12. BUILDING SCHOLASTIC VARIABLES

Any professional in a position of responsibility in a school must have a way to keep track of what the school produces in student achievement. The only way to account for what is produced is to have a means for measuring scholastic growth in the areas for which the school is responsible such as arithmetic and reading. We will call these growth areas, variables and speak of the school's job as the increase of students' standings on scholastic variables.

Although there is a great deal of information about how children are supposed to develop, and what kinds of stimulation is supposed to encourage them, unless school effects can be made explicitly quantitative, it is impossible to evaluate school success. A school has to account for its educational efforts. To do this, the school has to construct scholastic variables on which the results of teaching can be measured and devise ways to measure these results.

THE MEASUREMENT PROBLEM

No school can neglect the measurement problem. Schools have to deal with it because it is the only way they can report to themselves or to the people to whom they are responsible the extent to which they are accomplishing their reason for existing. Schools must be able to measure their students' achievement.

How can school variables be defined and measured? We are deluged with tests from competing publishers who claim that their products relate the scores of increasingly difficult tests and provide indications of growth in particular areas. We believe that scholastic growth can be manifest by performance on test items. There seems no doubt that useful and relevant information can be obtained by giving students carefully selected questions to answer and then observing how they answer them. Thus we expect to use test performance to infer students' standing on the scholastic variable provoked by the test questions.

The trouble is that test publishers offer contradictory systems for quantifying test performance. The translation from one system to the other is neither definite nor agreed upon. Connecting test publishers' measures over the years of development is also difficult. Their equating systems are not convincing. Their reporting units of percentiles and grade equivalents are misleading.

Disagreement among test publishers is not the only problem. Difficulty in equating forms over the years of development is another. Local school dissatisfaction with national test items is yet another. The definition of an educational variable provided by a publisher, although marginally acceptable in New Jersey, may not be relevant to a school in Oregon. But no school dares to go off on its own without maintaining some connection to other schools. Neither does any school want to capitulate to a "national" standard imposed by some publisher. National tests offer a kind of comparability but lack relevance and flexibility. Local tests offer relevance and flexibility, but lack comparability. What is needed is a measurement system based on students' responses to test questions as the essential observation with tests made up of items focused on common scholastic variables of interest to the school but with results

that can be compared from school to school. The ingredients of these tests must come from the school using them as well as from other reputable sources.

This flexibility, however, requires an objective method of constructing scholastic variables and checking consistencies that is accessible and workable for any school. We cannot know ahead of time whether there will be agreement among local definitions of the scholastic variables. Whether or not local schools are working with the same scholastic variables as state or national agencies is something that can only be established empirically. There also has to be a routine and objective way to find out how each test variable is working from moment-to-moment and place-to-place. Since the only way disagreements among differing agencies can be resolved is by an empirical check, the system of checking must be acceptable to all parties, even though they may disagree on the content of some items.

The way in which the relevance of items for a test is determined must be equally agreeable to national and local groups. It must have a methodological basis which transcends arguments about content. It must result in an objective measure which is immune to political manipulation.

Measurements can only be made through some kind of test situation. Tests can be valid if they are properly constructed. To be generally accepted, the test ingredients must represent both local and national wisdom and intention. The validity of items must be verifiable in some way equally satisfactory to all. Also, any measure, being an estimate rather than the thing itself, must be qualified by a standard error, the relevant index of its reliability as an agent of measurement.

To accomplish this it is necessary to develop banks of calibrated and validated items. These banks must consist of items which can be connected together in such a way that any selection from them can be used as a reasonable test for the common scholastic variable they, and all of the other items in the bank, define.

ITEM BANKS

This leads us to the concept of the item bank, with items contributed by local as well as national sources. National items would be items developed by expert teams. (See Choppin, 1968; Wright, 1977; Wright and Stone, 1979 and Wright and Bell, 1984 for an introduction to item banking.) Local items would be those items developed by school systems, by schools and even by an inspired teacher of the fourth grade who has insight into the scholastic development of the children in her class.

There must be room for all of these ingredients in the item bank. But having allowed this flexibility, there must be a method for checking whether each item is valid. It must also be possible for items that are valid to make up a test suitable to the occasion. Such a test must be equatable to any other test that might be constructed.

When a bank is well made and covers a wide range of the variable, then it is possible to have comparable measures available for individual children with whatever set of items they take and hence to follow student scholastic development longitudinally from the early grades. This requires an easy test that a second grader can take and another hard test measuring on the same scholastic variable but so much further along the variable that the same student can take it 10 years later and yet obtain a measure on the same scale and hence quantitatively comparable to the earlier measure. Items from these two tests could hardly be taken by both second and twelfth graders. Nevertheless, since we intend to compare

the measures implied by each of these tests and to be able to say in an objective way how much a student has grown on the scholastic variable in those 10 years, we must find a way to connect these items so widely separated in difficulty to the scale of a single common scholastic variable.

A school system cannot escape the responsibility of measurement. But measurement needs to have certain characteristics in order to be useful to the school system. An item bank, solves a number of crucial problems. The developmental range problem and the equated forms problem is solved, and when the bank consists of local as well as national items, the relevance problem is also solved.

ITEM ANALYSIS

The occasion on which a student responds to an item, which we are relying on to show us where the student stands scholastically, is fraught with a variety of potential influences. But when we actually ask a student to answer a specific question, we would like to arrange things so that almost all that occurs at the moment is just an expression of that student's particular latent ability on the variable probed by that item. We are trying to provoke in the student's response a clear instance of this latent ability by means of the latent difficulty of the item that has been chosen. How well a student does on items of known difficulty can then be used to infer the student's measure on the latent variable.

However, when a student answers an item, there are the inevitable influences of motivation and distraction, as well as incidental elements in the item itself, which impede and facilitate the student's ability to solve it. Suppose it is a mathematics word problem. If the student is a good reader, it may be easier to do this item than if the student is a poor reader. It would be unfortunate if we failed to learn about a student's mathematical competence because reading difficulties on math items obscured the evidence the student would otherwise provide about math competency.

There are also administration and targeting difficulties which affect how students respond to items: guessing (on items too hard for them), sleeping (on items too easy for them), fumbling (on how the form is to be filled in), plodding (too slowly for the testing time and so not finishing) and bias (for and against success), all of which can interfere with measurement.

The system used for measurement must have a way to protect itself and its users against being mislead by unexpected disturbances in the observations from which the measure is estimated. The system must be able to detect spoiled measures. Once a test has been administered, we must be able to detect improbable divergence from expectation, to catch and correct for the influences of guessing, sleeping, fumbling, plodding and bias. We must be able to identify any secondary factors which interfere with performance on each item.

The measure estimated from a score on a test is an inexact estimate. We need to know not only the validity of the item responses on which the measure is based but the reliability, the error, of the measure.

A measurement project has two parts, item banking and person measuring. What is needed to manage these two parts is a common system which underlies both of them and so connects them together. The only hope we have of succeeding with a measurement project is to deduce a model for what we want to happen when a person encounters an item, a model formulated in the simplest practical terms, which also implements the basic requirements of measurement.

If we do not have a model, we cannot tell how to connect items together in the bank or how to free individual measurements from the particular items which happen to be used on a test. If we do not know what to expect, we have no way to tell whether a response is unexpected. We must be able to calculate from a model what we expect the answer to be so that we can observe whether a particular answer is surprising. The detection of irregularities requires a frame of reference by which a surprise can be defined.

This leads to the realization that, as far as measurement is concerned, it is not only sufficient but also necessary to pursue and enforce the fiction that each item can be characterized by a difficulty and nothing else and each person can be characterized by an ability and nothing else. We know that other factors always play a part, but with a simple model as our guide, we can always tell whether or not those other factors have spoiled the use of our simple model as a means for calibrating items and measuring persons.

When a simple model is put forward, that is not to say that what it is applied to is thought to be simple. Rather it is to assert that only through the construction of successful approximations to a simple model have we any chance of proceeding coherently and of making progress in managing a measurement project.

It is also not to say that when a student takes an item nothing is observed but the student's ability and the item's difficulty. Instead, it is our plan to make an effort to arrange and maintain things so that when a student takes an item most of what is observed is the expression of the student's ability against the difficulty of the item so that the observed response is dominated by student ability and item difficulty. Then, if something else happens, we can use the frame of reference of our simple model to identify the disturbance and to make correction for it.

THE MEASUREMENT MODEL

The traditional true score model specifies the observed score of a person taking a test as the sum of a true score and an error term:

$$x = t + e$$

where

x = OBSERVED SCORE
t = TRUE SCORE
e = ERROR

But we know that raw scores cannot be linear in what they represent and there is no useful theory for how big the true score error term should be. What we need, instead, is a different model which not only specifies that the person has an ability which is expressed in his behavior, but also that each item has a particular difficulty which is also expressed in any responses to that item, including the given response. Finally, we want a model which specifies how much deviation from expectation is reasonable and how much is excessive.

THE RASCH MODEL

The Rasch model (Rasch, 1960/1980) is a binomial probability model for a dichotomous right/wrong response. The Rasch model specifies that the probability of a right answer is defined by

the difference between person ability and item difficulty. Then, when the probability of a right answer is calculated to be near zero, but a right answer is nevertheless observed, that right answer is obviously surprising. Being able to estimate the probability of a right answer enables us to be precise about the extent of our surprise.

The discrepancy between observation and expectation can be put into a standard form so that we can have a standard reference distribution for it. This quantifies the extent of our surprise. We will be surprised when a person of low ability achieves something that requires exceptional ability. When a person attempts an item many units harder than he is able and nevertheless gets it right, that right answer might have a probability of occurring less than five in 100 times. In that case, we might take the position that our surprise has become too large for comfort. Thus we have a means for being explicit about the extent of our surprise and, if we can agree among ourselves as to what level of improbability is unacceptable, then we have an explicit and public rule which we can apply to validate any observed response.

This enables us to take an objective stand with respect to what to do about correct answers to items too many units above a person's ability. Using the natural log odds units (logits) of Rasch measurement, a difference of three logits would produce an improbability of .05. In particular, we may decide to use such improbable answers only for diagnostic purposes and to exclude them from our measure of the person.

When an unexpected response occurs, we do not ignore it, what we do is to decide what to do with it. We might decide to use it in the score, or to delete it. We might decide to use it to diagnose the person or to diagnose the item. Both can be useful. When only one person uses one item unexpectedly, that, in itself, will not tell us whether the person or the item produced the unexpected condition. If we suspect it was the item, we will look at the responses of other people to see whether that item continues to behave poorly, e.g. for many boys, or for many fourth grade boys, or for whatever condition we suspect might make the item irregular. If, on the other hand, we are making an individual study of a child and are concerned about brain damage, emotional disturbance, a fixation, or an inhibition, then we could become especially interested in the diagnostic potential of unexpected responses, and might even seek to provoke such responses for diagnostic reasons.

A careful study of items is beneficial to any school system. It can produce uniform content-free public decision rules that can be applied fairly and without prejudice.

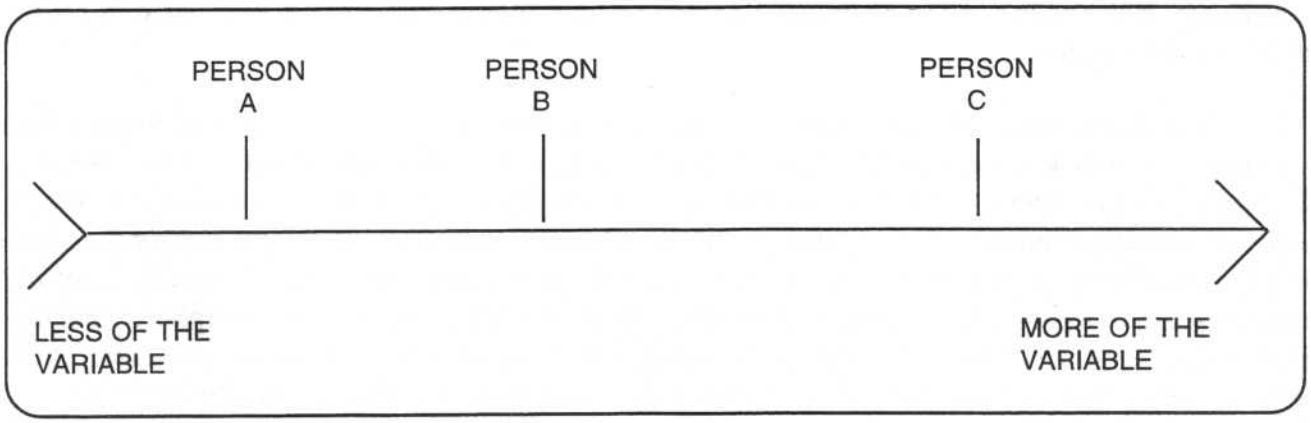
MEASUREMENT CRITERIA

LINEARITY

When we think about a variable, we have in mind the straight line so well represented by the familiar yardstick. One direction of this line represents more of the variable; the other, less. Person measures are locations along the interval scale of this line. This simple idea is illustrated in Figure 1.

That we employ the idea of a straight line when we think about variables like height and weight is obvious. But the relevance of this idea may not be as obvious when we speak of constructs such as intelligence or attitude. Nevertheless, we betray our reliance on this simple and useful idea whenever

Figure 12.1
Positions of persons A, B, C on the line of a variable.



we say that one person has a more positive attitude than another, or whenever we report an intelligence score for a child.

Our inevitable reliance upon this simple idea was noted long ago by L. L. Thurstone:

> The very idea of measurement implies a linear continuum of some sort such as length, price, volume, weight, age. When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind. We judge in a similar way qualities such as mechanical skill, the excellence of handwriting, and the amount of a man's education, as though these traits were strung out along a single scale, although they are, of course, in reality scattered in many dimensions. As a matter of fact, we get along quite well with the concept of a linear scale in describing traits even so qualitative as education, social and economic status, or beauty. A scale or linear continuum is implied when we say that a man has more education than another, or that a woman is more beautiful than another, even though, if pressed, we admit that the pair involved in each of the comparisons have little in common. It is clear that the linear continuum which is implied in a "more and less" judgment is conceptual, that it does not necessarily have the physical existence of a yardstick (Thurstone, 1928a, p. 532).

INVARIANCE OR OBJECTIVITY

When we measure a variable such as verbal ability, the measures we obtain must not depend upon the particulars of the items administered. Our ability measures must be freed of the particulars of the items taken in the same way that measures of height have a meaning which is independent of the particular yardstick used to obtain them.

Thurstone saw the necessity of this in 1926, and described the following requirements of a satisfactory measuring method:

> It should be possible to omit several test questions at different levels of the scale without affecting the individual score. It should not be required to submit every subject to the whole range of the scale. The starting point and the terminal point, being selected by the examiner, should not directly affect the individual score (Thurstone, 1926, p. 446).

Thurstone also pointed out the accompanying necessity of being able to obtain difficulty estimates for items which are freed from the particulars of the calibrating sample:

> One of the first requirements of a solution is that the scale values of the statements of opinion must be as free as possible, and preferably entirely free, from the actual opinions of individuals or groups. If the scale value of one of the statements should be affected by the opinion of any individual person or group, then it would be impossible to compare the opinion distributions of two groups on the same base (Thurstone, 1928b, p. 416).

And in the same year:

> The scale must transcend the group measured. One crucial experimental test must be applied to our method of measuring attitudes before it can be accepted as valid.
>
> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measuring function is so affected, the validity of the instrument is impaired or limited. If a yardstick measured differently because of the fact that it was a rug, a picture, or a piece of paper that was being measured, then to that extent the trustworthiness of that yardstick as a measuring device would be impaired. Within the range of objects for which the measuring instrument is intended, its function must be independent of the object of measurement (Thurstone, 1928b, p. 547).

The criteria for "measurement" are: logical ordering, linear scales, and objective comparisons. A model is needed which enables observations to be transformed into measures which meet these requirements.

In the early 1950's Georg Rasch (1960/1980) undertook to obtain measures of reading ability which were independent of the difficulty of the test taken:

> In a concrete formulation of this problem I imagined - in good statistical tradition - the possibility that the reading ability of a student at each stage, and in each of the two above-mentioned dimensions, could be characterized in a quantitative scale, but by a positive real number defined as regularly as the measurement of a length (Rasch, 1977, p. 59).

Rasch coined the term "specific objectivity" to describe comparisons among persons which are independent of the item parameters, and comparisons among items which are independent of the person parameters.

THE ITEM BANKING MODEL

Item banking can be accomplished with Rasch's psychometric methods. His measurement model describes the probable outcome of any encounter between a person and an item as entirely determined by two parameters - the "ability" of the person, represented by b; and the difficulty of the item, represented by d. If we use the numeric labels $x = 1$ to represent a correct answer and $x = 0$ to represent an incorrect answer, then Rasch's model for the probability of response x is:

$$P\{x=0,1|b,d\}=\exp[x(b-d)]/[1+\exp(b-d)]$$

$$\text{or} \quad \log\left[\frac{P_{x=1}}{P_{x=0}}\right]=b-d$$

Rasch specifies that the log odds (logits) that a person with ability b answers correctly an item with difficulty d correctly be dominated by the difference (b-d) between person ability b and item difficulty d. This positions persons by their ability and items by their difficulty on the interval scale of a single variable which they share. The result is probabilities of potential interactions between persons and items which are positioned along one common line and specifications of expectations for all possible responses.

Because the parameters b and d in Rasch's model appear as separate terms in a linear function, they can be separated in the application of the model. The difficulty calibrations of the items can be estimated in a way which frees them from the ability distribution of the persons used and the ability measures of the persons can be estimated in a way which frees them from the difficulty distribution of the items they happen to take. This produces the "sample-free" item calibration and "test-free" person measurement (Wright, 1968) which Thurstone demanded.

The sufficient statistics for these results are the test score for each person and the number of persons who respond correctly to each item, the sample score for each item. But these scores are not yet calibrations or measures because they are nonlinear on the variable they are intended to measure and also sample and test dependent. The Rasch measurement procedure, however, can use these familiar raw scores to construct sample-free item calibrations and test-free person measures on a common linear scale.

Each item's raw score is specific to the ability distribution of the sample used on that item, but the linear Rasch item calibrations are adjusted so that the effects of this ability distribution are removed. The resulting sample-free item difficulties can be used to define a general variable of meaning which can reach beyond the particular occasion of calibration.

Each person's raw score is specific to the pattern of item difficulties in the particular test he or she takes, but the linear Rasch ability measures are adjusted so that the effects of this item difficulty distribution are removed and the person's ability is generalized onto the variable defined by the whole set of calibrated items.

Whether any particular set of calibrations and measures are in fact test-free and sample-free can be verified at each step by simple methods (Wright and Stone, 1979). Verification of fit to the Rasch measurement model provides an explicit quantitative definition of item function validity and person performance validity and enables continuous quality control over item calibration and person measurement.

With a workable calibration procedure and a method for the evaluation of fit, it becomes practical to turn our attention to a critical examination of the calibrated items to see what it is that they imply about the possibility of a variable of some useful generality. We can find out whether our calibrated items spread out in a way that shows coherent and meaningful direction. We can examine

the hierarchy of item content and evaluate the extent to which this order indicates a line of increasing competence of recognizable meaning.

DEFINING A VARIABLE

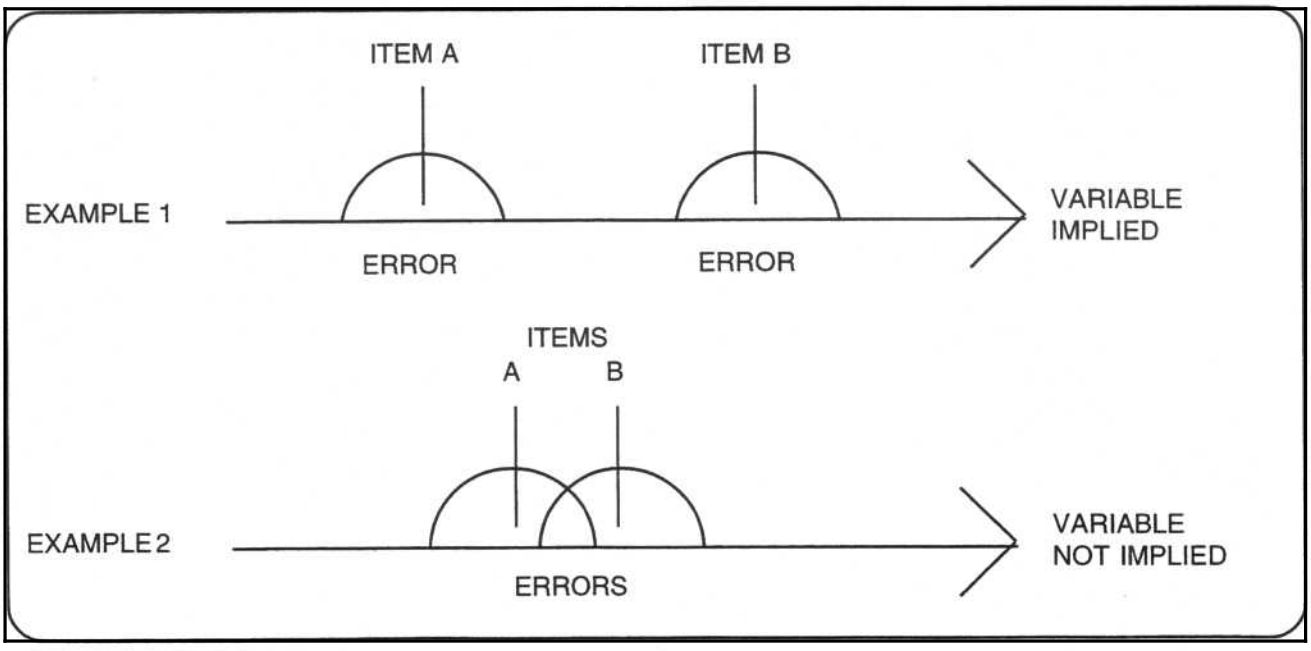
Our intention now is to show how calibrated items can be used to define a variable and how to find out whether the resulting operational definition of the variable makes content and construct sense. We begin by examining the degree to which the spread of item difficulties exceeds the standard error of their estimates, that is, the degree to which the data have given a direction to the variable. Consider, for example, the estimates of two item difficulties with their respective standard errors of estimation. In order for these two items to define a line between them, the difference between their estimates must be substantially greater than the standard error of this difference! Only when the two estimates are well separated by several calibration standard errors can we begin to see a line between the two items suggesting a direction for the variable defined by their content and order.

If, however, when we compare two item difficulty estimates, each bracketed by a standard error or two, they overlap substantially, then we cannot assume that the two values differ in difficulty, and as a result, cannot see a direction for the variable. Instead, the items define a point without direction. If the items do not spread out, then what have we defined? Only a point, perhaps on some variable, perhaps not. But the extent and hence the meaning of the variable is still missing.

Figure 12.2 illustrates this idea. In the first example we have items A and B separated from each other by several standard errors. Even with two items we see a direction to the variable as pointed out by these two items. In the second example, however, we find the two items so close to each other that, considering their standard errors, they are not separable. We have a point. But no direction has been established and so no quantifiable concept of the variable has as yet been implied. Only when items can be separated along the line representing the variable of interest have we begun to realize a construct.

Figure 12.2

Defining a variable.



In this discussion we have introduced a method by which objective scholastic variables can be constructed. Developing banks of Rasch calibrated items is the method. Item analysis is the tool by which these banks are built. The measurement model of George Rasch provides the means by which we construct these measurements. It provides a workable calibration procedure and a method for the evaluation of fit. Successful item bank construction can meet the criteria that Thurstone stated in defining the requirements of measurement - valid ability scales which transcend their particular items.

In the accompanying chapters we explore each of the above areas in detail.

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware