20. VALIDITY

According to the Standards for Educational and Psychological Testing (1985), validity is "the most important consideration in test evaluation" (p. 9). Validity deals with the meaning of inferences drawn from test scores. The Standards emphasize that it is the inferences that are validated and not the test. The idea is that no test is valid or invalid in itself. Only its use in some application merits a designation of validity.

This primer discusses how validity is addressed in Rasch measurement. We explain how the types of validity discussed in the Standards are handled in a Rasch analysis of item response data.

TRADITIONAL VALIDITY

The three types of validity discussed in the Standards are (1) content related, (2) criterion related and (3) construct related. Validity itself, however, is held to be unitary. The Standards advance these types as related facets of a single problem. The types must be combined to validate the information obtained from the application of a test.

It is easy to become confused as to what is meant by validity because the three types are different in meaning and method. While the virtue of a single term "validity" is agreed upon by everyone, how to connect that term to the analysis of data is not. There are substantial and puzzling questions as to what is referred to, how it can be implemented in practice and what the results of implementation mean.

Since it is not clear what additional data are required to determine validity, that is what criteria are relevant, it is easy to become confused about what should be done. There is no unique or objective way to determine what the right criterion would be. There are always many possibilities. How do we establish which criteria are necessary, which are optional, which are decisive, which are only advisory? When criteria differ, how do we decide which one to use. Attempts to base validity on external criteria have raised more problems than they have solved. Many articles lament this dilemma (Bechtoldt, 1959; Beck, 1950; Campbell & Fiske, 1959; Cronbach & Meehl, 1955).

REAPPRAISING VALIDITY

The only way to escape this fruitless muddle is to focus on the data that are available, namely the actual responses of individuals to items and then to ask: What is there in these data that could answer validity questions?

What can we get from analyzing the data we have that could tell us about validity? When we look at responses to test items, two, and only two, types of data relevant to validity emerge. The first type concerns the ordering and spacing of items and persons which are produced by the analysis of item responses. The actualization of this kind of validity depends on prior knowledge of item content and person characteristics and, most of all, on clear intentions concerning what variable is to be defined and measured.

ORDER VALIDITY

The relation between item content and the empirical difficulty order of the items produced by the way persons respond to them either verifies, improves on or contradicts the intended definition and hence meaningfulness of the variable which the items are intended to implement. We expect one digit integer addition to be easier (have less difficulty) than division involving decimals and both to be easier than any problem involving a quadratic equation. It is almost impossible to write mathematics items without knowing in advance their intended and expected difficulty ordering.

For a spelling variable, we anticipate sequences of increasing difficulty like "cat", "wagon", "friendship", "meretricious" as defining a spelling variable that could be extended by adding easier and harder words as well as enriched by adding words at intermediate levels of difficulty between these four.

The way to begin this kind of thinking out of a variable is to write or select an initial item and then to write or find another item which we expect, according to our theory, to be easier or harder - continuing in this way, item after item, extending and filling in, step-by-step, until a detailed definition of the intended variable is laid out.

This simple beginning can lay out an orderly and meaningful item definition of an entire variable. We need only to apply a theory of what we are trying to do and to know our measurement intentions in order to think out an expected difficulty order for the items we plan to use. Then, when we use our items with persons, we can compare this intended and expected conceptual order with the empirical order actually provided by the data to see how well our expectations are confirmed. *Item order validity* operationalizes two of the Standard's three types of validity: content validity and construct validity.

Should we discover, however, that we are unable to imagine any canonical order for our items, then we are forced to admit that we do not understand the variable we are trying to define or how our items are supposed to implement its definition. We are forced to realize that we still have more work to do on our variable, by thinking it through more carefully before our purpose will become clear enough to us for useful action. Even in the earliest stages of variable construction we must have some idea of how to write items in an orderly fashion or else our measurement project cannot thrive. We must know ahead of time the difference between an easy and a hard item. We must know our purpose.

The difficulty order of items defines the variable's meaning and hence its content and construct validity. The ability order of persons that is produced by their performance on a test specifies the consequences of measuring on the variable and so determines the variable's utility. Relevant concomitant person orders such as those produced by age, school grade, civil service rating, or any other characteristics which ought to correlate with our intended measure, can help us to learn about its utility and so might be referred to as background criteria for our variable. But the variety of possibilities guarantees that no single criterion can be decisive. Nevertheless, to the extent that there is a part to be played by "criterion validity" in the evaluation of the utility of our variable, it is to be found in *person order validity* - the way persons are ordered by their measures.

CRITERION VALIDITY IN VARIABLE MAPS

The criterion validity of the Standards presupposes the existence of an external criterion sufficiently well established to serve as the base against which the test can be compared. The correlation coefficient is usually used as the index by which this comparison is evaluated. Two strategies are usually employed.

1. A test is designed to predict some already known criterion and the correlation with this criterion is taken to indicate the degree of criterion validity.

2. One test form is correlated with another test form to indicate their degree of consistency with one another.

The apparent simplicity of these approaches is flawed by the problem of the criterion. Is the criterion valid? Can it serve as a stable base? Does the correlation between two test forms address any substantive question about validity?

Criterion validity is better addressed by building an item map of the variable and then augmenting this map with the values of whatever concomitant criteria can be gathered along with the test data. All criteria can be located on this variable map together with the item calibrations and person measures.

When collecting test data we can record the associated person information of gender, age, school and scholastic level. The levels of these criteria can be plotted along with item calibrations and person measures on the variable map to show how these criteria relate to persons measures and also to item content.

We can formulate hypotheses about any criteria that we imagine might be relevant to these item calibrations and person measures and determine from the relative locations of these criteria on the map exactly how they are, or are not, related to the item calibrations and person measures.

The variable map is the best way to assemble and picture relevant criteria together with item calibrations and person measures. The map gives us a definitive and detailed picture whereas correlations only indicate the presence of some general relationships.

FIT VALIDITY

The second type of validity has to do with response pattern consistency for items and also for persons. This kind of validity comes from the fit of the observed person-item responses to a useful definition of measurement and hence to the estimated values of item calibrations and person measures. Although the necessity of this kind of response performance validity for persons and items was explained and satisfied by L. L. Thurstone in the 1920's (i.e., Thurstone & Chave, 1929), it is not mentioned in the Standards.

Item and person fit statistics are always necessary. The absence of fit statistics implies the absence of a model for what we expect - a lack of awareness of what we are trying to do. If we do not know what to expect, we cannot hope to explain what happened or know how to use the results.

Point-biserial coefficients (conventional item discrimination) have been used as item fit statistics for decades, though few practitioners have much idea as to what the statistical model for point biserials might be or what that signifies for the interpretation of their data. No one knows what size coefficients to seek or to act on.

When we take as the working requirement, however, that item responses shall be summarized by right answer counts i.e. raw scores, then we can deduce from that ubiquitous practice the necessary and sufficient measuring model. That necessary model is the Rasch model in which person raw scores or percent corrects can be used as the sufficient statistics for estimating person abilities and item p-values can be used as the sufficient statistics for estimating item difficulties (Andersen, 1977).

The mathematical form of the measurement model is deduced from the canonical requirements for measurement (Wright & Stone, *Deducing the Measurement Model*, Chapter 4). The measurement model specifies how to apply these measurement requirements to the data. It specifies what kind of relationship must be approximated between the observed data and the estimated measures in order for valid calibrations and measures to result.

ITEM FIT

When a Rasch analysis is made of item response data, it follows naturally to analyze the extent to which each person's response to each item fits the Rasch model expectation. An item fit statistic is calculated for each item. This summarizes the extent to which the sample's pattern of response to that item is consistent with the way these people have responded to the other items. This gives us "consistency" fit statistics for each item and for each person and also for any subsets of items and persons which might interest us (see Identifying Item Bias, Chapter 8).

The conventional approach to item fit has been the point-biserial correlation coefficient. Item misfit is thought to be indicated by a low point-biserial. It is equally true, however, that a high point-biserial coefficient can also indicate item misfit. This dilemma was identified by Loevinger (1947) as the attenuation paradox (Tucker, 1953; Andrich, 1982).

PERSON FIT

Although hardly anyone computes a person point biserial, the motivation to do so is even greater than the motivation to compute item point biserials. Of course, the attenuation paradox applies to person, as well as, item data. In a Rasch analysis a person fit statistic is calculated for each person. This fit statistic summarizes the extent to which that person's pattern of performance on the test is consistent (or inconsistent) with the way these test items are usually used by people responding to them.

When a person does some lucky guessing and so manifests some unexpected right answers on items that ought to be too hard for that person, we may doubt the validity of their performance and hence question the meaning of their score and measure. How much of the score tells us what they know and how much tells us that they are lucky guessers? When we examine the particular items on which they have failed, we may conclude that their score contains some lucky guesses and is thus misleading and hence somewhat invalid. At this point, however, our attempts to measure these

"lucky guessers" need not cease. Our measurement model enables us to know what we are doing. We can use its fit statistics to identify the lucky guesses and its item-free estimation procedure to re-measure the "lucky guessers" on the basis of their answers to the items on which their right answers were not lucky guesses.

SUMMARY

Rasch measurement helps us to see that there are two, and only two, types of validity that can be evaluated from item response data: (1) the ordering of items and persons and (2) the fit of items and persons.

Order Validity

1.1 "Meaning" validity from the calibration order of items. This implements the *content* and *construct* validities of the Standards.

1.2 "Utility" validity from the measurement order of person characteristics. This implements the *criterion* validity of the Standards.

Fit Validity

2.1 "Response" validity determined from the discrepancy between a particular response and its expectation. This identifies individual observations the values of which contradict their use in the estimation of useful measures or calibrations.

2.2 "Item Function" validity determined by an analysis of the validities of the sample of responses to that item, i.e. item fit. This identifies for review and revision items which may not be working the way we intend them to.

2.3 "Person Performance" validity determined by an analysis of the validities of the responses of that person, i.e., person fit. This identifies for review and diagnosis person's who may not have taken this test in the way we expected them to.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware