

A RASCH MODEL FOR PARTIAL CREDIT SCORING

GEOFF N. MASTERS

UNIVERSITY OF CHICAGO

A unidimensional latent trait model for responses scored in two or more ordered categories is developed. This "Partial Credit" model is a member of the family of latent trait models which share the property of parameter separability and so permit "specifically objective" comparisons of persons and items. The model can be viewed as an extension of Andrich's Rating Scale model to situations in which ordered response alternatives are free to vary in number and structure from item to item. The difference between the parameters in this model and the "category boundaries" in Samejima's Graded Response model is demonstrated. An unconditional maximum likelihood procedure for estimating the model parameters is developed.

Key words: latent trait, Rasch model, ordered categories, partial credit.

1. Introduction

In 1960 Rasch introduced a model for the analysis of dichotomously-scored responses. When data fit this model, item parameters can be estimated independently of the characteristics of the calibrating sample and person parameters can be freed from the difficulties of the items taken [Rasch, 1960, 1977]. Since the introduction of this model, other Rasch models which share this potential have been developed for other observation formats. These include models for the analysis of counts [Rasch, 1960], repeated trials [Rasch, Note 1; Andrich, 1978a] and rating scales [Andrich, 1978b].

In this paper a Rasch model is developed for the analysis of partial credit data. In common with all Rasch models, the parameters in this "Partial Credit" model appear additively in the exponent of the model and so can be separated and estimated independently of each other. This separability results in sufficient statistics for the model parameters and makes possible objective comparisons of persons and items from graded responses.

Section 2 distinguishes partial credit scoring from three other sources of ordered category data: repeated trials, counts and rating scales. Section 3 summarizes the concept of parameter separability and discusses its implications for measurement. Section 4 reviews the traditional "category boundaries" approach to the analysis of ordered category data used by Thurstone [Edwards & Thurstone, 1952] and Samejima [1969]. Section 5 develops the Partial Credit model and Section 6 describes its relationship to Andrich's [1978b] Rating Scale model. Section 7 gives an unconditional maximum likelihood procedure for estimating the parameters in the Partial Credit model. Section 8 applies the model to the analysis of a prekindergarten screening test.

2. Observation Formats

Observation formats which record ordered levels of response can be classified into four general types:

Preparation of this paper was supported by grants from the Spencer Foundation and the National Institute for Justice. I would like to thank Professor Benjamin D. Wright of the University of Chicago for his very kind help with the various drafts of this paper.

Requests for reprints should be sent to Geoff Masters, Department of Education, University of Chicago, 5835 S. Kimbark, Chicago, 60637.

Repeated Trials

Repeated trials data result when respondents are given a fixed number of independent attempts at each item on a test. The observation x is the number of successes on the item and takes values from 0 to the number of attempts m . This format is useful for tests of psychomotor skills in which the observation is a count of the number of times in m attempts that a task is successfully performed. Under this format the order in which successes occur is considered irrelevant. Only the number of successes in m attempts is recorded, and these m attempts are assumed independent of each other. The ordered response categories are defined by the counts 0, 1, 2, \dots m .

Counts

A second general type of ordered category data results when there is no upper limit on the number of independent successes (or failures) a person can make on an item. Under this format the observation x may be a count of the number of times a person completes a task in a specified period of time, or a count of the errors a person makes in reading a passage on an oral reading test. The ordered response categories are defined by the counts 0, 1, 2, \dots ∞ .

Rating Scales

A third type of ordered category data comes from rating scales in which a fixed set of ordered response alternatives is used with every item. When performances are being rated, judges may be provided with response alternatives like

POOR FAIR GOOD

Or, when self-reports on attitude are sought, alternatives like

STRONGLY STRONGLY
DISAGREE DISAGREE AGREE AGREE

may be provided. The distinguishing feature of this third observation format is the use of a single set of response categories with every item in the scale.

Partial Credit

The fourth general type of data comes from an observation format which requires the prior identification of several ordered levels of performance on each item and thereby awards partial credit for partial success on items. The usual motive for partial credit scoring is the hope that it will lead to a more precise estimate of a person's ability than a simple pass/fail score. Three items for which partial credit might be awarded are shown in Figure 1. The numbers 0, 1, 2, and 3 in Figure 1 indicate only the ordering of the response categories, and are not used as category "weights".

The highest level of performance on each item in Figure 1 is Level 3. Levels 2, 1 and 0 represent decreasing levels of performance, with Level 0 being the lowest level of performance possible on each item. Under this format the number of steps into which an item is divided and the relative difficulties of these steps can vary from item to item.

A model for the analysis of counts was used by Rasch in the 1950's [Rasch, 1960]. A model for repeated trials data was proposed by Rasch [Note 1] and investigated by Andrich [1978a, 1978c], and a related model for the analysis of rating scales has been explored by Andrich [1978b, 1978d, 1979] and Masters [Note 2]. The model developed in this paper for the analysis of partial credit data is an extension of Andrich's Rating Scale model to situations in which response alternatives are free to vary in number and structure from item to item.




<u>MATHEMATICS ITEM:</u>		$\sqrt{7.5 / 0.3 - 16} = ?$	
Failed		0	
$7.5 / 0.3 = 25$		1	
$25 - 16 = 9$		2	
$\sqrt{9} = 3$		3	
<u>SCREENING TEST ITEM:</u>			
Draw a circle.			
0	1	2	3
No response	Scribble, no resemblance to circle	Lack of closure much overlap, more than 1/3 of figure distorted	Closure, no more than 1/8" overlap, 2/3 figure round
			
(from Mardell and Goldenberg, 1972)			
<u>GEOGRAPHY ITEM:</u>			
The capital city of Australia is			
a. Wellington		1	
b. Canberra		3	
c. Montreal		0	
d. Sydney		2	

FIGURE 1
Three items for which partial credit scoring might be used

3. Parameter Separability

The necessity of being able to free estimates of item parameters from the transient characteristics of calibration samples was noted by Thurstone in 1928 in his discussion of a method for estimating "scale values" for the statements on an attitude scale:

One of the first requirements of a solution (to the problem of constructing a rational method of assigning values for the base line of a scale of opinion) 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 [1928, p. 416]

Tests and questionnaires can be used to make meaningful comparisons among persons only if the item scale values have a generality which extends beyond the calibration sample used to obtain them. Thus an essential requirement of a latent trait model is that it be able

to estimate item parameters independently of the characteristics of the calibration sample. This can be achieved in practice if the person parameters in the model can be conditioned out of the estimation equations for the items. Andersen [1973a] shows that this in turn depends upon the existence of sufficient statistics for the person and item parameters in the model.

Dichotomously-Scored Responses

Of the various latent trait models which have been proposed for dichotomously-scored responses, only *one* has sufficient statistics for person and item parameters. This model, which was developed by Rasch in the 1950's [Rasch, 1960], specifies the probability of a correct response to an item as an exponential function of the difference between person ability β_n and item difficulty δ_i :

$$\pi_{1ni} = \frac{\exp(\beta_n - \delta_i)}{1 + \exp(\beta_n - \delta_i)} \quad (1)$$

where π_{1ni} is person n 's probability of success on item i ,
 β_n is the ability of person n ,
 δ_i is the difficulty of item i ,
 and $\pi_{0ni} = 1 - \pi_{1ni}$ is person n 's probability of failing item i .

This is the only latent trait model for dichotomously-scored responses for which the number of successes r_n is a sufficient statistic for the person parameter β_n . This model permits the elimination of the person parameters β_n , ($n = 1, 2, \dots, N$) from the estimation equations for the items, and so makes possible item estimates which are freed of the particulars of the calibration sample.

Polytomously-Scored Responses

Parameter separability is not confined to Rasch's Dichotomous model. Rasch realized that "the possibility of separating two sets of parameters must be a fundamental property of a very important class of models" [1977, p. 66]. Members of this family of "Rasch models" share the potential for providing measures of ability which are freed of the difficulties of the items taken, and item difficulty estimates which are freed of the abilities of the persons in the calibrating sample.

One member of this family of models was proposed by Rasch [1961] for polytomously-scored responses. The unidimensional form of this model gives the probability of person n responding in the x 'th category to item i as

$$\pi_{xni} = \frac{\exp[\kappa_x + \phi_x(\beta_n - \delta_i)]}{\sum_{k=0}^m \exp[\kappa_k + \phi_k(\beta_n - \delta_i)]} \quad x = 0, 1, \dots, m \quad (2)$$

where κ_x is a parameter associated with response category x , and ϕ_x is a nonparametric "scoring coefficient". Rasch deduced (2) as the necessary and sufficient form for a unidimensional model for $m + 1$ response categories which permits separation of a single position parameter for each item and a single position parameter for each person.

Andersen [1977] has demonstrated that for the raw score $r_n = \sum_{i=1}^L x_{ni}$ to be a sufficient statistic for β_n in (2), the coefficients $\phi_0, \phi_1, \dots, \phi_m$ must have equidistant values (e.g., successive integers). Andrich [1978b] has provided a useful interpretation of the category parameters κ_x , ($x = 0, 1, \dots, m$) when the response categories are successive alternatives on a rating scale.

Various elaborations of (2) have been proposed. The alternative parameters $\beta_{nx} = \beta_n \phi_x$ and $\sigma_{ix} = \delta_i \phi_x - \kappa_x$ were suggested by Andersen [1973b]. With these doubly-subscripted person and item parameters (2) becomes

$$\pi_{xni} = \frac{\exp(\beta_{nx} - \sigma_{ix})}{\sum_{k=0}^m \exp(\beta_{nk} - \sigma_{ik})} \quad x = 0, 1, \dots, m \quad (3)$$

in which there are $m + 1$ parameters for each person and $m + 1$ parameters for each item. An application of this model is described by Andersen [1980].

Fischer [1977] proposed that the item parameter σ_{ix} in (3) be rewritten as a weighted sum of some other "explanatory" parameters $\eta_{1x}, \eta_{2x}, \dots, \eta_{Mx}$:

$$\sigma_{ix} = \sum_h^M w_{ih} \eta_{hx} + c_x$$

where w_{ih} is a weight for parameter η_{hx} and c_x is a normalizing constant. To my knowledge, this polytomous generalization of the linear logistic test model [Fischer, 1972] has not been applied.

The model introduced in this paper for the analysis of partial credit data is a member of this family of models with separable parameters. Like (2), and unlike the models of Andersen and Fischer, the Partial Credit model contains a singly-subscripted person parameter. However, the Partial Credit model is not developed as an elaboration of (2), but is constructed from the basic building block of all Rasch models: Rasch's simple logistic expression.

4. The Category Boundaries Approach

Most of the ordered response category literature is based on an approach which dates back to the psychophysics experiments of the nineteenth century. This approach views ordered response alternatives as regions of a continuum separated by "boundaries". Thurstone's scaling work of the 1920's and Samejima's Graded Response model [1969] use this approach.

The Partial Credit model developed in this paper does *not* parameterize boundaries between regions of a continuum, and so, represents a break with this traditional approach. Before developing the Partial Credit model, Samejima's Graded Response model will be reviewed to clarify the distinction between these two approaches and their two quite different sets of parameters.

Samejima's Graded Response Model

Samejima's [1969] model is an extension of Thurstone's method of successive intervals to the analysis of graded responses on educational tests. The operating assumption of Samejima's model is that when person n encounters item i , a latent random variable ε_{ni} is induced. According to the Graded Response model, the probability of this variable taking a value greater than the k 'th "category boundary" associated with item i depends on the ability β_n of the person, the value of the k 'th item boundary λ_{ik} and the item's discrimination α_i . If ε_{ni} is assumed logistically distributed, then in general the model probability of person n responding *in or above* category k to item i is specified as

$$\pi_{kni}^* = \frac{\exp[\alpha_i(\beta_n - \lambda_{ik})]}{1 + \exp[\alpha_i(\beta_n - \lambda_{ik})]}$$

where λ_{ik} is the boundary between categories k and $k - 1$ associated with item i (Figure 2).

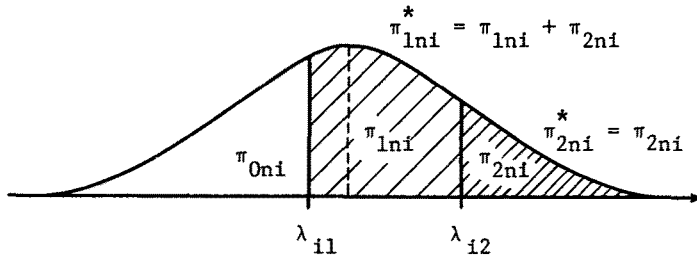


FIGURE 2
Cumulative probabilities π_{1ni}^* and π_{2ni}^* for person n taking item i

When this model is applied, an attempt is made to estimate an ability β_n for each person, a discrimination α_i and a set of boundaries $\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{im}$ for each item.

Samejima's Graded Response model provides no simple general expression for the probability π_{kni} of person n responding in category k to item i . Instead, this probability must be obtained by subtracting cumulative probabilities. In the three-category case, the three category probabilities are

$$\begin{aligned}\pi_{0ni} &= 1 - \pi_{1ni}^* = \frac{1 + \exp[\alpha_i(\beta_n - \lambda_{i2})]}{\Psi} \\ \pi_{1ni} &= \pi_{1ni}^* - \pi_{2ni}^* = \frac{\exp[\alpha_i(\beta_n - \lambda_{i1})] - \exp[\alpha_i(\beta_n - \lambda_{i2})]}{\Psi} \\ \pi_{2ni} &= \pi_{2ni}^* = \frac{\exp[\alpha_i(\beta_n - \lambda_{i2})] + \exp[\alpha_i(2\beta_n - \lambda_{i1} - \lambda_{i2})]}{\Psi}\end{aligned}\quad (4)$$

where Ψ is the sum of the numerators.

These category probabilities π_{0ni} , π_{1ni} and π_{2ni} are plotted for a range of abilities in Figure 3.

From Figure 3 it is seen that for three categories the item "boundaries" λ_{i1} and λ_{i2} are the points where $\pi_{0ni} = 0.5$ and $\pi_{2ni} = 0.5$. As the distance between λ_{i1} and λ_{i2} is increased in Figure 3, the probabilities π_{0ni} and π_{2ni} decrease, and π_{1ni} increases for every value of β_n . In other words, as the "width" of the middle category is increased, every person becomes more likely to respond in this category. As λ_{i1} and λ_{i2} are brought closer together, a

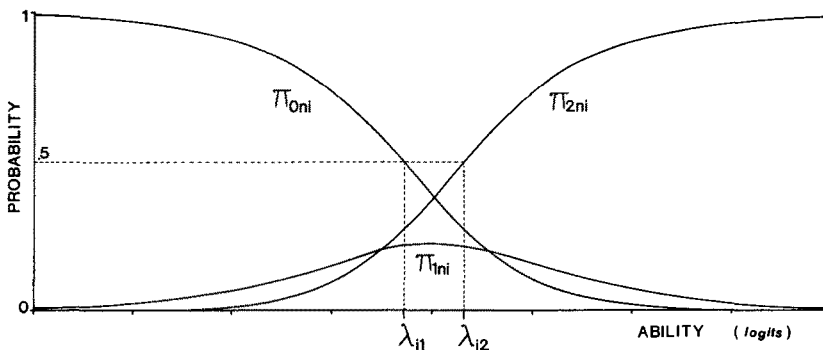


FIGURE 3
Probability curves and category boundaries λ_{i1} and λ_{i2} for item i

response in the middle category becomes less likely until, in the limit, $\lambda_{i1} = \lambda_{i2}$, $\pi_{1ni} = 0$ for every value of β_n , and the three categories collapse to a 0/2 dichotomy.

While the category "boundaries" λ_{i1} and λ_{i2} are readily located in Figure 3, when there are more than three response categories, only the first and last boundaries have obvious locations in a probability plot such as this. For more than three categories there is no simple connection between the probability curves for the categories and the "boundaries" between these categories.

Attempts to represent ordered response categories as regions of a continuum separated by ordered "category boundaries" have a long history. However, the form of the Graded Response model, even in the absence of a discrimination parameter, prevents the algebraic separation of person and item parameters. As a result, no sufficient statistic exists for either the person or item parameters in the model, and the person parameters cannot be conditioned out of the estimation equations for the items. To obtain *separable* person and item parameters, and hence the possibility of "specifically objective" comparisons, a different approach to parameterizing ordered response categories must be taken. Such an approach is now described.

5. The Partial Credit Model

Ordered Performance Levels

When performances on an item are recorded in the $m + 1$ ordered levels 0, 1, \dots , m , it is convenient to think in terms of m "steps" which have to be taken to complete the item. For the three-step mathematics item

$$\sqrt{\frac{7.5}{0.3}} - 16 = ?$$

considered in Figure 1, the first step is to make a 1 rather than a 0 on the item by solving the division $7.5/0.3 = ?$; the second step is to make a 2 rather than a 1 by solving the subtraction $25 - 16 = ?$, and the third step is to make a 3 rather than a 2 by solving $\sqrt{9} = ?$.

These three steps must be taken in order. The second step in the item can be taken only if the first has been completed, and the third step can be taken only if the first two steps have been completed. These three steps are pictured in Figure 4.

One-Step Items

The questions on most educational tests are treated as one-step items. When only one step is identified in an item, performance on the item is scored as either failure or success.

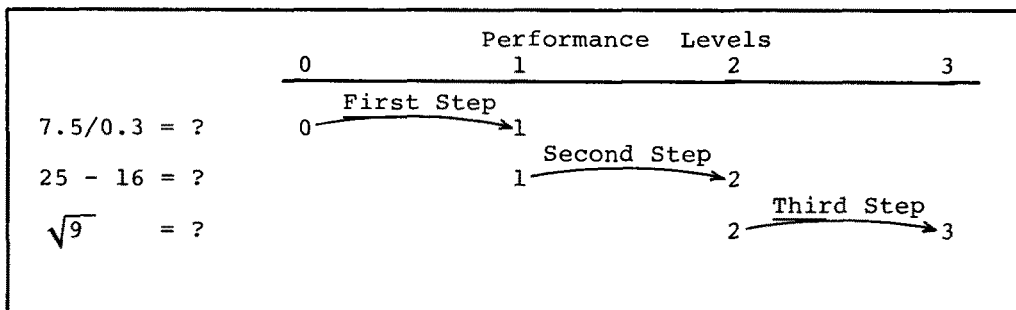


FIGURE 4

A three-step interpretation of the mathematics item in Figure 1

These dichotomously-scored observations can be analyzed with Rasch's Dichotomous model (1), which for the discussion to follow is usefully rewritten as

$$\frac{\pi_{1ni}}{\pi_{0ni} + \pi_{1ni}} = \frac{\exp(\beta_n - \delta_{i1})}{1 + \exp(\beta_n - \delta_{i1})}.$$

This makes explicit that (1) specifies the probability of person n succeeding on item i given that only two outcomes are possible. The item difficulty δ_i is rewritten δ_{i1} to make it explicit that this is the difficulty of the *first* (and in this case, only) step in item i .

m-Step Items

The structure of the mathematics item in Figure 4 invites a "step" interpretation. But this idea can be applied to any item with ordered response alternatives. For an item on an attitude questionnaire, "completing the k 'th step" means choosing the k 'th response alternative over the $(k - 1)$ 'th in response to the item. Thus a person who chooses to AGREE with a statement on an attitude questionnaire when given the ordered categories

STRONGLY DISAGREE	DISAGREE	AGREE	STRONGLY AGREE
0	1	2	3

to choose among, can be considered to have chosen DISAGREE over STRONGLY DISAGREE (first step taken) and also AGREE over DISAGREE (second step taken), but to have failed to choose STRONGLY AGREE over AGREE (third step rejected).

As the steps within an item must be taken in sequence, it is possible to infer from a person's response to an item the steps that the person must have taken in order to arrive at their response. This is illustrated in Table 1.

On the right of Table 1 the scores of ten persons on a three-step item i are shown. Each score x_{ni} can be interpreted as the number of steps completed by person n taking item i , or,

TABLE 1
Pass/Fail Scores for a Three-Step Item i

Person n	Performance Levels				Score x_{ni}
	0 y_{0ni}	1 y_{1ni}	2 y_{2ni}	3 y_{3ni}	
	First Step	Second Step	Third Step		
1	1	1	1	1	3
2	1				0
3	1	1	1		2
4	1	1			1
5	1				0
6	1	1	1	1	3
7	1	1			1
8	1	1	1		2
9	1				0
10	1	1	1	1	3
Count:	S_{i0} =10	S_{i1} =7	S_{i2} =5	S_{i3} =3	

if the ordered performance levels are labelled 0, 1, 2, and 3, as the highest performance level in item i reached by person n . The dichotomous variable y_{kni} in Table 1 takes the value 1 if person n reaches the k 'th performance level in item i , and 0 otherwise. In this case, seven persons completed the first step and reached level 1 ($y_{1ni} = 1$). Of these, five also completed the second step and reached level 2 ($y_{2ni} = 1$), and of these, three went on to complete the third step and reached level 3 ($y_{3ni} = 1$).

With the column of person scores on the right of Table 1 replaced by a column of 0/1 scores for each performance level, it becomes possible to apply Rasch's Dichotomous model (1) to the analysis of these data. In particular, an expression can be written for the probability of person n reaching performance level k in item i :

$$P\{y_{kni} = 1 \mid \beta_n, \lambda_{ik}\} = \frac{\exp(\beta_n - \lambda_{ik})}{1 + \exp(\beta_n - \lambda_{ik})} \quad (5)$$

where β_n is the ability of person n
and λ_{ik} is the difficulty of reaching level k in item i .

As the number of persons reaching level k can never be greater than the number reaching level $k - 1$ in an item,

$$S_{i1} \geq S_{i2} \dots \geq S_{im},$$

and, because S_{ik} is a sufficient statistic for λ_{ik} in (5), estimates for the item "levels" must always be ordered

$$\hat{\lambda}_{i1} \leq \hat{\lambda}_{i2} \dots \leq \hat{\lambda}_{im}.$$

In other words, it must always be easier to reach level $k - 1$ in an item than to reach level k . These "level" difficulties $\lambda_{i1}, \lambda_{i2}, \dots, \lambda_{im}$ have exactly the same interpretation as the ordered "category boundaries" in the Graded Response model.

It is a requirement of Rasch's Dichotomous model that each dichotomously-scored observation y_{kni} be governed by only one person parameter and only one item parameter, and that it be independent of all other influences. But it is impossible for a person in Table 1 to reach level 3 ($y_{3ni} = 1$) if they do not first reach level 1 ($y_{1ni} = 1$), and then level 2 ($y_{2ni} = 1$). This hierarchical dependence among levels violates the intention of the model that $P\{y_{3ni} = 1\}$ be governed by β_n and λ_{i3} only, and raises a serious question about the utility of the Dichotomous model in this context.

This observation provokes an alternative approach to the three-step item in Table 1. Rather than thinking in terms of three ordered "level" difficulties (λ_{i1} the difficulty of reaching level 1; λ_{i2} the difficulty of reaching level 2, and λ_{i3} , the difficulty of reaching level 3), we could think instead in terms of the individual difficulty of each successive step in the item.

The *third* step in item i , for example, is from level 2 to level 3. The difficulty of this third step governs how likely it is that a person who has already reached level 2 will complete the third step to level 3. Another way of saying this is that the difficulty of the third step governs how likely it is that a person will make a 3 *rather than* a 2 on item i . A simple model for this third step in item i is

$$\Phi_{3ni} = \frac{\pi_{3ni}}{\pi_{2ni} + \pi_{3ni}} = \frac{\exp(\beta_n - \delta_{i3})}{1 + \exp(\beta_n - \delta_{i3})}. \quad (6)$$

In this expression $\pi_{2ni} + \pi_{3ni}$ is person n 's probability of scoring either 2 or 3 on item i , and

$\pi_{3ni}/(\pi_{2ni} + \pi_{3ni})$ is the probability of person n completing the third step in item i to score 3 rather than 2. As before, β_n is the position of person n on the variable, but now δ_{i3} is defined as the difficulty of the third "step" in item i . (Notice that while δ_{i3} governs person n 's probability of completing the step from level 2 to level 3, it says nothing about their probability of reaching level 2 in the first place. That depends on the person's ability and the difficulties of the first two steps in the item. Equation (6) gives person n 's probability of completing the third step *if they reach level 2*).

The *second* step in item i is from level 1 to level 2. (A person cannot make a 0 by failing the second step. Neither can they make a 3 by completing this step). A simple expression can also be written for a person's probability of making a 2 rather than a 1 on item i :

$$\Phi_{2ni} = \frac{\pi_{2ni}}{\pi_{1ni} + \pi_{2ni}} = \frac{\exp(\beta_n - \delta_{i2})}{1 + \exp(\beta_n - \delta_{i2})} \quad (7)$$

where β_n continues to be the position of person n on the variable, and δ_{i2} is now the difficulty of the second "step" in item i .

Similarly, the *first* step in item i is to make a 1 rather than a 0:

$$\Phi_{1ni} = \frac{\pi_{1ni}}{\pi_{0ni} + \pi_{1ni}} = \frac{\exp(\beta_n - \delta_{i1})}{1 + \exp(\beta_n - \delta_{i1})}. \quad (8)$$

This last expression is identical to the Rasch model for a one-step item. The only differences are that now $\pi_{0ni} + \pi_{1ni} < 1$, since more than two response categories are provided, and δ_{i1} , while still the difficulty of the first "step" in item i , is not the difficulty of the only step in the item.

Finally, as person n must make one of the four possible scores on item i ,

$$\pi_{0ni} + \pi_{1ni} + \pi_{2ni} + \pi_{3ni} = 1. \quad (9)$$

Equations (6), (7), (8) and (9) are readily solved to obtain one general expression for the probability of person n scoring x on item i :

$$\pi_{xni} = \frac{\exp \sum_{j=0}^x (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})} \quad x = 0, 1, \dots, m_i \quad (10)$$

where for notational convenience, $\sum_{j=0}^0 (\beta_n - \delta_{ij}) \equiv 0$. This is the Partial Credit model. It gives the probability of person n scoring x on the m_i -step item i as a function of the person's position β_n on the variable and the difficulties of the m_i "steps" in item i . The observation x is a *count* of the successfully completed item steps, and only the difficulties of these x completed steps appear in the numerator of the model.

Unlike the item "levels" λ_{ik} , ($k = 1, 2, \dots, m$) in (5), each of which represents the difficulty of reaching performance level k in item i , the individual step difficulties δ_{ik} , ($k = 1, 2, \dots, m$) can be separated from, and estimated independently of the person parameters in the model. Before exploring (10) further, the separability of the parameters in this model is demonstrated.

Establishing Separability and Sufficiency

Under the Partial Credit model, the probability of a person n making any particular response vector (x_{ni}) on an L -item test is

$$P\{(x_{ni}) | \beta_n, ((\delta_{ij}))\} = \prod_{i=1}^L \left[\frac{\exp \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})} \right] \\ = \frac{\exp \sum_{i=1}^L \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})}{\Psi_n} \quad (11)$$

where

$$\Psi_n \equiv \prod_{i=1}^L \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right].$$

If the "score" r_n of person n on an L -item test is defined as the count of the total number of item steps completed by person n (i.e., $r_n = \sum_i x_{ni}$) then the probability of person n making the score r is

$$P\{r | \beta_n, ((\delta_{ij}))\} = \frac{\sum_{(x_{ni})}^r \exp \sum_{i=1}^L \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})}{\Psi_n} \\ = \frac{\exp(r\beta_n)}{\Psi_n} \sum_{(x_{ni})}^r \exp \left(- \sum_{i=1}^L \sum_{j=0}^{x_{ni}} \delta_{ij} \right) \quad (12)$$

where $\sum_{(x_{ni})}^r$ denotes the sum over all those response vectors which produce the score r .

The conditional probability of response vector (x_{ni}) , given the score r , is obtained by dividing (11) by (12):

$$P\{(x_{ni}) | r, ((\delta_{ij}))\} = \frac{P\{(x_{ni}) | \beta_n, ((\delta_{ij}))\}}{P\{r | \beta_n, ((\delta_{ij}))\}} \\ = \frac{\exp(r\beta_n) \exp \left(- \sum_{i=1}^L \sum_{j=0}^{x_{ni}} \delta_{ij} \right)}{\exp(r\beta_n) \sum_{(x_{ni})}^r \exp \left(- \sum_{i=1}^L \sum_{j=0}^{x_{ni}} \delta_{ij} \right)} \\ = \frac{\exp \left(- \sum_{i=1}^L \sum_{j=0}^{x_{ni}} \delta_{ij} \right)}{\sum_{(x_{ni})}^r \exp \left(- \sum_{i=1}^L \sum_{j=0}^{x_{ni}} \delta_{ij} \right)}. \quad (13)$$

The person parameter β_n does not appear in (13). By conditioning on the person's score r , the person parameter is eliminated from this conditional probability expression. This means that if a person makes a score of r on an L -item test, under the Partial Credit model, the way in which this score is made is not governed by the person's ability, but depends only on the relative difficulties of the steps in the L items. In other words, a person's score vector (x_{ni}) contains no more information about the person's ability β_n than we already have in the person's test score r_n , which is thus a sufficient statistic for β_n .

The conditional probability of an entire matrix of responses $((x_{ni}))$ given the vector of person test scores (r_n) can now be written:

$$P\{((x_{ni})) | (r_n), ((\delta_{ij}))\} = \prod_{n=1}^N \left[\frac{\exp\left(-\sum_i^L \sum_{j=0}^{x_{ni}} \delta_{ij}\right)}{\sum_{(x_{ni})}^{r_n} \exp\left(-\sum_i^L \sum_{j=0}^{x_{ni}} \delta_{ij}\right)} \right].$$

The fact that the person parameters do not appear in this expression means that the step difficulties can be estimated independently of the abilities of the persons in the calibrating sample.

Similarly, by conditioning on the sufficient statistics for the item step difficulties, a conditional probability expression containing only the person parameter can be obtained.

The model probability of observing a particular N -person vector of responses (x_{ni}) to an item i is

$$\begin{aligned} P\{(x_{ni}) | (\beta_n), (\delta_{ij})\} &= \prod_{n=1}^N \left[\frac{\exp \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})} \right] \\ &= \frac{\left[\exp\left(\sum_{n=1}^N x_{ni} \beta_n\right) \right] \left[\exp\left(-\sum_{n=1}^N \sum_{j=0}^{x_{ni}} \delta_{ij}\right) \right]}{\Psi_i} \end{aligned} \quad (14)$$

where

$$\Psi_i \equiv \prod_{n=1}^N \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right].$$

The probability of observing some particular vector of counts $(S) = S_{i1}, S_{i2}, \dots, S_{im}$ for this item i is

$$P\{(S) | (\beta_n), (\delta_{ij})\} = \frac{\left[\sum_{(x_{ni})}^{(S)} \exp\left(\sum_{n=1}^N x_{ni} \beta_n\right) \right] \left[\exp\left(-\sum_{n=1}^N \sum_{j=0}^{x_{ni}} \delta_{ij}\right) \right]}{\Psi_i} \quad (15)$$

where $\sum_{(x_{ni})}^{(S)}$ denotes the sum over all those response vectors which produce the item count vector (S) . Dividing (14) by (15) yields the probability of the vector of responses (x_{ni}) given (S)

$$\begin{aligned} P\{(x_{ni}) | (S), (\beta_n)\} &= \frac{P\{(x_{ni}) | (\beta_n), (\delta_{ij})\}}{P\{(S) | (\beta_n), (\delta_{ij})\}} \\ &= \frac{\exp\left(\sum_{n=1}^N x_{ni} \beta_n\right)}{\sum_{(x_{ni})}^{(S)} \exp\left(\sum_{n=1}^N x_{ni} \beta_n\right)} \end{aligned} \quad (16)$$

in which the item step difficulties do not appear. By conditioning on the observed vector of item counts (S) , the item parameter has been eliminated entirely. This means that under the Partial Credit model, all the information available in a data matrix about the difficulties of the item steps is contained in a simple count of the number of persons completing each step in an item. No further information about the step difficulties can be obtained by keeping

track of any other aspect of the performances of individuals, and so, the item counts ((S_{ij})) contain all the information available in the data matrix about the step difficulties ((δ_{ij})).

The implication of being able to eliminate the item parameters from (16) is that person measures can be freed of the particulars of the items used. When data fit the Partial Credit model, the potential exists for making statistically equivalent measures of person ability from tests of different lengths containing items which differ in step structure and difficulty. The separability of the parameters in this model makes it a member of the Rasch family of models.

Item Steps and Model Probabilities

To illustrate the relationship between the item parameters $\delta_{i1}, \delta_{i2}, \dots, \delta_{im}$ in the Partial Credit model and the model probabilities $\pi_{0ni}, \pi_{1ni}, \dots, \pi_{mni}$, a two-step item i is now considered in which the first step is somewhat easier to complete than the second ($\delta_{i1} = -1.0$ logits and $\delta_{i2} = +1.0$ logits). The model probability curves for this item are shown in Figure 5.

At the top of Figure 5 the probability of completing the first item step (i.e., the probability of responding in category 1 *rather than* in category 0) and the probability of completing the second item step (i.e., the probability of responding in category 2 *rather than* in category 1) are plotted for a range of abilities. In general, the probability of completing the k 'th step in item i is described by a simple logistic ogive of slope one and location δ_{ik} .

The category probability curves for this item are plotted at the bottom of Figure 5. A comparison of these curves with the curves in Figure 3 shows that the item parameters in the Partial Credit and Graded Response models have quite different meanings. In the

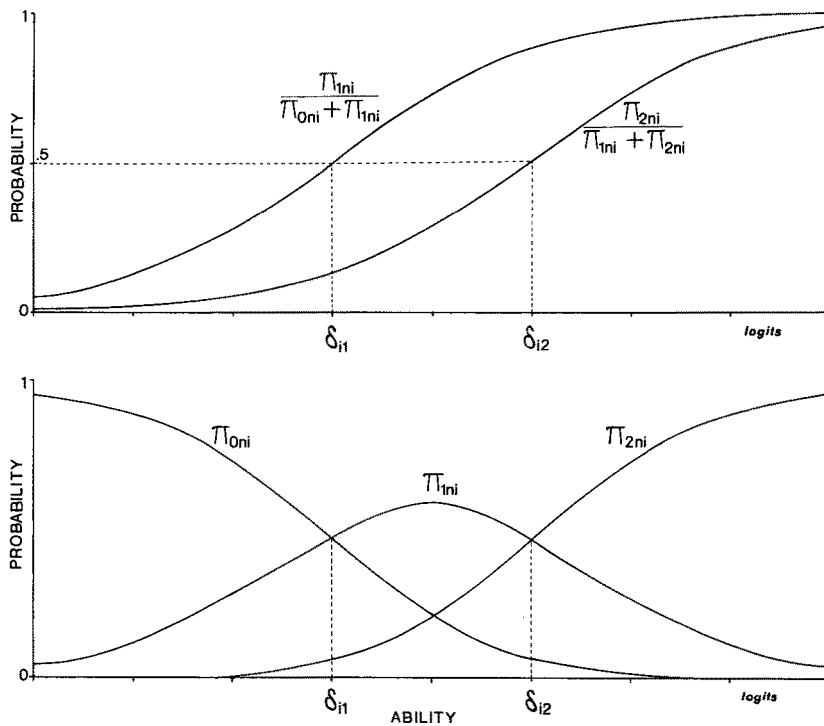


FIGURE 5
Category probability curves ($\delta_{i1} < \delta_{i2}$)

Partial Credit model, the k 'th item "step" δ_{ik} is the point on the ability continuum where the probability curves for categories $k - 1$ and k intersect (i.e., where $\pi_{k-1, ni} = \pi_{kni}$). In the Graded Response model, the k 'th item "boundary" λ_{ik} is the point on the continuum where $\pi_{kni}^* = \sum_{j=k}^{mi} \pi_{jni} = 0.5$.

As the step difficulties δ_{i1} and δ_{i2} are brought closer together in Figure 5, every person's probability of completing *only* the first step in item i (i.e., every person's probability of scoring 1) decreases. When the second step is made *easier* than the first ($\delta_{i2} < \delta_{i1}$), the probability curve for the middle response category drops still further, and every person becomes even less likely to complete only the first step (Figure 6).

6. A Special Case: Rating Scales

The Partial Credit model imposes no particular expectations concerning the relative difficulties of the steps within any item. Some will be easy, others hard, depending upon the item subtasks, and quite regardless of the necessary order in which the steps must always be taken.

In a rating scale, however, the ordered response levels are not defined by a series of item subtasks, but by the fixed set of ordered rating points used with the items. As the same set of rating points is used with every item, the *relative* difficulties of the steps within each item should not vary greatly from item to item.

This expectation can be incorporated into the Partial Credit model by resolving each item step into two components so that

$$\delta_{ij} = \delta_i + \tau_j$$

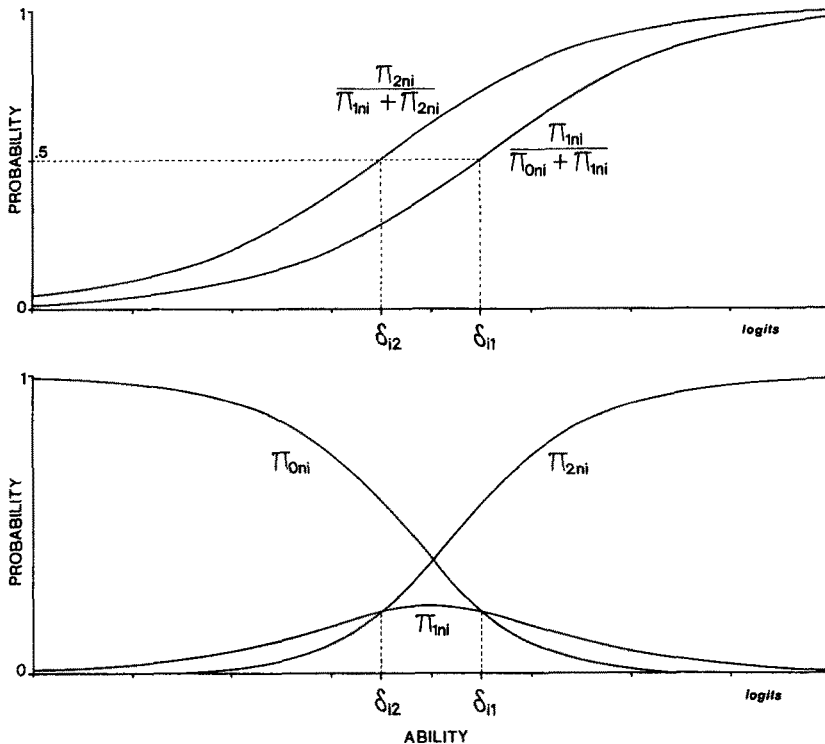


FIGURE 6
Category probability curves ($\delta_{i2} < \delta_{i1}$)

where δ_i is the location or "scale value" of item i on the variable and τ_j is the location of the j 'th step in each item relative to that item's scale value. Under this simplification, the only difference remaining among items is the difference in location on the variable. The pattern of item steps around this location, which is supposed to be determined by the fixed set of rating points used with the items, is described by the parameters $\tau_1, \tau_2, \dots, \tau_m$, and is estimated once for the entire item set. This simplification is illustrated in Figure 7.

With $\delta_{ij} = \delta_i + \tau_j$ the Partial Credit model becomes

$$\pi_{xni} = \frac{\exp \sum_{j=0}^x [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^m \exp \sum_{j=0}^k [\beta_n - (\delta_i + \tau_j)]} \quad x = 0, 1, \dots, m.$$

This Rating Scale model [Andrich, 1978b] can be rewritten

$$\pi_{xni} = \frac{\exp \left[- \sum_{j=0}^x \tau_j + x(\beta_n - \delta_i) \right]}{\sum_{k=0}^m \exp \left[- \sum_{j=0}^k \tau_j + k(\beta_n - \delta_i) \right]}$$

or, as Rasch's general unidimensional model (3):

$$\pi_{xni} = \frac{\exp[\kappa_x + \phi_x(\beta_n - \delta_i)]}{\sum_{k=0}^m \exp[\kappa_k + \phi_k(\beta_n - \delta_i)]} \quad \text{where } \kappa_x \equiv - \sum_{j=0}^x \tau_j;$$

and $\phi_x \equiv x$.

When this model is applied to the analysis of a rating scale, a position on the variable β_n is estimated for each person n , a scale value δ_i is estimated for each item i , and m response "thresholds" $\tau_1, \tau_2, \dots, \tau_m$ are estimated for the $m + 1$ rating categories. Applications of this model to the analysis of attitude questionnaires and performance ratings are described by Andrich [1978d, 1979], Masters [Note 2] and Wright and Masters [Note 3, 1982].

7. Estimation

An unconditional maximum likelihood procedure for estimating the parameters in the Partial Credit model is now outlined. This procedure is based on Wright and Panchapakesan's [1969] estimation algorithm for Rasch's Dichotomous model. A conditional esti-

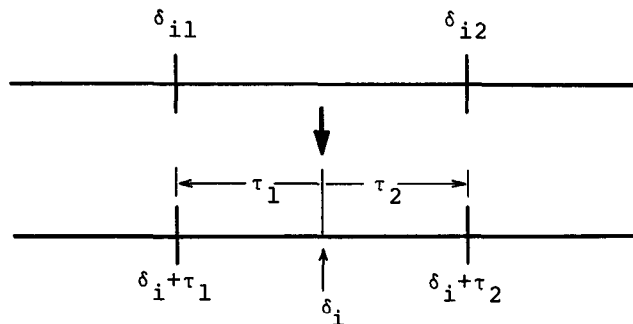


FIGURE 7
The Rating Scale simplification

mation procedure for the Partial Credit model is described by Wright and Masters [1982].

The likelihood of the data matrix $((x_{ni}))$ is the continued product of the probability π_{xni} over all values of n and i :

$$\Lambda = \prod_n \prod_i \pi_{xni} \exp \sum_n \sum_i \sum_{j=0}^{x_{ni}} (\beta_n - \delta_{ij})$$

$$= \frac{\prod_n \prod_i \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right]}{\prod_n \prod_i \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right]}.$$

Taking logarithms,

$$\lambda \equiv \log \Lambda = \sum_n \sum_i x_{ni} \beta_n - \sum_n \sum_i \sum_{j=1}^{x_{ni}} \delta_{ij} - \sum_n \sum_i \left[\log \sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right]$$

in which $\sum_{j=0}^{x_{ni}} \delta_{ij} = \sum_{j=1}^{x_{ni}} \delta_{ij}$ because $\delta_{i0} \equiv 0$.

The log likelihood can be simplified by first letting $r_n = \sum_i x_{ni}$. This is person n 's "score" on the L -item test, and is defined as a count of the total number of completed item steps. To further simplify the log likelihood it is noted that $\sum_j^{x_{ni}} \delta_{ij}$ is the sum of the difficulties of the steps in item i completed by person n . These completed step difficulties can be summed over all N persons to obtain $\sum_n \sum_j^{x_{ni}} \delta_{ij}$, the sum of the difficulties of all steps completed by the sample of N persons. Since S_{ij} is defined as the number of persons completing step j on item i , this sum can be rewritten $\sum_n \sum_j^{x_{ni}} \delta_{ij} = \sum_{j=1}^{m_i} S_{ij} \delta_{ij}$. (This can be seen in Table 1 where summing the difficulties of the completed steps across each row and then summing over the persons gives the same result as counting the number of persons completing each step and weighing the step difficulties by these counts to form the sum $S_{i1} \delta_{i1} + S_{i2} \delta_{i2} \dots + S_{im_i} \delta_{im_i}$). With these simplifications the log likelihood becomes

$$\lambda = \sum_n r_n \beta_n - \sum_i \sum_{j=1}^{m_i} S_{ij} \delta_{ij} - \sum_n \sum_i \log \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right].$$

The fact that the observation r_n appears only once in this expression, in combination with the parameter β_n , and that the observation S_{ij} appears only once, in combination with the parameter δ_{ij} , permits the objective comparison of persons and items. The log likelihood takes this form *only* when the parameters of the model are linear in the argument of an exponential.

The first and second derivatives of λ with respect to β_n and δ_{ij} are simplified by

$$\frac{\partial \log \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij}) \right]}{\partial \beta_n} = \frac{\sum_{k=0}^{m_i} k \exp \sum_{j=0}^k (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (\beta_n - \delta_{ij})}$$

$$= \sum_{k=0}^{m_i} k \pi_{kni} = \sum_{k=1}^{m_i} k \pi_{kni}$$

and

$$\begin{aligned} \frac{\partial \log \left[\sum_{k=0}^{m_i} \exp \sum_{h=0}^k (\beta_n - \delta_{ih}) \right]}{\partial \delta_{ij}} &= \frac{- \sum_{k=j}^{m_i} \exp \sum_{h=0}^k (\beta_n - \delta_{ih})}{\sum_{k=0}^{m_i} \exp \sum_{h=0}^k (\beta_n - \delta_{ih})} \\ &= - \sum_{k=j}^{m_i} \pi_{kni} \end{aligned}$$

in which the difficulty δ_{ij} of step j appears only in those terms for which $k \geq j$ so that the derivative of $\sum_{k=0}^{m_i} \delta_{ik}$ with respect to δ_{ij} truncates the summation $\sum_{k=0}^{m_i}$ to $\sum_{k=j}^{m_i}$.

With these results the first derivatives of λ with respect to β_n and δ_{ij} are:

$$\frac{\partial \lambda}{\partial \beta_n} = r_n - \sum_i^L \sum_{k=1}^{m_i} k \pi_{kni} \quad n = 1, N. \quad (17)$$

$$\frac{\partial \lambda}{\partial \delta_{ij}} = -S_{ij} + \sum_n^N \sum_{k=j}^{m_i} \pi_{kni} \quad i = 1, L; j = 1, m_i. \quad (18)$$

In (17) $\sum_{k=1}^{m_i} k \pi_{kni}$ is the number of steps person n is expected to complete in item i . When summed over items this becomes the number of steps person n is expected to complete on the L -item test (i.e., the expected value of r_n).

In (18) $\sum_{k=j}^{m_i} \pi_{kni}$ is the probability of person n completing at least j steps in item i . When summed over the N persons, this becomes the number of persons expected to complete at least j steps in item i (i.e., the expected value of S_{ij}).

The second derivatives of λ with respect to β_n and δ_{ij} are:

$$\frac{\partial^2 \lambda}{\partial \beta_n^2} = - \sum_i^L \left[\sum_{k=1}^{m_i} k^2 \pi_{kni} - \left(\sum_{k=1}^{m_i} k \pi_{kni} \right)^2 \right] \quad (19)$$

$$\frac{\partial^2 \lambda}{\partial \delta_{ij}^2} = - \sum_n^N \left[\sum_{k=j}^{m_i} \pi_{kni} - \left(\sum_{k=j}^{m_i} \pi_{kni} \right)^2 \right]. \quad (20)$$

Equations (17) and (18) can be solved by an iterative procedure such as Newton-Raphson. The person and item estimates are improved using

$$\begin{aligned} b_n^{t+1} &= b_n^t - \frac{r_n - \sum_i^L \sum_{k=1}^{m_i} k P^t}{- \sum_i^L \left[\sum_{k=1}^{m_i} k^2 P^t - \left(\sum_{k=1}^{m_i} k P^t \right)^2 \right]} \quad n = 1, N \\ d_{ij}^{t+1} &= d_{ij}^t - \frac{-S_{ij} + \sum_n^N \sum_{k=j}^{m_i} P^t}{- \sum_n^N \left[\sum_{k=j}^{m_i} P^t - \left(\sum_{k=j}^{m_i} P^t \right)^2 \right]} \quad i = 1, L; j = 1, m_i \end{aligned}$$

where b_n^t is the estimate of β_n after t iterations, d_{ij}^t is the estimate of δ_{ij} after t iterations, and P^t is the estimated probability of person n responding in category k to item i .

To overcome the indeterminacy in the scale origin, the mean step difficulty $d..$ is set

equal to zero. Asymptotic estimates of the standard errors are given by

$$SE(b_n) = \left[\sum_i^L \left(\sum_{k=1}^{m_i} k^2 P - \left(\sum_{k=1}^{m_i} kP \right)^2 \right) \right]^{-1/2}$$

$$SE(d_{ij}) = \left[\sum_n^N \left(\sum_{k=j}^{m_i} P - \left(\sum_{k=j}^{m_i} P \right)^2 \right) \right]^{-1/2}.$$

Andersen (1973a) has shown that item estimates obtained using an unconditional maximum likelihood procedure contain a slight bias. This bias is effectively removed in the dichotomous case by multiplying the estimates by $(L - 1)/L$ (Wright & Douglas, 1977). Initial studies with data simulated to fit the Partial Credit model suggest that this same correction may be appropriate for removing bias in item step estimates when $m > 1$.

8. Example

The Partial Credit model has been used to analyze performances on a screening test (DIAL, Developmental Indicators for the Assessment of Learning) constructed to identify learning problems among prekindergarten children [Mardell & Goldenberg, 1972, 1975]. The data are from five hundred children between two-and-a-half and five-and-a-half years of age observed in Northbrook, Illinois in 1978. Performances on the fourteen items in the Fine-motor/Cognitive section of the test are analyzed here. The DIAL users' manual defines four performance levels for each item. In Item 2, for example, each child is asked to construct a three-block tower, a three-block bridge and a six-block pyramid, in that order. Each new structure is attempted only after the preceding ones have been completed. The four levels of performance on Item 2 are shown in Figure 8.

The partial credit analysis of these data provides three step estimates d_{i1} , d_{i2} and d_{i3} for each item on the DIAL test. These estimates, which govern the probability of scoring 1 rather than 0, 2 rather than 1, and 3 rather than 2 on each item, are shown in Table 2 together with their calibration errors and a statistic summarizing the fit of each item to the Partial Credit model.

The three step estimates for each item define a unique set of probability curves for the four performance levels in that item. The probability curves for Item 2 ("Building blocks") are shown at the top of Figure 9. The estimates $d_{21} = -.91$, $d_{22} = -.93$ and $d_{23} = 1.29$ are located at the intersections of curves 0 and 1, 1 and 2 and 2 and 3. For children with estimates below $-.9$ logits the most probable score on Item 2 is 0. Children with ability estimates below $-.9$ logits will most probably not even complete the three-block tower. Children with estimates between $-.9$ logits and 1.3 logits will most probably complete both the three-block tower and the three-block bridge to score 2 on Item 2, and children with ability estimates greater than 1.3 logits will most probably complete all three structures to score 3.

The estimated probability curves for two other DIAL items (Item 4, "Copying shapes" and Item 8, "Sorting blocks") are shown in Figure 9. The three items in Figure 9 have quite

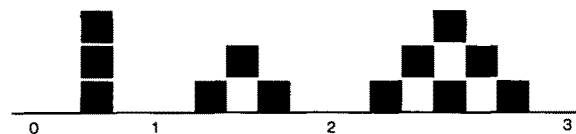


FIGURE 8
Item 2: building blocks

TABLE 2
Item Statistics for DIAL Test

Item Number i	Step Estimates			Step Errors			Mean Square v_i	Error q_i	Fit t_i
	d_{i1}	d_{i2}	d_{i3}	s_{i1}	s_{i2}	s_{i3}			
1	-1.33	-1.28	-1.03	.55	.35	.18	.96	.16	-.21
2	-.91	-.93	1.29	.35	.21	.11	.84	.08	-2.02
3	-.91	.98	.21	.26	.17	.13	1.04	.09	.48
4	-2.25	1.21	3.47	.35	.12	.12	.74	.06	-4.81
5	-1.34	1.72	3.40	.24	.12	.12	.72	.06	-5.28
6	1.81	1.07	1.46	.15	.15	.12	1.31	.08	3.79
7	.32	.86	2.21	.18	.14	.11	1.12	.07	1.71
8	.58	.63	-.49	.22	.19	.15	.96	.10	-.37
9	-1.76	-.09	.19	.41	.20	.13	1.13	.10	1.38
10	-2.50	-.85	2.28	.55	.19	.11	.92	.06	-1.27
11	-1.49	-.83	2.66	.38	.18	.11	1.24	.07	3.33
12	-2.20	-1.33	-.48	.67	.32	.15	1.22	.13	1.67
13	-2.25	-1.80	1.66	.65	.27	.11	.96	.07	-.60
14	-.54	-2.11	.74	.43	.30	.12	1.08	.10	.88
Mean		0.00			.24		1.02	.09	-.10
S.D.		1.61			.15		.18	.03	2.65

different patterns of probability curves. These different patterns can be understood by examining the details of the scoring schemes for Items 2, 4 and 8.

The probability curves for Item 2 ("Building blocks") show that the completion of only the first step in this item is a relatively improbable event. This is because the first two steps in Item 2 are about equally difficult. A child with a good chance of completing the three-block tower also has a good chance of completing the three-block bridge. Because the third step is significantly more difficult than the first two, however, it takes significantly more ability to succeed on this third step. The combined effect of an easy second step and a hard third step makes a score of 0 or 2 more likely than a score of 1 on Item 2.

In Item 4 each child is shown four shapes, a circle, a cross, a square and a triangle, and is asked to copy each shape with a pencil. Three points can be earned on each shape, giving a maximum of twelve points for the item. Four performance levels on Item 4 are then defined in terms of these points:

Item 4	Performance Level			
	0	1	2	3
Minimum Points Needed:	0	1	6	10.

A child who makes "no response" to a shape earns no points. One point is earned for simply marking the paper (e.g., scribbling). Two points are earned for a rough but recognizable copy of the shape, and three points are earned for an accurate copy. When the points earned on all four shapes are summed and converted to a score of 0 to 3 on the item, we see that to score 1 on Item 4 a child needs only to mark the paper *once* in response to any of the four shapes. This is an extremely easy first step, and should be failed only by children who do not cooperate. To score 2 on Item 4 a child must earn at least six points. This requires *recognizable* copies of at least two shapes, and so, defines a much higher level of functioning than scribbling. Finally, to score 3 on Item 4 a child must earn at least ten

points, for example by making *accurate* copies of two shapes and recognizable copies of two shapes. This requires a relatively high degree of coordination, and defines a still higher level of functioning.

The probability curves for Item 4 are different from the curves for Item 2 in that there is now a wide range of abilities for which 1 is the most probable score. In Item 2 the first two steps are about equally easy, and so, the probability of completing only the first step is never high. In Item 4, however, the first step is very easy and the second step is relatively hard, and so, for children between $d_{41} = -2.25$ logits and $d_{42} = +1.21$ logits, the most probable outcome on Item 4 is the completion of only the first step (scribbling).

In Item 8 each child is given a pile of twenty-four blocks and asked to arrange the blocks into six four-block squares, each of a different color. One point is earned for identifying and grouping any four blocks of the same color. An additional point is earned if the four blocks are arranged in a square. Since six squares are possible, a total of twelve points can be earned. Four performance levels on Item 8 are then defined in terms of these points:

Item 8	Performance Level			
	0	1	2	3
Minimum Points Needed:		+ 1 point	+ 5 points	+ 3 points
	0	1	6	9

To score 1 on Item 8 a child needs one point. This can be earned by finding any four blocks of the same color. To score 2 a child must earn another five points by grouping the remaining five colors or by arranging some color groups into squares. Finally, a score of 3 requires three more points. These can be earned by arranging three color groups into squares. The estimated difficulties of the three steps in Item 8 are $d_{81} = .58$, $d_{82} = .63$ and $d_{83} = -.49$.

The second step in Item 8 is not significantly more difficult than the first, making the grouping of only *one* color a relatively improbable event. A child with a good chance of finding four blocks of the same color has a good chance of finding all blocks of the same color. The third step in Item 8 is to score only three more points to make a total of nine rather than six points. This step is estimated to be easier than either of the preceding steps, meaning that every child is estimated to be less likely to complete the second step than to complete the third step *if they reach it*. Consequently, the completion of only two steps is also unlikely. Children with abilities below .2 logits where the '0' curve crosses the '3' curve for Item 8 will most probably fail to group any blocks of the same color, and so, score 0 on Item 8. Children with abilities above .2 logits will most probably earn nine or more points, and so, score 3.

At the bottom of Figure 9 the sample ability distribution is shown. Below this distribution Items 2, 4 and 8 have been used to mark out the ability variable that these items help to define. The lowest ability level is defined by the first step in Item 4 from "no response" to "scribbling". Even children with abilities as low as -2.2 logits are more likely to scribble than not to respond. Higher ability levels are defined by the completion of three-block structures like the tower and the bridge, and the sorting of colored blocks. The highest ability level is defined by step 3 in Item 4 from "recognizable" to "accurate" drawings of simple shapes. Only children with ability estimates above $+3.5$ logits are likely to make "accurate" drawings of simple shapes.

The first step in Item 4 is very easy for these children. Unless this item is to be used with children of much lower abilities, Item 4 could probably be made more informative by

making its first step more difficult. Item 8 presents a different problem. The second and third steps in this item are no more difficult than the first, meaning that a child with a good chance of completing the first step in Item 8 has a good chance of completing all three steps. It should be possible to make Item 8 more informative by making its second and third steps harder to complete.

The fit of each item to the Partial Credit model is summarized in Table 2. The fit analysis of item i is based on estimates of the expected value E_{ni} , expected variance W_{ni} and expected kurtosis C_{ni} of each child's score on item i :

$$E_{ni} = \sum_{k=0}^m k P_{kni}$$

$$W_{ni} = \sum_{k=0}^m (k - E_{ni})^2 P_{kni}$$

$$C_{ni} = \sum_{k=0}^m (k - E_{ni})^4 P_{kni}$$

where P_{kni} is the estimated probability of child n scoring k on item i . When performances on item i fit the Partial Credit model,

$$v_i = \frac{\sum_n^N (x_{ni} - E_{ni})^2}{\sum_n^N W_{ni}}$$

should be approximately distributed as a mean square with expected value one and expected variance

$$q_i^2 = \frac{\sum_n^N (C_{ni} - W_{ni}^2)}{\left(\sum_n^N W_{ni}\right)^2}.$$

This can be standardized to

$$t_i = (v_i^{1/3} - 1) \left(\frac{3}{q_i} \right) + \frac{q_i}{3}$$

which, when data fit the model, has a mean near zero and a standard deviation near one [Wright & Masters, 1982].

Two of the DIAL items have fit values greater than +3.0. Two others have fit values less than -3.0. The item with the most positive fit value is Item 6 ($t = 3.79$). The task in Item 6 is to touch the thumb of each hand to the fingers of that hand in sequence. If a child touches all fingers of one hand, but not in sequence, he scores 1 on Item 6. If he touches the fingers of one hand in sequence, he scores 2. Finally, if he touches the fingers of both hands in sequence, he scores 3.

To explore the misfit of Item 6 the 500 children were divided into ten ability strata and the proportions of children in each ability group scoring 0, 1, 2 and 3 on Item 6 were compared to the corresponding model probabilities [Masters & Wright, Note 4]. Figure 10 shows the proportion of children in each ability group succeeding on the *third* step in Item 6 (i.e., the proportion of children completing two steps who went on to complete the third). The operating curve for this step is shown. The proportion of low ability children completing this step is higher, and the proportion of high ability children is lower than expected.

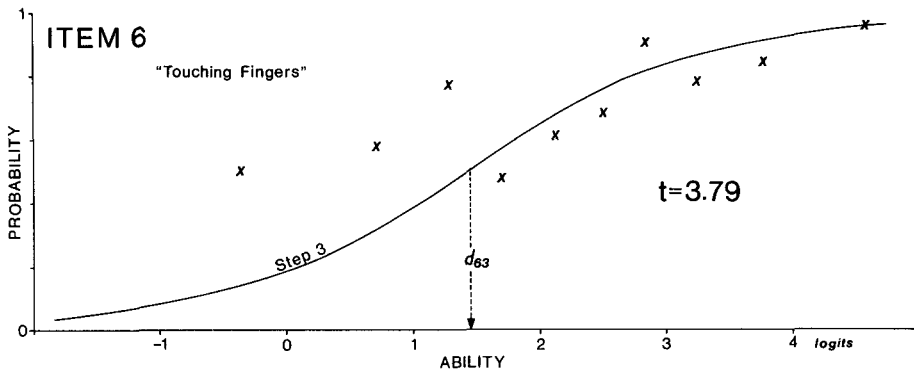


FIGURE 10
Observed proportions and model probabilities for step 3 in Item 6

The ambidextrousness needed to complete this task does not seem to be closely related to the fine-motor variable defined by the majority of these items. Item 6 should probably be revised so that it no longer assesses ambidextrousness.

The item with the most negative fit statistic is Item 5 ($t = -5.28$). This item is identical to Item 4 except that it asks the child to copy the letters *E*, *N*, *D* and *S*. The first step is to mark the paper. The second is to draw "recognizable" copies of at least two letters, and the third is to draw "accurate" copies of at least two letters. When the proportions of children scoring 0, 1, 2 and 3 on Item 5 are compared to the model probabilities, the misfit of this item can be traced to its *second* step from "scribbling" to "recognizable" letters. Figure 11 shows the proportion of children in each ability group completing this second step. Among children who score 1 or 2 on Item 5, fewer low ability children and more high ability children than expected score 2. In other words, performances on this step from scribbling to recognizable drawings are more highly correlated with total test score than has been modelled.

The misfit picture for Item 5 is typical of items which interact with specific instruction. Item 4 has a very similar misfit picture. These are the only items on the test which ask children to draw—skills which are highly correlated with school exposure. The problem with including items like 4 and 5 in a test dominated by items which do not function in this way is that they disorganize the definition of the variable. The estimated difficulty of the second step in Item 5 interacts with the ability level of the children in the calibrating

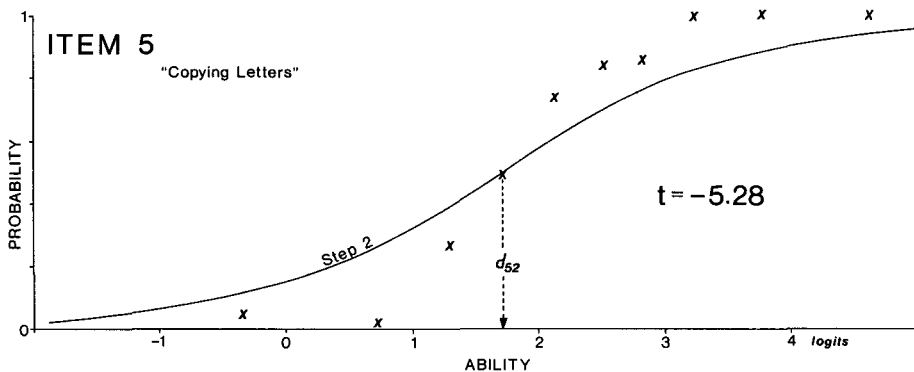


FIGURE 11
Observed proportions and model probabilities for step 2 in Item 5

sample. When this step is calibrated on high ability children, who tend to be older and hence more likely to have been taught how to copy, the step is estimated to be relatively easy. When calibrated on low ability children, who tend to be younger and hence less likely to have been taught how to copy, it is estimated to be relatively hard.

This example brings out three important features of a partial credit analysis. First, the pattern of probability curves for the performance levels in an item is fixed by the difficulties of the item's subtasks, and so can vary from item to item (Figure 9). Each item's pattern can be understood by examining the details of its scoring scheme.

Second, when a test is constructed of items upon which more than two levels of performance are possible, the latent ability variable is usually best understood not in terms of a single "location" parameter for each item, but in terms of the transitions or "steps" between adjacent performance levels. Thus while it would be possible to characterize Items 2, 4 and 8 in Figure 9 by single item location parameters δ_2 , δ_4 and δ_8 (corresponding to the average difficulty of the steps in each item), these item location parameters would not be as useful as the individual step difficulties for understanding the ability variable that these items define. Estimating Item 2 to be "easier" or "harder" than Item 4 is not particularly informative here. Because the first step in Item 4 is so easy, it is easier to earn some credit on Item 4. Because the third step in Item 4 is so difficult, it is easier to earn full credit on Item 2.

A further advantage of identifying and parameterizing individual item steps in a measurement model is that it provides an opportunity to discover particular steps, like step 1 in Item 4 and steps 2 and 3 in Item 8, which could be revised to make the item more informative.

Third, in this test it was possible to trace the source of an item's misfit to a particular step in that item. When performances on individual item steps are studied, the patterns of misfit familiar in the analysis of dichotomously-scored items are found. Item steps which interact with ability level to become "poorly discriminating" result in large positive fit values. Item steps which interact with ability level to become "too discriminating" result in large negative fit values. Both types of misfit confound the definition of this ability variable.

9. Summary

A unidimensional Rasch model for ordered response categories is developed. This "Partial Credit" model provides the opportunity to estimate item parameters independently of the characteristics of the calibrating sample, and to free person measures from the particulars of the items taken.

The Partial Credit model is developed by parameterizing the difficulties of a series of "steps" in each item. The difficulty of the k 'th step in an item is supposed to govern the probability of responding in category k rather than in category $k - 1$, and person n 's probability of completing this k 'th step is specified as

$$\phi_{kni} = \frac{\pi_{kni}}{\pi_{k-1, ni} + \pi_{kni}} = \frac{\exp(\beta_n - \delta_{ik})}{1 + \exp(\beta_n - \delta_{ik})}$$

where π_{kni} is the probability of person n responding in category k to item i , β_n is the ability of person n , and δ_{ik} is the difficulty of the k 'th "step" in item i .

From this expression, and with the requirement that person n must respond in one of the available categories (i.e., $\sum_{k=0}^{m_i} \pi_{kni} = 1$), it follows that

$$\pi_{kni} = \frac{\exp \sum_{j=0}^k (\beta_n - \delta_{ij})}{\sum_{h=0}^m \exp \sum_{j=0}^h (\beta_n - \delta_{ij})} \quad k = 0, 1, \dots, m_i.$$

This is the Partial Credit model. The separability of the parameters in this model results in sufficient statistics for person ability and step difficulty. For a person's ability the sufficient statistic is the count of the total number of steps the person completes, and for an item's step difficulties the sufficient statistics are counts of the number of persons completing each step. The separability of the model parameters permits person abilities to be eliminated from the estimation equations for the items entirely, thereby making possible sample-free estimates of step difficulty.

The crucial difference between the Partial Credit model and Samejima's [1969] Graded Response model is the *separability* of the parameters in the Partial Credit model. In Samejima's model, the item parameters are "category boundaries" which, because they are based on *cumulative* probabilities, are always ordered $\lambda_{i1} \leq \lambda_{i2} \cdots \leq \lambda_{im}$. But these ordered "category boundaries" cannot be separated from the person parameters in the model, and so their estimates can be expected to have a limited generality.

To achieve separable parameters the Partial Credit model is expressed in terms of the difficulties of the individual "steps" in each item. These item steps need not be ordered in difficulty. The k 'th step in an item may or may not be more difficult than the $k - 1$ 'th, depending on the subtasks in the item, and quite regardless of the order in which the steps must always be taken.

REFERENCE NOTES

1. Rasch, G. Objektivitet i samfundsvideenskaberne et metodeproblem. Paper presented at the University of Copenhagen, 1972 (mimeo).
2. Masters, G. N. A Rasch model for rating scales. Doctoral dissertation, University of Chicago, 1980.
3. Wright, B. D. & Masters, G. N. The measurement of knowledge and attitude. *Research Memorandum No. 30*, MESA Psychometric Laboratory, Department of Education, University of Chicago, 1981.
4. Masters, G. N. & Wright, B. D. A model for partial credit scoring. *Research Memorandum No. 31*, MESA Psychometric Laboratory, Department of Education, University of Chicago, 1981.

REFERENCES

- Andersen, E. B. *Conditional inference and models for measuring*, Copenhagen: Mentalhygiejnisk Forlag, 1973a.
- Andersen, E. B. Conditional inference and multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1973b, 26, 31-44.
- Andersen, E. B. Sufficient statistics and latent trait models. *Psychometrika*, 1977, 42, 69-81.
- Andersen, E. B. *Discrete statistical models with social science applications*. Amsterdam: North-Holland Publishing Company, 1980.
- Andrich, D. A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology*, 1978a, 31, 84-98.
- Andrich, D. A rating formulation for ordered response categories. *Psychometrika*, 1978b, 43, 561-573.
- Andrich, D. Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, 1978c, 38, 665-680.
- Andrich, D. Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, 1978d, 2, 581-594.
- Andrich, D. A model for contingency tables having an ordered response classification. *Biometrics*, 1979, 35, 403-415.
- Edwards, A. L. & Thurstone, L. L. An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika*, 1952, 17, 169-180.
- Fischer, G. H. A measurement model for the effect of mass-media. *Acta Psychologica*, 1972, 36, 207-220.
- Fischer, G. H. Some probabilistic models for the description of attitudinal and behavioral changes under the influence of mass communication. In W. F. Kempf and B. H. Repp, *Mathematical models for social psychology*. Vienna: Hans Huber, 1977.
- Mardell, C. & Goldenberg, D. S. *DIAL: Developmental Indicators for the assessment of learning*. Highland Park, Ill.: DIAL Inc., 1972.

- Mardell, C. & Goldenberg, D. S. For prekindergarten screening information: DIAL. *Journal of Learning Disabilities*, 1975, 8, 140-147.
- Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut, 1960 (Chicago: University of Chicago Press, 1980).
- Rasch, G. On general laws and the meaning of measurement in psychology. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1961, 321-333.
- Rasch, G. On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 1977, 14, 58-94.
- Samejima, F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement No. 17*, 1969.
- Thurstone, L. L. The measurement of opinion. *Journal of Abnormal and Social Psychology*, 1928, 22, 415-430.
- Wright, B. D. & Douglas, G. A. Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, 1977, 37, 47-60.
- Wright, B. D. & Masters, G. N. *Rating scale analysis*. Chicago: MESA Press, 1982.
- Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, 29, 23-48.

Manuscript received 5/7/81

Final version received 1/4/82