JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 14, NUMBER 3 FALL 1977

MISUNDERSTANDING THE RASCH MODEL

BENJAMIN D. WRIGHT
University of Chicago

Whitely and Dawis (1974) recently discussed, in this journal, the nature of objectivity with the Rasch model. They made several good points. However, their paper contained some unfortunate misunderstandings and errors.

One important misunderstanding concerns the estimation procedure necessary to calibrate items according to the Rasch model. The authors recommend a two-stage estimation procedure which is actually unnecessary and impractical. As a result they conclude that only huge sample sizes make it possible to apply the Rasch model to real data. Since the Rasch model can be and has been applied productively to sets of data as small as 100 persons, and under those circumstances lead to useful results, it is important to correct that misunderstanding and the incorrect conclusion drawn from it.

Whitely and Dawis recommend beginning parameter estimation with a least squares method based on item by score-group frequencies. This crude method was first used by Rasch in the 1950's and presented in his 1960 book. It is neither the most convenient nor the most efficient method of estimation and it is unnecessary as a beginning. Far preferable is to begin directly with the unconditional maximum likelihood procedure described by Wright and Panchapakesan (1969) and Wright and Douglas (1975b). Most researchers applying the Rasch model today are using some version of the computer program for unconditional maximum likelihood estimation written by Wright and Panchapakesan in 1965. The danger to them of being misled by the recommendation of Whitely and Dawis is slight.* However, newcomers to the topic who take the Whitely-Dawis recommendation seriously will be led astray.

Perhaps because they focused their attention on the least squares method of estimation, Whitely and Dawis conclude that huge sample sizes are necessary for useful item calibration using the Rasch model. This conclusion is based on their incorrect statement that each and every possible score group has to be inhabited by a substantial number of persons for the estimation procedure to proceed. This is untrue. In fact, a set of items may be calibrated even when all the persons in the calibration sample have earned one and the same score. Even the least squares method allows estimation of the relative difficulties of a set of items from merely the log-odds success that any single score group obtains on each item. No additional score groups are necessary for a set of unbiased least squares estimates. The same, of course, is true for the unconditional maximum likelihood procedure.

It is only when the researcher wishes, and wisely so, to test the fit of his data to the Rasch model, that more than one score group is needed. In order to evaluate fit, at least two score groups must be available so that the separate estimates of each item difficulty within each score group can be compared with one another to see if they are statistically equivalent across groups. If they are, then the results demonstrate con-

*Readers interested in applying the Rasch model and lacking a computer program can get an efficient program, at cost, from Benjamin Wright, University of Chicago, Department of Education, 5835 South Kimbark Avenue, Chicago, Illinois 60637.

sistent sample-free item calibration over the two score groups examined. If they are not, that is evidence of an instability in item difficulty (i.e., of an interaction between item and score level) which implies that the data collected are not suitable as they stand for this kind of item calibration.

Incidentally, when this kind of misfit is encountered, it is sometimes thought a good reason for concluding that the data collected may not be suitable for any kind of measurement. That is the conclusion which follows if the measurement sought is to be objective, in the sense ordinarily meant when scientists "measure."

When scientists measure they intend their measurements to be objective in the sense of being generalizable beyond the moment of measurement. This means that, whatever parameters are thought to characterize the measuring instruments, they must remain relatively stable through the range of intended application and must not interact substantially with the objects being measured. It also means that the parameters intended to describe the process of measurement can be estimated successfully.

In the various efforts to conceptualize what happens when a person takes a test item, it is sometimes popular to think of three item characteristics, namely, difficulty, discrimination and guessing. Difficulty, of course, is the parameter traditionally represented by the percent correct in the "standard" sample and is explicit in the Rasch model. But discrimination and guessing are quite another problem.

The estimation of discrimination, while frequently attempted, is clouded by uncertainties as to whether it can in fact be reliably estimated. The values actually obtained for a particular set of items are highly sensitive to the particular distribution of person abilities which happen to occur in the calibrating sample. In addition, when iterative solutions to the estimation problem are attempted, they tend to diverge at the extremes.

The estimation of guessing is even more obscure. Those who view it as the lower asymptote of the item characteristic curve believe very large and widely spaced samples are required to estimate it usefully. Others believe that guessing is done by persons, rather than by items. But they require long and widely spaced tests to get at guessing for each person.

In practice, it is not at all difficult to focus calibration data and item selection so that disturbances in measurement caused by possible variations in item discrimination or guessing become non-significant. The consequence of such explicit caretaking can be a pool of calibrated items selected because they are found to function usefully according to the Rasch model. Everyone who is content to use test scores which are unweighted by item discrimination and/or uncorrected for guessing, the practice followed by nearly all practitioners, is assuming that their items are in fact working in just the way modeled by Rasch, whether they realize and capitalize on that assumption or not.

Another point made by Whitely and Dawis (1974) in their section, "Item Calibration and Unweighted Score Groups" on page 168 is that, when the least squares method is used, the estimates are based on equally-weighted score groups rather than equally weighted persons. That is true of the least squares estimation procedure, but it is not true of the maximum likelihood estimation procedure. In the maximum likelihood procedure, each person is weighted equally, not each score group.

In their next section, "Anchoring and Interpreting Ability Scores," Whitely and Dawis say that "the key to the sample-invariant interpretability of ability scores, ... is the manner in which scores can be anchored." This is a trivial observation and, in fact, is *not* "the key" to sample-invariant interpretability. Sample-invariant interpretability depends on a demonstration that the difficulties of the items in question remain statistically equivalent over the various kinds of persons to be measured with these items. This is exactly the condition investigated when evaluating the data for fit to the model.

Anchoring the scales is a trivial matter of establishing a reference point. Inevitably it must be done arbitrarily. All the researcher is obliged to do is to make the arbitrary decision public.

In the second paragraph of the same section on page 169, Whitely and Dawis assert that anchoring can be "accomplished by setting the mean simple likelihood for the item set equal to one." This would be an unwise step to take, and would not be consistent with the multiplicative structure of the simple likelihood. Were it a question of anchoring simple likelihoods, the usual and most convenient step to take would be to set the continued product of the likelihoods of the standardizing set of items, not the mean, equal to one. Setting the continued product equal to one corresponds to setting the mean of the log likelihoods equal to zero which, in fact, is the arbitrary but convenient anchoring procedure routinely used in the estimation of Rasch item difficulties.

The Whitely and Dawis discussion of the interpretation of Rasch measures as basically domain referenced gets off to a good start. But when they assert that "the interpretability of Rasch ability parameters is, of course, slightly more involved than the direct interpretability of percentage correct scores" they neglect to mention that every Rasch ability measure can be compared directly with the Rasch difficulties of any criterion set of items. The probability of a person at a particular ability level succeeding on a particular set of items can be calculated directly from the model. This gives a result which is the expected percentage correct on those items for that ability; it is equivalent in interpretation to the "percentage correct score" and is backed up by a clear model of what is taking place.

This brings out a simple but important reason for using the Rasch model. The traditional index of item difficulty, percent correct in some "standard" sample, depends for its interpretation on knowledge of the ability distribution of that sample. If a smarter or dumber sample is used, then the traditional index of item difficulty changes. Rasch difficulties, on the other hand, depend on the sample-free calibrations of a pool of items which, together, provide the operational definition of a variable. These difficulties do not change with the calibrating sample, nor with the person measured.

In their discussion of equivalency and precision on page 170, Whitely and Dawis assert that "the equivalency of Rasch items subsets falls under the more limited definition of equivalent measures." Their definition of "less limited" is that parallel forms should have equal variances or reliabilities for any group tested. This, however, is not primarily a property of a test form, but more a property of the sample of persons to whom the test is given. Equal variances in samples of persons is irrelevant when developing generality in an item pool and a dangerous illusion when relied upon to provide equivalent forms. Perhaps the single most important consequence of the explicit measurement model proposed by Rasch is that it shows how to select and calibrate a pool of items. Every and any subsequent selection of items from this pool produces a test from which scores can easily be converted into measures on the latent variable common to the items in the pool. This is true generality in measurement, and provides the least limited form of test equivalence possible. In contrast, it is exactly the traditional method for equating forms which is severely limited by requirements which are neither relevant nor necessary.

Whether or not subsequent measurements made with various selections of items from the calibrated pool have comparable precision; i.e., similar standard errors of measurement, depends not on the calibration or construction of the item pool but on the appropriateness to the measurement target of the items selected. The best set of items to select for a target depends not only on where that target is thought to be centered, but also on how much uncertainty there is as to its probable location. The less certain we are, the wider test we need use to minimize the average expected standard error. Wright and Douglas (1975a) develop the specifics of best test design at length. In brief, the idea is to minimize the standard error of measurement by selecting items which are centered on the hypothesized location of the intended target and spread out to cover optimally the prior uncertainty about this location.

Far more influential on measurement precision than test center or width, however, is test length. The final step in best test design is to determine a useful balance between the number of items a person might reasonably be asked to attempt and the precision the examiner thinks he must obtain to make the measurement worthwhile.

On page 171 Whitely and Dawis (1974) provide a Formula 6 [probably borrowed from Formula 23, page 35 of Wright and Panchapakesan (1969)] which they refer to as a "standard error." It is actually a variance. The formula is incorrectly rendered. A similar mislabeling and error appears in their Formula 7.

In their discussion of Formula 7 Whitely and Dawis say "the inverse of the predicted cell frequencies are summed over items." This is not what is done. It is also not true that estimated abilities for scores belonging to large score groups have smaller standard errors than those for scores belonging to small score groups. In fact, the size of the calibrating sample score group does not appear in the formula for the standard error of measurement.

This error continues in the next paragraph on page 171 where Whitely and Dawis refer to "the average size of measurement error for a group." "Group" is an irrelevant reference. The measurement error pertains to a score only and not in any way to the number of persons who happened to obtain that score in some calibrating sample. In fact, one of the beauties of the Rasch model is that, once items are calibrated, an ability measure equivalent to any score can be estimated, regardless of whether or not any person so far encountered has ever actually earned that score. The standard error of this measurement estimate depends primarily on the length of the test, somewhat on the extent to which the implied ability is central among the item difficulties being used, and only in a very minor way on the precision of item calibration.

Finally, in discussing these incorrectly rendered formulas, Whitely and Dawis fail to bring out their major characteristic, namely that the standard error of item calibration is dominated by the reciprocal of the square root of sample size, while the standard error of person measurement is dominated by the reciprocal of the square root of test length.

The correct formulas for these standard errors are:

$$SE(b_r) = \sqrt{C_r/L} \qquad r = 1, L - 1 \tag{1}$$

and

$$SE(d_i) = \sqrt{C_i/N} \qquad i = 1, L \tag{2}$$

where $SE(b_r)$ is the estimated standard error of measurement of estimated ability b_r from score r,

$SE(d_i)$ is the estimated standard error of calibration for estimated difficulty d_i for item i,

L = test length,

N = size of calibration sample,

$$C_r = L\left[\sum_i^L P_{ri}(1 - P_{ri})\right]^{-1},$$

$$C_i = N\left[\sum_r^{L-1} P_{ri}(1 - P_{ri})n_r\right]^{-1},$$

n_r = the number of persons in the calibration sample with a score of r,

and $P_{ri} = \exp(b_r - d_i)/[1 + \exp(b_r - d_i)]$ is the Rasch model estimated probability of a right answer for a person with estimated ability b_r on an item with estimated difficulty d_i.

The coefficient C_r is the reciprocal of the average information concerning ability b_r provided by the L-item test on which the score r was observed. The coefficient C_i is the reciprocal of the average information concerning difficulty d_i provided by the N persons in the calibrating sample. Since these coefficients are the reciprocals of averages of $P(1 - P)$, they are limited to values between 4 and 9 for P's in the range $1/8$ to $7/8$. This means that when items and persons are appropriate to one another, so that the probability of success is between $1/8$ and $7/8$, the possible values of the C's can only vary between 4 and 9 while test length typically varies between 20 and 100 and calibration-sample size can vary between 100 and more than 500. Thus it is sample size or test length which dominates the possible values of the standard errors of calibration or measurement.

These formulas for the standard errors give us an opportunity to answer the question of how large a sample is needed to calibrate items satisfactorily with the Rasch model. Wright and Douglas (1975a, p. 34) have shown that the value of C is 6 or less for tests of typical width (5 logits or less) and relative scores in the range of 20 to 80 percent correct. If we use 6 as our working value for C, we can calculate the sample size necessary to obtain a particular level of calibration precision. Thus the sample size N necessary to obtain a desired standard error of calibration SE becomes;

$$N = 6/SE^2 \tag{3}$$

If in addition, we wish to evaluate item fit by splitting the calibrating sample into two approximately equal size groups, (perhaps a dumb group and a smart group, or boys and girls) and making a statistical comparison of the consequent pair of independent items calibrations, then the standard error of this difference SED becomes;

$$SED = \sqrt{2}\sqrt{6/(N/2)} = 2SE$$

and the necessary sample size for a given level of item calibration control; i.e., desired SED, becomes;

$$N = 24/SED^2 \tag{4}$$

The implications for item calibration sample size can be seen in Table 1.

To settle the question of calibration sample size completely we must trace the influence of calibration imprecision as manifested in SE on measurements subsequently

Table 1

Sample Sizes Needed for Various Calibration Precisions and Tests of Fit

Total Sample Size N	Calibration Precision SE	Two Group Test of Item Fit SED
100	.25	.5
150	.20	.4
250	.15	.3
600	.10	.2

made with the calibrated items. Wright and Douglas (1975a, pp. 35–39) have shown that when the disturbances in item calibrations tend to be random, when at least 30 items are used for measurement and when the test center is within 2 logits of the target center, then the standard deviation of the disturbance in item calibration SE can be as large as 0.5 without causing a bias in measurement larger than 0.1 logits of ability or 0.3 standard errors of measurement. If the control of calibration precision can be further narrowed to a standard deviation of 0.25, then the maximum bias in standard errors of measurement drops to 0.2. These findings, coupled with the information in Table 1, lead to the conclusion that calibration sample sizes of 500 are more than adequate in practice and that useful information can be obtained from samples as small as 100.

The conclusion of the Whitely and Dawis (1974) paper contains a mixture of true and untrue statements. It is true that unless the data are carefully constructed to yield objectivity they will not necessarily be found to have objectivity when the attempt is made to fit a Rasch model to them. The opportunity to determine whether or not there is a possibility of objective measurement in some data, by checking their fit to the Rasch model, represents the model's most important contribution to scientific method.

The Whitely-Dawis implication on page 176, that traditional equivalent forms are in some way better because they maximize both precision and statistical equivalency, is incorrect. The precision of a measurement depends on the relevance of the items to the target of measurement and on the number of items used. For dichotomous items, this situation is effectively the same, no matter what method or model for measurement is used. The more-essential question is, do the set of items all bear on a single common latent variable? If they do, then the Rasch model is the necessary and sufficient conceptualization. If they do not, then the set of items contain a mixture of variables and there is no simple, efficient or unique way to know their utility for measuring anything.

Whitely and Dawis' final statement reads "routine administration of tests by computer would be part of the necessary technological sophistication." This is incorrect. There are numerous paper and pencil group-administered tests now in use which have been composed out of items calibrated by the Rasch model and which can be used for estimating measurements by means of simple tables of score-measure equivalents, without any recourse to computers.

REFERENCES

RASCH, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Nielsen & Lydiche, 1960.

WHITELY, S. E., & DAWIS, R. V. The nature of objectivity with the Rasch model. *Journal of Educational Measurement,* 1974, **11,** 163–178.

WRIGHT, B. D., & DOUGLAS, G. A. *Best test design and self-tailored testing* (Research Memorandum No. 19). Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1975. (a)

WRIGHT, B. D., & DOUGLAS, G. A. *Better procedures for sample-free item analysis* (Research Memorandum No. 20). Chicago: Statistical Laboratory, Department of Education, University of Chicago, 1975. (b)

WRIGHT, B. D., & PANCHAPAKESAN, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 1969, **29,** 23–37.

AUTHOR

WRIGHT, BENJAMIN D. *Address:* The University of Chicago, Chicago, IL 60637. *Title:* Professor of Education and Behavioral Science. *Degrees:* B.S. Cornell University, Certificate Chicago Institute for Psychoanalysis, Ph.D. University of Chicago. *Specialization:* Measurement and Data Analysis; Psychoanalytic Psychology.