity) and supported by calculations (quantitative validity).

By *utility* we mean the application of the variable to whatever circumstances appear useful for investigating relationships between this variable and others.

Discussions about utility arise from applications of the variable to any number of circumstances. There is no limit to the number of questions and answers that might be aroused by applying the variable to different settings. There is no end to the investigation of utility, although not all situations merit equal attention.

Every conceivable application gives an answer to utility. Unless one application becomes and remains the main consideration, there is no way that utility can be interpreted as an essence of validity. Consequently, utility is not inherent to the integrity of the variable.

In some applications, the variable may have only one use. In such a narrowly defined circumstance, we might consider utility as relevant to the validity of the variable. The difficulty with this approach, however, is that inevitably the criterion changes.

THE STANDARD ERROR

The main consideration facing us in any measurement application is to determine what standard error must be achieved in a designated application. The question becomes one of measurement precision, a question about reliability. We want to know how many items we need in a given region of the variable to obtain measurement precise enough to meet a current need.

THE RELIABILITY COEFFICIENT

The reliability coefficient is commonly used to indicate how well the variable correlates with itself over independent applications as it were. The traditional reliability coefficient, however, is a confusing mixture of qualitative validity and measurement precision. It remains confusing until it is decomposed into its constituent elements.

The traditional reliability coefficient is a mixture of item fit and prediction because it is composed of point-biserial correlations between the item responses and total score. Unfortunately, it mixes these two issues in such a way that one cannot explicate the component parts. Furthermore, the two issues are quite different from one another.

Precision of measurement depends on nothing more than the number of items administered, the extent to which the items are on target and the degree to which the respondent uses the items coherently. It is a straightforward calculation that can be done for any example.

BEST TEST DESIGN

Best test design depends on relating the characteristics of test design T (H, W, L - height, width and length) to the characteristics of the target G (M, S, D - location, dispersion and distribution) so that the SEM is minimized in the region of the variable where the measurements are expected to take place.

A test design can be defined completely by three test characteristics; height, width and length:

H = the height of the test on the variable, the average difficulty of its selected items,
W = the width of the test in item difficulties, the range of its item difficulties, and
L = the length of the test in number of items.

A target specification states where on the variable we suppose the target to be:

M = our best determination as to target location,
S = our best determination as to target dispersion, and
D = our best determination as to target distribution.

A best test is one which measures best in the region where measurements are expected to occur. Measuring best means measuring most precisely. A best test design T (H ,W, L) is one with the smallest error of measurement SEM over the target G (M, S, D) for a given test length.

If we know what precision in measures we want and have enough items near where we think our targets will be found then the standard error of measurement in logits is almost always well approximated by 2.5 over the square root of the number of items administered. (For further details see Wright and Stone, 1979, pp. 129-140.)

There is nothing more to investigate and nothing else to compute. It is a straightforward calculation of measurement precision for the proposed test and application.

THE TEST PLAN

Qualitative considerations need to precede the collection of data because they direct the construction and development of items from the plan for the collection of data.

In item writing we can specify preliminary anchor item values determined by our theory underlying their construction. Then we can plot the resulting empirical values against the intention of the items to evaluate whether resulting values relate to the previous specification according to theory. This approach connects intention and realization.

Realization is never exact. But it can meet the guidelines and be supported by the collected values showing how on target the numerical values of the items are relative to the intended plan. We expect the plan to be substantiated to some degree, we also expect further refinement.

MISFIT ANALYSIS

When our intentions are supported by data we get construct encouragement. If there are some discrepancies they teach us something about the relationship. If there are misfits of certain items, these teach us something about what it is that we are working towards.

The item fit statistics are most important. But, unfortunately, they are overwhelmed by the reliability coefficient in the traditional approach.

In Rasch measurement, we isolate the misfit statistic so that we can see which items demonstrate quantitative validity and also where they appear in the hierarchy of qualitative validity. Fit statistics are diagnostic of validity. They guide the measurement process by detecting *lack of fit* and *too good fit*. The former identifies discrepancies between our intention and the results. The latter identifies circumstances too good to be true and hence, suspicious. Both need further investigation. (See Chapter 17 (p.143), Information and Misfit Analysis.)

The confrontation of qualitative and quantitative validity provides opportunities in data analysis as we learn more about our data and as we resolve the discrepancies appearing in the fit analysis. Then we take steps to see whether we can make them more in agreement.

CONCOMITANT AND PREDICTIVE VALIDITY

A second kind of validity, concomitant and predictive, is how the task arranges the people.

The motivation now is to measure differences between people. We began by addressing the items, but the ultimate purpose is measuring the people. Now our concern is how the people are spread out along the variable.

This is the relationship between the standard error of the test and the standard deviation of the people we are measuring. This can never be addressed in general. It depends upon what problem is posed, whether our attention is over a wide or narrow range of persons, over all people, or only 4 year old children.

The answer depends upon the application of the characteristics of the test. Whether the test is good enough for the circumstances. In general, we need enough precision for each level of growth that is being studied.

A good measurement strategy is to first use a pilot location test, followed by a specific test to target the discovered location with greater precision than can be achieved solely by the pilot test. The pilot test determines the general location on the variable and the second test more precisely targets that location. The combination of pilot and target tests is less wasteful of time, achieves the desired measurement precision, and generally uses fewer items.

However, what we are addressing now is the test's utility, not its validity. Questions about how people are spread out over the variable are secondary to questions about the integrity of the instrument, the idea of the variability and its development.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware