INVESTIGATING FIT WITH THE RASCH MODEL

Benjamin Wright and Ronald Mead (1979?)

Most disturbances in the measurement process can be considered a form of multidimensionality. The settings in which measurement are made always involve a multitude of factors. Successful measurement depends on achieving sufficient control with respect to the observations taken, so that their variations, dominated by the positions of persons (and items) differ along a single variable. Even though persons differ in many ways, their measurement becomes possible when one of these dimensions dominates the behavior provided by the items used for measuring. Analogously, even when items differ on a number of dimensions, this can be used for measuring if the responses of persons are dominated by only one of these dimensions. Thus measurement can succeed inspite of multidimensionality when the multidimensionality is not shared actively by both persons and items. We will illustrate this with some examples.

Case I: Two types of items

Suppose we wish to measure "general mental ability," and to do this, we construct an instrument containing both reading and math items. While this instrument might be considered two dimensional, measurement with it could succeed in situations where either

1) there is no variable which affects the probability of success on the reading items differently than on the math items,
2) or math ability and reading ability are so highly correlated in the population that they do not appear different.

In either case we should not care whether the measurement were made entirely with reading items, entirely with math items, or any mixture in between, since all items would measure the "same" variable. In the first case, there is only one variable. In the second, there are two but since they are highly correlated They act as one and we can measure math ability with reading items and reading ability with math items, if we choose. It does not matter whether we call the resulting measure math, reading or general ability. However, if we try to assert that both types of items are necessary for a "fair" measurement and become involved in setting the correct proportion of each, we have admitted the multidimensionality of the situation and should instead measure the two variables separately with items appropriate to each.

It is only possible to measure a person, who always has many different abilities on one variable by carefully constructing an instrument which addresses just that one variable. We may sometimes get by with a multidimensional instrument, since the two alternatives above--one variable versus two highly correlated variables--are not distinguishable in data, but, when we use an instrument of items readily classifiable into two or more types, its (effective) unidimensionality must be corroborated for each new sample.

Case II: One type of item with extraneous variables

A contrasting case can be illustrated by considering the measurement of, say, problem solving ability with an instrument composed of written word problems. Proficiency on this instrument requires many abilities in addition to problem solving, not the least of which is the ability to read the

language in which the problems are written. If the reading ability of every person is well above the "readability" of the problems, differences among items or persons in this respect will not affect performance on the instruement. However, if any person has difficulty reading the problems, his measure of problem solving ability will be biased downward by this extraneous factor. His probability of success will be influenced by the interaction between his reading ability and the readability of the problem. This can, of course, be eliminated by carefully regulating readability to be beneath the reeding ability of the target population.

This case differs from the preceding one in that each item has a "difficulty" on two variables. As long as all persons are sufficiently able readers, the instrument can be used to measure problem solving ability. Theoretically, at least, such an instrument could also be used to measure reading ability among very able problem solvers who were poor readers.

Random guessing on multiple-choice items is another example of extraneous variation. Persons succeed on difficult items more often than their abilities predict. This makes them appear more able when more difficult items are administered since their success rate does not decrease as difficulty increases. A similar but opposite effect occurs when able persons become careless with easy items, making them appear less able than they are.

Such items "measure" two variables--the ability of interest and the tendency to guess or to become careless. The "guessingness" of the item may or may not be a simple

function of the difficulty on the main variable but for the person two different variables are involved. The measure of either variable is threatened by the presence of the other.

These forms of multidimensionality have in common that different subsets of the items produce non-equivalent estimates of person ability and different subsamples of personsproduce different estimates of item difficulties. This contradicts the Rasch requirement that ability measures be independent of the items administered and item calibrations be independent of the persons used. In order to see how to avoid such disruptions we need to study possible sources of disturbance to develop analytic procedures that detect and assess the importance of disruptions when present.

Unequal Item Discriminations

No discussion of disturbances in Rasch measurement is complete without metion of item "discrimination." Rasch's derivation of what is required in order to achieve objectivity (i.e., measures of person ability that are freed from the particular sets of items administered), and calibrations of items that are freed from the particular samples of persons used, lead to amodel which rules against a parameter for item discrimination. If measurement objectivity is to be achieved the situation must be arranged so that a parameter for discrimination is not necessary.

When the problem is approached from other perspectives, for example, when the observations are considered so inviolate and valuable that the data allowed to determine the form of the model, regardless of the effect on the measurement process, item discrimination is almost always

included as a parameter. A model with an additional parameter, such as discrimination, will always recover the observed data more precisely than one without, but it is not at all clear when that is done that status the resulting "estimates" of discrimination can have in our thinking about the generalizable and reproducible attributes of the situation. It remains to be settled whether discimination "estimates" pertain to a stable, meaningful parameter that characterizes future outcomes of similar situations or whether they are only transiently useful as a descriptive statistic for diagnosing the trouble in one set of observations.

There are two distinctly different situations which would lead to unequal discriminations. First, the items may be influenced to different extents by factors other than the variables of interest. For example, with the problem solving test, if the items vary in readability and their readability is near enough to the reading level of the persons so that some persons have reading difficulty with some of the items, then the items will appear to vary in their power to discriminate along the problem solving scale. Although were the second variable centered, this apparent variation in discrimination would disappear. Items which no one is able to read will have no relationship to problem solving ability and items which everyone reads without difficulty will have the strongest relationship. Hence, the highest "discriminations" will be associated with items that are only influenced by the variable of interest

and the lowest will be for items most influenced by other factors.[1]

Alternatively, discriminations could vary if the items differ in the amount of random fluctuation associated with them. It should not he surprising that a completion item which requires the person to recall the correct response, discriminates more sharply than a multiple-choice item, which requires the person merely to recognize the correct answer. Recognition items give the person who does not recall, or even recognize, the correct response the opportunity to eliminate responses he knows to be incorrect, thereby increasing his probability of choosing the correct one. If his success at this is related to his position on the latent variable, not to his test-wiseness or any other extraneous factor, intelligent guessing of this sort need not interfere with the measurement process but does suggest the two types of items are of different quality.

More generally, items which require dramatically different behaviors from the person could be reasonably expected to vary in their discriminations. It would be rather surprising if, "application" items related to their variable with the same precision as "knowledge" items.

This suggests that it should be possible to distinguish items of unequal quality, if the inequality is due to the precision inherent in an item type, thus avoiding the necessity of parameterizing discrimination

---

[1] This assumes that items influenced by problem solving ability are numerous enought to dominate the operational definition of the variable measured by the instrument. Otherwise the variables would be a mixture of reading and problem solving and both types of items might discriminate poorly.

in the measurement model. If the inequality is due to the subtle presence of extraneous variables, the effect is not that of an item parameter but rather a transitory attribute of the local situation. This source of multidimensionality is a problem for any latent trait model and will always require a technique for discovering and controlling it.

The requirements of the Rasch model are little different than the recognized rules of good test construction. They are more explicit and more defensible because of their relationship to clear philosophy of measurement.

We will exploit the unique properties of the Rasch model to discover and evaluate disruptions in the measurement process. Fitting the model, and analyzing the observed residuals from it, in the difficulty metric, allows us to specify many useful questions in the familiar form of linear models. Since the Rasch model is the simplest possible latent trait model, this analysis gives the data the greatest possible opportunity to reveal important occasions of multidimensionality.

Motivation for the Use of Rasch Fit Analysis

Rasch (1960) and Wright (1967) have persuasively argued the advantages of specific objectivity in measurement over alternative approaches. With it, it is possible to estimate item parameters that are independent of the ability level of the calibrating sample and to obtain valid comparisons among persons regardless of the particular items administered. The mathematical form that the model necessarily takes to achieve this also eliminates the estimation and design problems that have traditionally troubled test users. (Andersen, 1972; Wright and Douglas, 1975)

The "cost" of these advantages is the exchange of assumptions about the distribution of person ability that is necessary for obtaining estimates of parameters with other latent trait models for the assumption of equal item discrimination implied by the Rasch model. Superficially this would seem a poor exchange; we are more apt to have a reasonable idea about the distribution than we are to have equal discriminations. However, the items unlike the Unknown distributions of person abilities, are under the control of the test constructor and with adequate resources, items can be developed which have relatively homogeneous discriminations.

There is also an important advantage in using the Rasch model as the basis of the analysis of fit that is distinct from its value as a measurement model. A more complex model will adapt itself more completely to specific sets of data and so obscure potentially interesting aspects of the situation which represent anomalies in the measurement process.

Including discrimination as a parameter will seem to "explain" many occurences of multidimensionality which in contrast will appear clearly as misfit with the Rasch model. Since when discrimination is used as a parameter there will appear to be no misfit present in these cases, we will be less likely to investigate further and will miss an opportunity to discover the reason for the interaction between persons and items. This too convenient mathematical "explanation" of lack of fit will also jeopardize the validity of future applications of the instrument.

The comparison with classical item analysis approaches is still more striking. The requirements necessary to use either the classical or Rasch approaches are essentially the same except that the Rasch model assumes the shape of the regression of the observed outcome on ability is logistic rather than linear. The advantages are the invariant Rasch paramaters and a rigorous mathematical model of the process which permits systematic evaluation of the data's fit to the model.

The requirements governing the use of the Rasch model are simply the logical extensions of traditional requirements but now facilitated by the explicitness of the model.

The traditional selection criteria are limited to attributing poor measurement to characteristics of items. Since it is difficult to address the issue of fit, item selection rules are limited to improving the total test reliability (Lord and Novick, 1968, Chap. 15). This is maximized if the "difficulty" (i.e., the proportion correct) for every item is 0.5 and the item discriminations (or biserial corelations) are as large as possible.

These criteria have several problems. An observed proportion correct of 0.50 refers to the applicability of the item to the mean of the distribution of persons. The item will still be "inappropriate" for persons in the tails of this distribution. In many applications, these people are our major concern.

Selecting items with high discriminations may also be misleading since exceptionally high discriminations are frequently due to the local and passing influence of an extraneous factor. Including such items

reduces the instruments reliability and validity in future applications. Finally, if we are totally successful in obtaining items of 0.5 difficulty and discriminations near one, the instrument of whatever number of items will end up functioning as a single item (Tucker, 1946) as far as measurement is concerned since each item will provide exactly the same information.

Fit analysis (and test design) based on the Rasch model incorporates and refines both of these rules. We can, if we have some notion of a person's ability, select items which minimize the standard error of measurement (i.e., maximize precision, reliability) for that person individually. This allows us to tailor the instrument to best achieve the objective of any measurement application. If we are chiefly interested in a particular person, we can design the test for him. If we are interested in a population we can design the test to best cover that range of ability or if we are interested in some external criterion, we can design the instrument to measure in that region. In no case are w& dependent on the proportion of some calibrating sample that succeeded on the item.

The tests of fit for the Rasch model will reject items with unusually low or unusually high discriminations. This screens out items that are weak or potentially misleading measures of the variables, as defined by the remaining items. It protects us from including items that are vulnerable to the influence of extraneous factors which happen to be related to the variable of interest in our sample. It selects from a set of items-that-we-think are relevant to our variable a homogeneous subset that can be reasonably accepted as measuring a single variable in the given sample in a generalizable way.