THE RASCH MODEL FOR ITEM ANALYSIS

Bruce Choppin

CSE Report No. 219
1983

Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles

The project presented or reported herein was performed pursuant to a grant from the National Institute of Education, Department of Education. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education, and no official endorsement by the National Institute of Education should be inferred.

TABLE OF CONTENTS

1. Definition of the Model

The so-called Rasch model now widely employed for item analysis, is only one of a complete family of models described by Rasch in his 1960 text. All may be properly called "Rasch Models" since they share a common feature which Rasch labeled "specific objectivity". This is a property of most measurement systems which requires that the comparison of any two objects that have been measured shall not depend upon which measuring instrument or instruments were used. It is a familiar feature of many everyday physical measurements (length, time, weight, etc.). In the context of mental testing, it means that the comparison of two individuals who have been tested should be independent of which items were included in the tests. Traditional test analysis based on "true scores" does not have this property since "scores" on one test cannot be directly compared to "scores" on another. (The peculiar virtues of specific objectivity and the conditions needed to achieve it are discussed later in this report.)

Mathematical Representation

The Rasch model is a mathematical formulation linking the probability of the outcome when a single person attempts a single item to the characteristics of the person and the item. It is thus one of the family of latent-trait models for the measurement of achievement, and is arguably the least complex member of this family. In its simplest form it can be written:

$$\text{Probability } [X_{vi} = 1] = \frac{A_v}{A_v + D_i}$$

where, X_{vi} takes the value 1 if person v responds correctly to item i, and zero otherwise,

A_v is a parameter describing the ability of person v,

and D_i is a parameter describing the difficulty of item i.

In this formulation, A and D may vary from 0 to . A transformation of these parameters is usually introduced to simplify much of the mathematical analysis. This defines new parameters for person ability ( ) and item difficulty ( ) to satisfy the equations:

$$A_v = W \quad \text{and } D_i = W \quad \text{for some constant } W.$$

Figure 1

$$A_v = W^{\alpha_v} \quad \text{and } D_i = W^{\delta_i} \quad \text{for some constant } W.$$

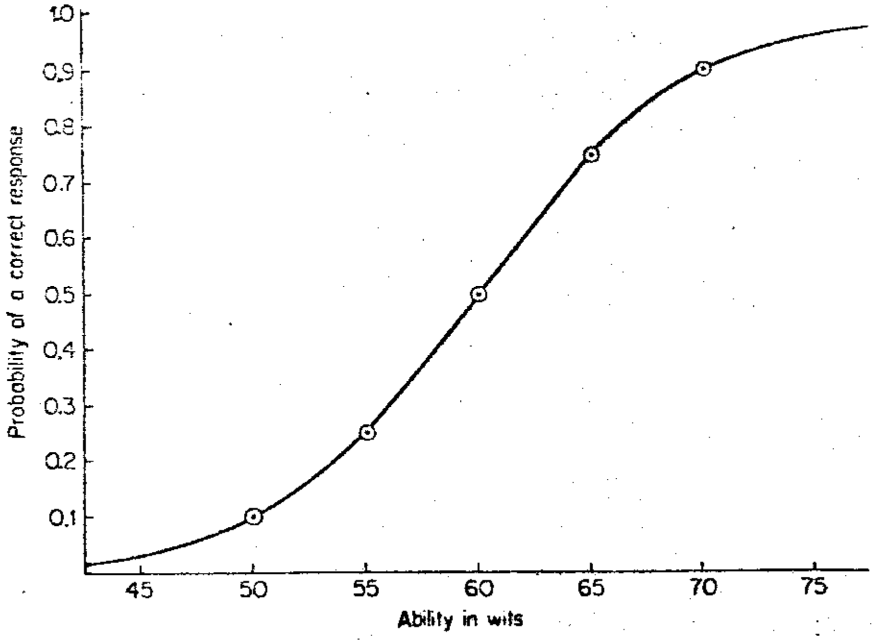


Figure 1 : Item Characteristic Curve (wits) for the Rasch Model

A further simplification, introduced by Rasch himself and used widely in the literature, is to fix the constant W to the natural logarithmic base, e. In this case the model can be written:

$$\text{(2)} \qquad \text{Probability } [X_{vi} = 1] = \frac{e^t}{1 + e^t}, \text{ where } t = (\alpha_v - \delta_i).$$

In this formulation, α and δ can take all real values and measure ability and difficulty respectively on the same "logit" scale. The sign of the expression (α - δ) in any particular instance indicates the probable outcome of the person-item interaction. If $\alpha > \delta$ then the most probable outcome is a correct response. If $\alpha < \delta$ then the most likely outcome is an incorrect response. It should also be noted that the "odds" for getting a correct response (defined as the ratio of the probability for getting one to the probability for not getting one) take on a particularly simple form:

$$\text{Odds } [X_{vi} = 1] = \frac{\dfrac{e^t}{1 + e^t}}{1 - \dfrac{e^t}{1 + e^t}} = e^t$$

$$\text{or } t = \log_e(\text{odds})$$

For this reason, the Rasch model is sometimes referred to as the "log-odds" model.

Alternative Units

As stated above, the model based on the exponential function yields measures of people and items on a natural scale, whose unit is called a "logit". Rasch himself used the model in this form, and most of Wright's publications also make use of it. Mathematically and computationally the logit is convenient, but as an operational unit it has two drawbacks. First, a change in achievement of one logit represents a considerable amount of learning. Studies in various parts of the world indicate that in a given subject area, the typical child's achievement level would rise by rather less than half a logit in a typical school year. In practice, many of the differences in achievement level that we need to measure are much less than this, as is the precision yielded by our tests, so results are commonly expressed as decimal fractions rather than as integers.

Secondly, logits are usually ranged around a mean of zero (this is a matter of convention rather than necessity) so that half of all the values obtained for parameters are typically negative. In general, teachers dislike dealing with negative numbers, and the prospect of having to explain to an anxious parent what Jimmy's change in math achievement from -1.83 logits to -1.15 logits actually means is too much for most of them.

The solution for practical applications of the Rasch scaling technique is to use a smaller and more convenient unit. This is accomplished by setting W to some value other than e. A number of alternatives have been suggested, but the unit in the widest use after the logit is obtained by setting $W = 3^{0.2}$. This unit is known as the "wit" in the United Kingdom and United States, and as the "bryte" in Australia. Wits are typically centered around 50 with a range from about 30 to 70. One logit is equal to 4.55 wits. For many purposes it is sufficient to report wits as integers. The particular value for W is chosen so as to provide a set of easily memorized probability values, as can be seen in the Table 1.

Table 1
The Relationship of Logits and Wits to the
Probability of Correct Response

(α - δ) Measured in Logits	(α - δ) Measured in Wits	Probability of a Correct Response
-2.198	-10	0.10
-1.099	-5	0.25
0	0	0.50
+1.099	+5	0.75
+2.198	+10	0.90

It must be emphasized that the choice of a unit for reporting is an arbitrary matter. Most of the theoretical work on the model, and all the computer programs for parameter estimation in common use, work in logits--translating to wits or some other scale for reporting only if desired.

Analytic Possiblities

Parameter estimation is a difficult issue in latent-trait theories. That for the Rasch model a variety of different estimation algorithms (at least six) have become available in the last fifteen years results from the mathematical simplicity of the Rasch formulation.

The basic equation models only the outcome of one particular item-person interaction, but since it does so in terms of a probability function, it is necessary to accumulate data from several such interactions in order to estimate parameters or test the fit of the model itself.

For example, the accumulation of responses of one individual to a set of items may be used to estimate the ability parameter for the individual, and the pattern of responses by several individuals to two items may be used to estimate the relative difficulty of the two items. From a (persons-by-items) response matrix it is possible to estimate both sets of parameters (abilities and difficulties), and also to check on whether the model is an acceptable generating function for the data. This calibration of items, and the test of goodness-of-fit to the model, correspond to item analysis procedures in classical test theory (but see section 5(a)later in this report).

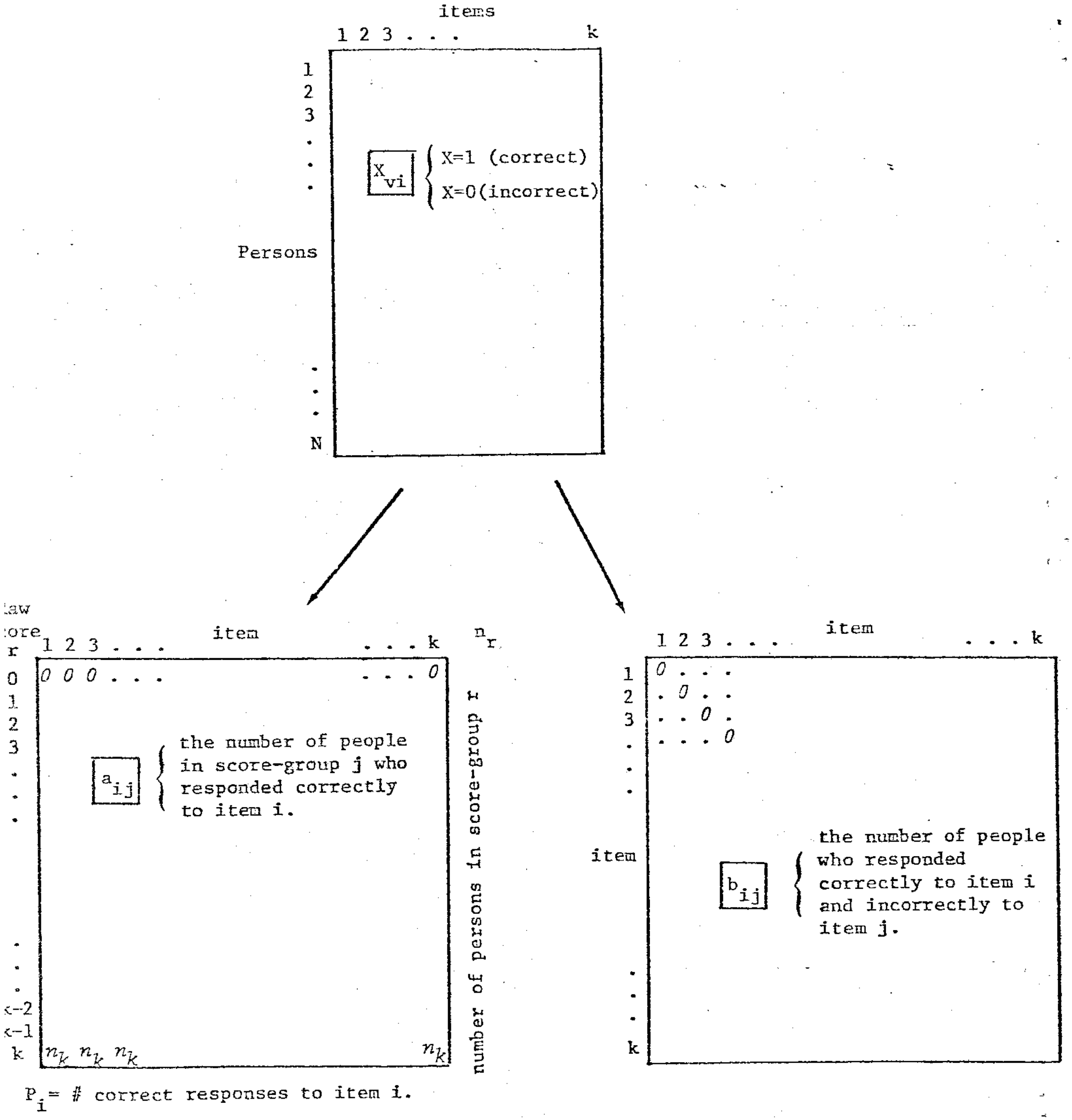
Once items have been calibrated, equations can be developed to predict the characteristics of tests composed of different samples of previously calibrated items, or the performance of previously measured people on new items. Although the simplest approach to statistical analysis requires a complete rectangular persons-by-items response matrix, other procedures are available to handle alternative data structures. For example, when a group of individuals take different but overlapping tests, the persons-by-items matrix will necessarily be incomplete, but it is still possible to calibrate the items and measure the people. An extreme example, in which a computer-managed adaptive test is individually tailored to each testee (such that the next item given depends on the responses to previous items), may lead to a situation in which every person tested may respond to a unique set of items. If the items have been calibrated in advance, it is possible to estimate the individual's ability parameter at each step of the sequence, and to discontinue testing when the ability has been measured with the desired degree of precision.

Estimation Techniques

Although this paper is not the place for a detailed presentation of the algebraic manipulation involved in the various algorithms for parameter estimation, an outline of the different approaches may be helpful.

Conventionally the starting point is taken to be a rectangular matrix of **persons** by **items** in which the elements are one if a particular person responded correctly to the appropriate item, zero if he responded incorrectly, and blank if the person was not presented with the item. Initially we shall restrict the discussion to complete matrices of ones and zeros such as occur when a group of N people all attempt a test of **k** items. In most applications N is usually much larger than **k** . Two summarizations of data contained in the **N x k** matrix leads to effective strategies for parameter estimation (see Figure 2).

One, known as the "score-group method," clusters together all those persons who had a particular raw score, and then counts within each cluster the number of correct responses to each item. This



2a : Score-group Summarization

2b : Pair-wise Item Summarization

Figure 2 : Data reduction strategies for Rasch parameter estimation.

produces a **score-group** by **item** matrix as in Figure 2A. The other method considers the items two at a time, and counts for each pair the number of persons who responded correctly to the first but incorrectly to the second. This is known as the "pair-wise" approach and produces an **item** by **item** matrix as in Figure 2B. (A parallel analysis comparing the people two at a time can be developed theoretically, but has found little practical application.) Both the score-group and the pair-wise approaches are described by Rasch in his 1960 book, but without the development of a maximum likelihood technique he was unable to exploit them.

The score-group method produces a **(k + 1)** by **k** matrix, but since raw scores of zero and k do not contribute to the estimation procedure, the summary yields k**(k - 1)** elements for use in the estimation algorithm. The pair-wise approach results in a **k** by **k** matrix in which the leading diagonal elements are always zero, so again there are **k(k - 1)** elements in the summary on which the estimation algorithm operates.

Analysis of the score-group matrix to separate information on and and thus obtain fully conditioned estimates for both the item difficulty parameters and the abilities associated with membership of score-group 1 through k - 1 is computationally demanding and expensive. The best available procedure has been programmed by Gustafsson (1977), but, though mathematically elegant and statistically sound, it is far too expensive for routine use. However, Wright has shown that estimates developed from the margins of the score-group matrix can be developed very easily using a maximum likelihood approach. Though the simultaneous estimation of both α and δ sets of parameters introduces a bias, a simple expansion factor applied to the results can largely correct for this (Wright & Douglas, 1977; Habermann, 1977), and this method is widely used in practice. When the data are summarized in a score-group fashion, they are convenient for checking the assumption of equal discriminating power between items, and the tests of fit developed by Wright and Mead (1976) concentrate on this.

By contrast, the pair-wise approach separates information about the δ's from information about the α's at the beginning. The matrix of counts summarized in Figure 2B has conditioned out all information about variations in δ, so that a fully conditional estimate of the δ's (either by maximum likelihood or least squares) can be obtained. The ability estimates for each individual are developed from solving iteratively the equation:

$$r - \sum_{i=1}^{k} \frac{W^{\alpha-\delta_i}}{1 + W^{\alpha-\delta}} = 0$$

where r is the raw score of the person, and the summation extends only over those items that were attempted.

The test of fit applied to the pair-wise summary matrix is not very sensitive to violations of the equal discrimination power assumption (see section 3), but instead focuses on the issue of local independence between items (Choppin & Wright, in progress). In practice, therefore, the two approaches may be regarded as complementary.

Though slower than the Wright estimation algorithm based on score-group marginals, the pair-wise approach has the considerable advantage of being able to handle incomplete data matrices--corresponding to all those applications in which not every person attempts every item. It is thus of particular interest in such fields as adaptive testing and item banking (Choppin, 1978, 1982).

2. The Measurement Philosophy and Primary Focus of Interest

Although it turns out that the mathematical details have much in common with those of "item response theory", Rasch derived his models from a very different standpoint. In the first paragraph of the preface to the book which launched his ideas on measurement (Rasch, 1960) he quotes approvingly an attack by B.F. Skinner on the application of conventional statistical procedures to psychological research.

> The order to be found in human and animal behavior should be extracted from investigations into individuals ... psychometric methods are inadequate for such purposes since they deal with groups of individuals. (Skinner, 1956, p. 221)

Group-centered statistics, which form the backbone of conventional psychometric practice (factor analysis, analysis of variance, etc.), require the clustering of individuals into discrete categories or populations, and further make assumptions about the nature of variation within these categories which Rasch viewed with grave distaste. The alternative was to develop methods which would work with individuals.

> Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments - tests, or items or other stimuli - within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class - measuring the same thing - independent of which particular individuals within the class considered were instrumental for the comparison. (Rasch, 1960, p. vii)

In this excursion into what he later calls "specific objectivity", Rasch is echoing a theme developed explicitly by L.L. Thurstone three decades earlier:

> A measuring instrument must not be seriously affected in its measuring function by the object of measurement. To the extent that its measurement 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, 1928, p.547).

Reliance on this form of analogy to the physical sciences is quite characteristic of latent trait measurement theorists. Wright (1968, 1977) also uses the yardstick as a convenient metaphor for a test item. Others (Eysenck, 1979; Choppin, 1979, 1982) have pointed out the similarities between the measurement of mental traits and the measurement of temperature. The underlying premise is that although psychological measurement may be rather more difficult to accomplish than is measurement in the fields of physics and chemistry, the same general principles should apply. Features which are characteristic of good measurement techniques in physics should also be found in the fields of psychology and education.

Rasch himself draws out the similarity between the development of his model, and Maxwell's analysis of Newton's laws of motion in terms of the concepts force and mass (Maxwell, 1876). The second law links force, mass and acceleration in a situation where although acceleration and its measurement have been fully discussed, the concepts mass and force are not yet defined. Rasch (1960, pp. 110-114) considers the necessity of defining the two concepts in terms of each other, and shows how appropriate manipulation of the mathematical model (the "law") and the collection of suitable data can lead to the (comparative) measurement of masses, and the (comparative) measurement of forces. He points out the close analogy to his item-response model which links ability, difficulty and probability. Ability and difficulty require related definitions since people need tasks on which to demonstrate their ability, and tasks only exhibit their difficulty when attempted by people. Since his model is "specifically objective", data can be collected so that the two sets of parameters are capable of separate estimation (as with force and mass).

This approach to measurement is the primary focus of interest for the Rasch model. Individuals are to be measured through the estimation of parameters characterizing their performance. These parameters shall be interpretable by comparison with the parameters estimated for other individuals (as in norm-referencing) and/or in conjunction with the parameter estimates for test stimuli (as in criterion-referencing).

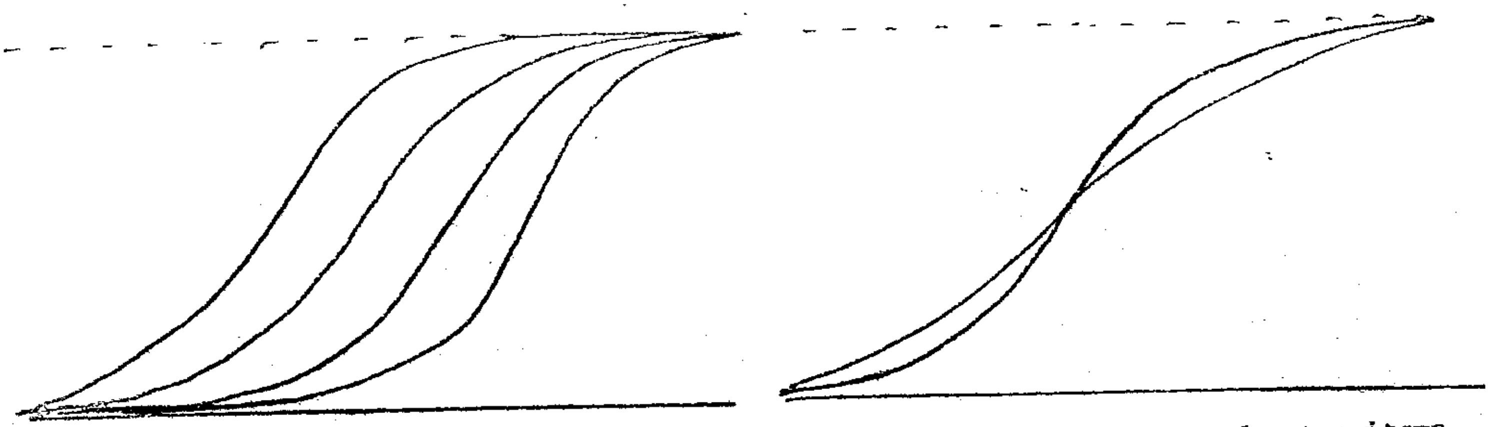
3. Assumptions made by the Rasch Model

The basic assumption is a simple yet powerful one that derives from the requirement of specific objectivity, so central to Rasch's thinking about measurement. It is that the set of people to be measured, and the set of tasks (items) used to measure them, can each be uniquely ordered in terms respectively of their ability and difficulty. (Ability and difficulty as already described.) This ordering permits a parameterization of people and tasks that fits the simple model defined in section 1 above.

The basic assumption has a number of important implications. One such assumption is that of local independence. The probability of a particular individual responding correctly to a particular item must not depend upon the responses that have been made to the previous items. If it did, then altering the sequence of items that made up a particular test would alter the ordering of people on the underlying trait (in violation of the basic assumption). Similarly, local independence requires that the response of an individual to a particular item is not affected by the responses given by other people to the same item. If it were, then it would be possible, by selective clustering of people, to change the ordering of items in terms of their difficulty (in violation of the basic assumption).

Another implication that follows from the basic assumption of the model is sometimes stated (rather confusingly) as "equality of discrimination". It must be emphasized that this does not mean that all items are assumed to have equal point-biserial correlation indices with total test score, or with some external criterion. Rather, it means that the signal/noise ratio represented by the maximum slope of the characteristic curve of each item is assumed to be the same for all items. If the slopes were not the same, then at some point the characteristic curves for two items would cross. This would mean that the ordering of the items in terms of difficulty for persons of lower ability would not be the same as the ordering for persons of higher ability (see Figure 3). This again violates the basic assumption.

Figure 3



(a) Characteristic curves for items that fit the Rasch Model.

(b) Characteristic curves for two items with different discriminations.

Uni-dimensionality is also a consequence of the basic assumption. If the performance of people on a set of items depended on their individual standing on two or more latent traits, such that the ordering of people on these latent traits was not identical, then it would be impossible to represent the interaction of person and task with a single person parameter for ability.

A further assumption and one which is mathematically very convenient, albeit somewhat unrealistic (at least on multiple-choice items), is that there is no random guessing behavior. The model requires that for any test item, the probability of a successful response tends asymptotically to zero as the ability of the person attempting it is reduced (see Figure 1).

Similarly, there is a built in assumption, which has been much less carefully explored, that as the ability of the person being considered increases, the probability of a successful response to any given item approaches one.

4. Strengths and Weaknesses and Gaps in the Development

The strong features of the Rasch model when compared with other measurement models are:

(a) The combination of specific objectivity, a property taken for granted in the field of physical measurement, and the model's mathematical simplicity.

(b) Deriving from this, the separability property which permits the estimation of person-parameters and item-parameters separately.

(c) The existence of several algorithms for parameter estimation some of which are extremely fast and which work well with small amounts of data.

(d) The inbuilt flexibility of the system. As with other latent trait models which are defined at the item level, there is no requirement that tests be of a fixed length or contain the same items.

(e) The close parallels that exist between the Rasch model and the conventional practice of calculating raw scores based on an equal weighting of items. Rasch models are the only latent-trait models for which the raw score, as conventionally defined, is a sufficient statistic for ability (and correspondingly the raw difficulty or p-value of an item is a sufficient statistic for Rasch difficulty).

Against this it must be admitted that there are areas of considerable weakness. The most serious focuses on the assumptions made by the model. These are, in general, too strong to carry full credibility. In practice some real data appear to fit the model rather poorly. The assumptions of local independence and of no guessing (which are crucial to the model) are not strictly met in practice. Although the psychometrician may be able to reduce the guessing problem through the avoidance of objective items, and may be able to structure the test and the conditions under which it is administered to improve local independence, in real life situations these problems are rarely completely eliminated. The model also demands (as do most others) uni-dimensionality (or, as Rasch calls it, conformability), and while the items that comprise many existing tests fail to meet this criterion, the problem is less critical. If one has control over the test construction phase of a measurement program, then it is possible to build sets of items which satisfy the uni-dimensionality assumption moderately well.

One feature of the model which has been described as a weakness (Goldstein, 1979; Divgi, 1981) is that it implies a unique ordering of items, in terms of their difficulty, for all individuals. This appears not to be sufficiently sensitive to the effects of instructional and curriculum variation, and stands, therefore, as an important criticism (but see Bryce, 1981).

The seriousness with which such objections need to be considered depends upon the nature of the measurement task being addressed. Most educational instruction programs aim at increasing the learning of the student and thus at increasing his ability to solve relevant test items. We would usually expect the ability to solve all relevant test items to increase--but the relative difficulty of the items could (and normally would) remain unchanged. While this is the dominant goal of instruction, the model can handle the situation appropriately, and the occasional changes in relative difficulty brought about by alternative curricula (see, for example, Engel, 1976 or Choppin, 1978) can shed considerable light on the real effects of the instructional program. If, however, a section of curriculum is aimed specifically at breaking down some piece of learning and replacing it with another (i.e. making some items more difficult to solve, and other easier) such as may occur during revolutionary changes in society, then we may well feel that the simple model proposed is inadequate to describe the situation. In this case the items measuring the "old" learning and the "new" do not seem to belong on the same scale. Such circumstances, however, are not routine in the United States.

Similarly, we find in general that the ordering of item difficulties is the same with respect to all students. Where one student differs significantly in finding some item much harder or easier than predicted by the model, then we have valuable diagnostic information about that individual (Mead, 1975). In practice we rarely find evidence for such differences, and where they do occur the interpretation is usually clear and direct (for example, the student missed instruction on a particular topic). If we were attempting to measure in an area where there was no common ordering of item difficulties for most students, then the model would appear quite inappropriate. Such situations may be simulated by creating test items whose solution depends upon luck or chance, but this is far removed from purposive educational testing.

Experience over the last two decades suggests that the simplification made by the model in requiring a unique ordering of items is met adequately in practice. Deviations, where they do occur, are indicators of the need for further investigation (Dobby & Duckworth, 1979; Choppin, 1977). There seems little reason, therefore, to regard this as a weakness of the Rasch approach.

5. Areas of Application

The basic form of the model proposed by Rasch, and described in section 1, dealt with the simplified situation where only two possible outcomes of a person attempting a test item were considered (i.e the response is scored "right" or "wrong"). For this reason, perhaps, most of the applications so far developed have been confined to the use of "objective" test items for the measurement of achievement since these are most naturally scored in this fashion.

(a) Item Analysis

The most frequent application of the model has been for item analysis. Users have wanted to confirm that the model fits data they have already accumulated for existing tests; they seek clues as to why particular tests are not functioning as well as they should; or in the construction of new tests they seek guidance as to which items to include and which to omit.

It is probably true to say, however, that the Rasch model has not proved particularly valuable in any of these three roles. It can detect lack of homogenity among items, but is probably less sensitive to this than is factor analysis. It can identify items that do not discriminate or for which perhaps the wrong score key has been selected, but it seems no more effective at this than is the more traditional form of item analysis. The exception to this generalization probably comes when tests are being tailored for a very specific purpose. Wright and Stone explore this in "Best Test Design" (1979). Careful adherence to all the steps they outline would probably yield a test with better characteristics for the specific and intended purpose than would a test produced on the basis of only traditional forms of item analysis and the crude criteria they employ.

(b) Scaling and Equating

A serious problem of traditional testing is that the "score" produced can only be interpreted in terms of the particular test used. The development of norms for standardized tests is an attempt to overcome this problem but this too has serious limitations. Latent trait scaling has been used to tackle this question directly. With the Rasch model, the raw scores on one test are mapped onto their latent trait scale, and different tests can of course have their scores mapped onto the same scale (provided always that the dimension of ability being measured is the same). The method has been used to compare "quasi-parallel" tests (e.g., Woodcock, 1973; Willmott & Fowles, 1974); to link the tests given at different stages of a longitudinal study (Engel, 1976; Choppin, 1978); and to check on the standardization characteristics of batteries of published tests (Rentz & Bashaw, 1976, 1977).

It should perhaps be noted that although equating using the Rasch model appears more flexible than traditional procedures in that only the difficulty level of the two tests is being compared and other characteristics such as test length, the distribution of item difficulties, etc. maybe quite different, there is an implicit assumption that the "discrimination power" (in the sense discussed above) of the items in the two tests are comparable. As a rule this implies that the item types are similar. Attempts to use the Rasch model to equate multiple choice and essay type tests on the same topic have led to inconsistent and bizarre results (Willmott, 1979; Vincent, 1980).

(c) Item Banking

Item banks take the equating of test scores to its logical limit by calibrating all possible performances on all possible tests composed of items drawn from a fixed set (the bank).

> When a family of test items is constructed so that they can be calibrated along a single common dimension and when they are employed so that they retain these calibrations over a useful realm of application, then a scientific tool of great simplicity and far reaching potential becomes available. The "bank" of calibrated items can serve the composition of a wide variety of measuring tests. The tests can be short or long, easy or hard, wide in scope or sharp in focus. (Wright, 1980).

An item bank requires calibration, and although in theory there are alternative approaches, in practice the Rasch model has proved by far the most cost effective and is the most widely used (Choppin, 1979).

(d) Quality of Measurement.

An important development that is facilitated by latent trait scaling is the calculation of an index to indicate the quality of measurement for each set of test data, and if necessary for each person attempting a test or for each item. The Rasch model, for example, yields an explicit probability for each possible outcome of every interaction of a person and an item. Where, overall, the probabilities of the observed outcomes are too low we may deduce that for some reason the Rasch model does not offer an adequate description of a particular set of data. If the probabilities are generally in the acceptable range, but are low for a particular item, then we may conclude that this is an unsatisfactory item. Perhaps it does not discriminate, or is addressing some different dimension of achievement. If the probabilities are generally acceptable but are low for a specific person, then we may conclude that this person was not adequately measured by the test (perhaps he guessed at random, was insufficiently motivated, or misunderstood the use of the answer sheet). The reporting for this person of a low measurement quality index would imply that the person's score should be disregarded and that a retest is appropriate.

A recent extension of this approach involves trying to identify within the vector of item responses from a particular individual those portions which provide reliable measurement information, on which items (or groups of items) the subject appears to have guessed at random, and how the total vector of responses may be selectively edited in order to provide a more reliable estimate of the subject's level of achievement.

6. Extensions to the Basic Model

Two types of adaptation and extension will be considered here. The first centers around the notion of sequential testing in which evidence of the level of ability of the subject is accumulated in Bayesian fashion during the test session and may be used to determine which items are to be attempted at the next point of the sequence and/or when to terminate the testing session. This approach relies upon the existence of difficulty calibrations for a pool or bank of test items. Most of the reseach that has been done so far has employed computers to manage the testing session: to select items for the subject to answer, to keep track of measurement quality, to generate up-to-date estimates of the ability of the subject (together with the appropriate standard errors) and to decide when the session should be terminated. Wright and Stone (1979) point out that individual people can do most of this for themselves if provided with suitable guidelines and computational aids, and in many circumstances making the learner responsible for evaluating his own learning is a useful thing to do.

The second area of development from the basic Rasch model is in the extension from simple dichotomous scoring of items (right-wrong) to a more complex system. Two separate situations need to be considered. The first is when an item is not answered completely but enough is done to earn some partial credit. Data would then consist of scores in the range 0 to 1 for each item. The other case is that which typically occurs with rating scales or attitude measures when the respondent is asked to choose one from among a finite number of discrete categories, and each category contains information about the standing of the respondent on some latent trait. Douglas (1982) has considered the theoretical implications of generalizing the basic Rasch model to include both these cases, and it turns out that almost everything that can be done for dichotomous items can also be done for these more complex methods of scoring. For the rating scale problem both Andrich (1977) and Wright and Masters (1982) have found it convenient to concentrate on establishing the location of thresholds (the point at which the probability for responding in one category passes the probability of responding in the next one - Figure 4). Wright and Masters have produced some interesting theorems about the importance of these thresholds being properly ordered, and about the spacing of thresholds that maximizes the information gained. There have been few practical applications of this approach to date.

Figure 4

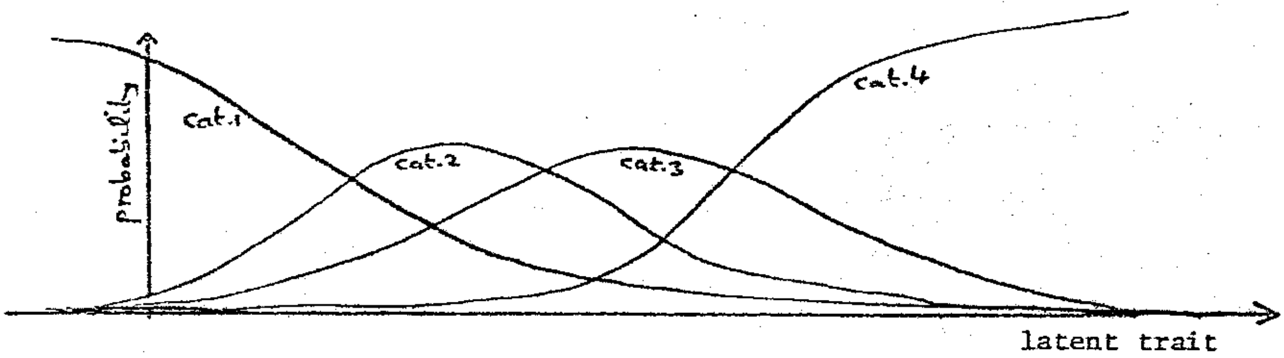


Figure 4 : The Probability of Responding in various categories.

For the analysis of "partial credit" data two computer programs (CREDIT by Masters and POLYPAIR by Choppin) have been devised and applied to real data sets. The latter program, for example, was used in the assessment of writing skills which forms part of the British National Assessment Program.

7. Points of Controversy

In some ways the Rasch model represents a revolutionary approach to educational measurement that discards many time-honored constructs in testing theory (e.g., true score, measurement error, and reliability). On the other hand, it can be viewed as providing a comprehensive and sound mathematical underpinning for the conventional practice of using raw scores, and shows that in most testing applications raw scores are all that are required. From this point of view the Rasch model may be seen as less radical than other latent trait models. Perhaps because the former view of the model was the first to catch the imagination in the United States and has dominated efforts to popularize it, it has been a subject of continuing controversy. The most strident arguments are not concerned with how best to use the Rasch model, but whether or not its use is ever appropriate.

To some extent the Rasch model has been central in the general attack on latent trait theory as applied to the measurement of student achievement. Goldstein (1979) who has led this attack on the other side of the Atlantic, stresses the fundamental difference between what he regards as well-ordered traits such as aptitude and intelligence on the one hand, and the complex pattern of behaviors that we call educational achievement on the other. In his view it makes no sense to apply *any* unidimensional model to the assessment of achievement.

Less extreme in their implications are the arguments within the latent trait camp about whether the Rasch (i.e., one-parameter) model is adequate for achievement testing, or whether a more complex (usually three-parameter) model is indicated.

It is important to differentiate two kinds of usage. One is in test construction where in general the users of Rasch models appear to be on firm ground in claiming that a strategy to develop and select items that conform to the Rasch model will produce better test instruments than would other more conventional strategies. The other type of usage is concerned with the analysis of existing test data (for example, the massive data sets of NAEP or the accumulated files of SAT material at ETS) where items are likely to be so varied (and in many cases so poor) that it is comparatively easy to show that the Rasch model is not appropriate. Devotees of the Rasch model react to this by dropping the non-fitting items (which may well be the majority) and working with those that are left--but this cavalier approach does not commend itself to many researchers. If one is interested in analyzing and scaling data sets which include some possibly very bad items, then something like the three-parameter model is going to be needed.

This difference of emphasis among the areas of application has its origins in contrasting views of measurement philosophy. As Hambleton (1983) makes clear, the Rasch model can be regarded as a special case of the three-parameter model when the discrimination parameters are held equal, and the "guessing" parameter is fixed at zero. Mathematically, this view is undoubtedly correct--but philosophically, it is very misleading. Rasch developed his model, in ignorance of Lord's seminal work on item characteristic curves, on the basis of a set of features which were necessary for an objective measurement system. For measurements with the required properties he found that his model, or a simple mathematical transformation of it, was the mathematically unique solution. The three-parameter model that forms the basis of Lord's Item Response Thoery is not, and cannot be, "specifically objective". Those whose main interest is in understanding existing data sets, and therefore in careful modeling of observed ICCs, see little benefit or relevance in speific objectivity. Those who wish to construct instruments to measure individuals optimally tend to prefer the approach which offers the stronger and more useful system. ICCs which reflect the behavior of inefficient or ineffective items have little interest for them. As has been suggested earlier in this paper, the Rasch model supports a range of applications which goes well beyond what a latent trait model that is not specifically objective can manage.

In the view of this writer, much of the energy which has fueled professional arguments over which is the better model (and the many research studies whose main goal was to compare the effectiveness of the two models in exploring a particular set of data) stem from a failure to appreciate that the two models are basically very different, and were developed to answer different questions. Neither is ever "true". Both are merely models, and it seems clear that in some applications one is of more use than the other and vice versa.

Among users of the Rasch model there is little that is currently controversial, due in no small part to the dominance of two computer programs now in use around the world (BICAL developed by Wright and his associates in Chicago, and PAIR developed by Choppin in London). One current issue that requires clarification concerns the status of "tests of fit". It is generally conceded by Rasch users that whereas better tests of fit are available for the Rasch model than for most other psychometric models, they still leave a lot to be desired. In most cases, showing that an item does not fit the model merely requires collecting a sufficiently large body of data. The area of disagreement lies between those who prefer to treat fit/misfit as a dichotomous categorization and draw up decision rules for dealing with test data on this basis, and those who prefer to regard degree of misfit as a continuous variable which needs to be considered in the context of the whole situation. The present writer belongs in the latter camp, but is prepared to admit that many of the "rules of thumb" that have been developed lack much theoretical or empirical basis.

References

Andrich, D. A rating formulation for ordered response categories. Psychometrika, 1978, 43, 561-73.

Bryce, T.G.K. Rasch-fitting. British Educational Research Journal, 1981, 7, 137-153.

Choppin, B. The national monitoring of academic standards. Paper read to National Council on Measurement in Education, Toronto, 1977.

Choppin, B. Item banking and the monitoring of achievement. Slough, England: National Foundation for Educational Research, 1978.

Choppin, B. Testing the questions: The Rasch formula and item banking. In M. Raggett (Ed.) Assessment and testing of reading, Ward Lock, London, 1979.

Choppin, B. The use of latent-trait models in the measurement of cognitive abilities and skills. In D. Spearitt (Ed.) The improvement of measurement in education and psychology, Melbourne: ACER, 1982.

Divgi, D.R. A direct procedure for scaling tests with latent trait theory. Paper read at the Annual Meeting of the American Educational Research Association, Los Angeles, 1981.

Douglas, G.A. Conditional inference in a generic Rasch model. In D.Spearitt (Ed.), The improvement of measurement in education and psychology. Melbourne, ACER, 1982.

Engel, I. The differential effect of three different mathematics curricula on student's achievement through the use of sample-free scaling. MA thesis, Tel Aviv University, 1976.

Eysenck, H.J. The structure and measurement of intelligence. Berlin: Springer-Verlag, 1979.

Goldstein, H. Consequences of using the Rasch model for educational assessment. British Educational Research Journal, 1979, 5, 211-220.

Gustafsson, J.E. The Rasch model for dichotomous items. Research Report 63. Institute of Education, University of Goteberg, 1977.

Habermann, S. Maximum likelihood estimates in exponential response models. Annals of Statistics, 1977, 77, 815-841.

Hambleton, R.K. Item response theory: The three-parameter logistic model. CSE Report No. 219. Los Angeles: UCLA Center for the Study of Evaluation, 1983.

Maxwell, J.C. Matter and motion. London, 1876.

Mead, R.J. *Analysis of fit to the Rasch model.* Doctoral dissertation, University of Chicago, 1975.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut, 1960. (Reprinted by University of Chicago Press, 1980)

Rentz, R.R., & Bashaw, W.L. *Equating reading tests with the Rasch model.* Athens, Georgia: Educational Resource Laboratory, 1975.

Rentz, R.R., & Bashaw, W.L. The national reference scale for reading: An application of the Rasch model. *Journal of Educational Measurement*, 1977, 14, 161-180.

Skinner, B.F. A case history in scientific method. *The American Psychologist*, 1956, *11*, 221-233.

Thurstone, L.L. The measurement of opinion. *Journal of Abnormal and Social Psychology*, 1928, *22*, 415-430.

Vincent, D. personal communication, 1980.

Willmott, A. Controlling the examination system. Paper presented at the Schools Council Forum on Comparability of Public Examinations, London, 1979.

Willmott, A., & Fowles, D. *The objective interpretation of test performance: The Rasch model applied.* Windsor: NFER Publishing Co., Ltd., 1974.

Woodcock, R.W. *Woodcock reading mastery tests.* Circle Pines, Minnesota: American Guidance Service, 1974.

Wright, B.D. Sample-free test calibration and person measurement. In *Proceedings of the 1967 invitational conference on testing problems.* Princeton, N.J.: Educational Testing Service, 1968.

Wright, B.D. Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 1977, *14*, 97-116.

Wright, B.D. Afterword. In G. Rasch *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press (1980 edition).

Wright, B.D., & Douglas, G.A. Conditional versus unconditional procedures for sample free item analysis. *Educational and Psychological Measurement*, 1977, *37*, 573-586.

Wright, B.D., & Masters, G. *Rating Scale Analysis*, Chicago: MESA Press, 1982.

Wright, B.D., & Mead, R.J. BICAL: Calibrating items with the Rasch model. *Research Memorandum 23*, Statistics Lab, Education Department, University of Chicago, 1976.

Wright, B.D., & Stone, M.H. Best Test Design, Chicago: MESA Press, 1979.