THE ASSESSMENT OF PARTIAL KNOWLEDGE

By Min-Ning Yu

余　民　寧*

摘　要

由於傳統的「答對才計分」(number－right scoring) 方式無法解決諸如猜題 (guessing) 和部份知識 (partial knowoldge) 的問題，爲此尋求更嚴謹，更精確的理論模式和估計考生實際能力的方法，便成爲迫切需要的課題；當代測驗理論中的潛在特質理論 (latent trait theory) 或試題反應理論 (item response theory) 便是我們所需要的。本文提出一個類似羅氏 (Rasch) 的對數模式 (logistic model)，即「部份得分模式」(partial credit model)，來補救傳統計分方法的缺失；並舉出一個實例說明：採用一種簡便方法 (即PROX法) 來估計考生能力和描述試題特性。本文評論「部份得分模式」的優缺點，並指出未來的研究方向。

Abstract

Because the traditional number-right scoring method cannot solve problems such as guessing and partial knowledge, searching for theoretically rigorous and more precisely estimating methods in measuring examinees' abilities are emergent. Latent trait theory or item response theory developed in modern test theory is the one that we need. This paper proposes a Rasch-type logistic model, partial credit model, to compensate for the defect of the traditional scoring method. An easier procedure, PROX, is used to illustrate how to assess item calibrations and person measures. Several applications and potential weaknesses of the partial credit model are briefly discussed. A future research is mentioned too.

* 作者現爲教育系副教授及兼代實驗小學校長

I. Introduction

With the increased use of computers in scoring students' or examinees' tests, educators' and psychologists' interests have come to be focused on developing more theoretically rigorous models and more precisely estimating methods in describing the ability or aptitude measures and the characteristics of test items. The development of item response theory (**IRT**) in the mental test literature reflects this trend and this importance (Hambleton & Cook, 1977; Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Hambleton & van der Linden, 1982; Traub & Lam, 1985).

Typically, the traditional number-right scoring method is used to estimate examinees' abilities or aptitudes. The sum of the correct responses on a given test is representative of an examinee's ability or aptitude. In fact, this method scores the multiple-choice tests with unit weights given to correct response choices (e.g., score as one) and with constant weights assigned to incorrect response choices (e.g., score as zero). This scoring mothod, however, cannot give a satisfactory estimate of an examinee's ability measures. On the contrary, it has at least two drawbacks in describing an examinee's ability measures: (a) It does not avoid the "guessing" problem which usually occurs in the lower ability examinees or in the extremely difficult items; (b) It cannot tell apart the examinees who really do not know the answers and are guessing wrong, from the examinees who know partially about the answers and are guessing wrong too. The latter question is usually termed the "partial knowledge" problem (Coombs, Milholland, & Womer, 1956; Dressel & Schmid, 1953; Lyerly, 1951). Partial knowledge, although not presenting full information about an examinee's complete ability, represents the partial result of instruction and learning. Its presence requires a more precisely estimating method to be used. Consequently, several alternative scoring methods have been proposed to compensate for the drawback of this scoring method.

The common use of remedies for the former drawback in the number-right scoring method is the formula score (Coombs et al., 1956; Glass & Wiley, 1964; Lord, 1963, 1964, 1975) or the correction for guessing (Cureton, 1966; Davis, 1959, 1967; Diamond & Evans, 1973; Jackson, 1955; Little, 1962; Lyerly, 1951; Sax & Collet, 1968; Stanley & Wang, 1968; Wang & Stanley, 1970). The correction for guessing uses the following formula:

$$S = R - \frac{W}{K - 1} \tag{1.1.1}$$

where S is an examinee's corrected score for guessing,

R is the number of items answered right,
W is the number of items answered wrong,
K is the number of choices for each item.

This equation is based on an assumption that all wrong answers are guessed wrong and that all correct answers are obtained either by "full" knowledge or by "lucky" guessing. The presence of omitted responses and partial knowledge is not taken into account. Obviously, this equation cannot give us any information about examinees' ability measures which are intermediate in scoring correct and scoring wrong items. Thus, this problem invites alternatives to the number-right scoring method.

Some alternatives to the number-right scoring method are the use of differential weighting schemes (Davis & Fifer, 1959; Hambleton, Roberts, & Traub, 1970; Hendrickson, 1971; Patnaik & Traub, 1973; Reilly & Jackson, 1973; Sabers & White, 1969), and confidence testing (de Finetti, 1965; Hambleton et al., 1970; Rippey, 1968; Shuford, Albert, & Massengill, 1966) for assessing examiness' partial knowledge (Coombs et al., 1956; de Finetti, 1965; Hambleton et al., 1970). The differential weighting schemes (or, alternatively, called option or choice weight scoring) refer to procedures whereby different weights are assigned to all the options or choices of an item. That is, the weights assigned to any one item need not be similar to those assigned to another item. Obviously, nondiscriminating options or choices may be given zero weights and a completely nondiscriminating item would be excluded from the test. The confidence testing requires an examinee to indicate his/her confidence in the correctness of each response choice, instead of simply selecting one response. It also uses the specialized scoring procedures and discourages guessing. The findings of these alternatives are usually interpreted in terms of test validity and reliability. However, they do not provide a convincing statement about examinees' partial knowledge by using the classical terminology (e.g., reliability, validity, item difficulty, and item discrimination ect.). Besides, they provide ability estimates neither with known statistical properties nor with standard errors of estimate associated with the estimated ability. This problem, as well as the preceding problem, invokes the consideration of using theorectically rigorous scoring models which are shown in modern test theory.

The above description about the assessment of partial knowledge belongs to the area of the classical test theory. Since it does not provide a satisfactory result, it is necessary to find other methods. Latent trait theory (**LTT**) or item response theory (**IRT**) developed in modern test theory is the one that we need (Allen & Yen, 1979; Baker, 1985; Crocker & Algina, 1986; Hambleton, 1983; Hambleton & Swaminathan, 1985; Hulin, Drasgow, & Parsons, 1983; Linn, 1989; Lord, 1980).

Thereafter, several authors (Bock, 1972; Huynh & Casteel, 1987; Jacobs & Vandeventer, 1970; Levine & Drasgow, 1983; Thissen, 1976) considered the use of information in wrong responses to improve the accuracy of ability estimation. Birnbaum's (1968) dichotomous model, providing estimates of ability based on right-wrong scoring of the test items, is a special case of Bock's (1972) general multiple category model that utilizes information in the pattern of wrong responses, as well as correct responses, in estimating ability. Such a model using categories or steps to build the latent trait models can be applied and extended to other ordered response cases, by assigning to each step or category a different weight or parameter, in order to assess the examinees' partial knowledge. Samejima's (1969) graded scores model initiated this kind of relevant research, which was followed by Samejima (1973a), Andersen (1973b, 1973c), Andrich (1978b, 1978d, 1982), Müller (1987), and synthesized directly to Samejima (1969) and expanded directly to Andrich (1978b) by a new term "partial credit model" (Masters, 1982; Wright & Masters, 1982). The other models used for rating scale data and counted events are the constrained versions of the partial credit model (Masters & Wright, 1984; Wright & Masters, 1982). The derivation of the partial credit model is described in detail in the next section.

II. The Formulation of the Partial Credit Model

The Derivation of the Model

The use of ordered performance levels or ordered response steps to indicate examinees' ability measures is a creative device for assessing an examinee's partial knowledge. Th epartial credit model, which requires an examinee's prior identification of several ordered levels of performance on each item and thereby awards partial credit for partial success on items, is a typical one of them. The primary goal for partial credit scoring is the hope that it will lead to a more precise estimate of an examinee's ability than the simple pass/failure or correct/wrong scoring methods, when the problem of partial knowledge is encountered (Masters, 1982; Wright & Masters, 1982).

A partial credit scoring example is shown in Figure 1.

Mathematics Item : $\sqrt{7.5/0.3 - 16} = ?$

Failed (No steps taken)	-----	0
7.5/0.3 = 25	----------	1
25 − 16 = 9	----------	2
$\sqrt{9}$ = 3	----------	3

Performance Levels

	0	1	2	3
7.5 / 0.3 = ?	0 —First step→ 1			
25 − 16 = ?		1 —Second step→ 2		
$\sqrt{9}$ = ?			2 —Third step→ 3	

(Sources : Masters, 1982, p.151 & p.155)

Figure 1. A three-step interpretation of a mathemtics item which using the partial credit scoring.

The numbers 0, 1, 2, and 3 are used to indicate only the ordering of the performance levels, and not used as the traditional "category boundaries" as done in Thurstone's scale model (Edwards & Thurstone, 1952) or in graded response model (Samejima, 1969). From this example, we see that the number of steps into which an item is divided and the relative difficulties of these steps are free to vary from item to item. The performance of an examinee in one of the ordered k + 1 levels 0, 1, 2, .., k on an item can be thought in terms of the last "step" that the examinee has completed or passed.

Recall that the well-known Rasch's dichotomous model (Rasch, 1960/1980) is expressed as

$$\phi_{ni1} = \frac{\exp(\beta_n - b_{i1})}{1 + \exp(\beta_n - b_{i1})} \qquad (2.1.1)$$

where ϕ_{ni1} is person n's probability of scoring 1 *rather than* 0 on item i, β_n is the ability of person n, and b_{i1} is the difficulty of the one step in item i.

For the one-step (i.e., only two performance levels, 0 and 1) item i, it is useful to introduce π_{ni0} as person n's probability of scoring 0 *rather than* 1 on item i, and π_{ni1} as person n's probability of scoring 1 *rather than* 0 on item i. Hence, this additional notation is identical to (2.1.1) since π_{ni1} is simply the probability ϕ_{ni1} of completing the first and only step in item i, and $\pi_{ni0} = 1 - \pi_{ni1}$. However, this notation will be convenient when we discuss more than two ordered performance levels (i.e., items with more than one step). So, (2.1.1) can be rewritten as

$$\phi_{ni1} = \frac{\pi_{ni1}}{\pi_{ni0} + \pi_{ni1}} = \frac{\exp(\beta_n - b_{i1})}{1 + \exp(\beta_n - b_{i1})} \tag{2.1.2}$$

which means that ϕ_{ni1} is person n's probability of scoring 1 *rather than* 0 on item i. Of course, when only two performance levels are possible, $\pi_{ni0} + \pi_{ni1} = 1$ and $\phi_{ni1} = \pi_{ni1}$.

Next, we consider a two-step item i with performance levels 0, 1, and 2. The probability of completing the first step in item i is identical to (2.1.2). However, from now on, $\pi_{ni0} + \pi_{ni1} < 1$ and b_{i1} still governs the probability of completing the first step to score 1 *rather than* 0. The second step from level 1 to level 2 can be taken only if the first step from level 0 to level 1 has been completed. A parallel expression for the probability of completing (or passing) this second step in item i is

$$\phi_{ni2} = \frac{\pi_{ni2}}{\pi_{ni1} + \pi_{ni2}} = \frac{\exp(\beta_n - b_{i2})}{1 + \exp(\beta_n - b_{i2})} \tag{2.1.3}$$

where ϕ_{ni2} is the probability of person n scoring 2 *rather than* 1 on item i, β_n is the same person ability, and b_{i2} is the difficulty of the second step on item i which governs the probability of completing the step from level 1 to level 2. The difficulty of this second step, b_{i2}, governs how likely it is that a person who has already reached level 1 will complete the second step to level 2. Another way of saying this is that a person will make a 2 *rather than* a 1 on item i.

Finally, the expression for the probability of completing the kth step in item i is

$$\phi_{nik} = \frac{\pi_{nik}}{\pi_{nik-1} + \pi_{nik}} = \frac{\exp(\beta_n - b_{ik})}{1 + \exp(\beta_n - b_{ik})} \quad k = 1, 2, \ldots, m_i \tag{2.1.4}$$

where ϕ_{nik} is the probability of person n scoring k *rather than* k − 1 on item i, β_n is the same person ability, and b_{ik} is the difficulty of the kth step on item i which govern the probability of completing the step from level k − 1 to level k.

We can draw a trace line for every step probability, ϕ_{nik}, $k = 1, 2, \ldots, m_i$, against its corresponding ability parameter expressed in the unit of logits, i.e., logit $= \beta_n - b_i$. Such a trace line is called operarting curve (Wright & Masters, 1982, p.42). The operating curves are very similar to item characteristic curves (ICCs). They are simple logistic ogives of the same slope which differ only in their location on the ability continuum. Hence, it is not necessary to draw the operating curves according to their step orders. Since the relative difficulties of these steps may vary from item to item, the positions of the curves fully depend on the step-difficulty orders, not the step-number orders. An example of the operating curves for Figure 1 is shown below.

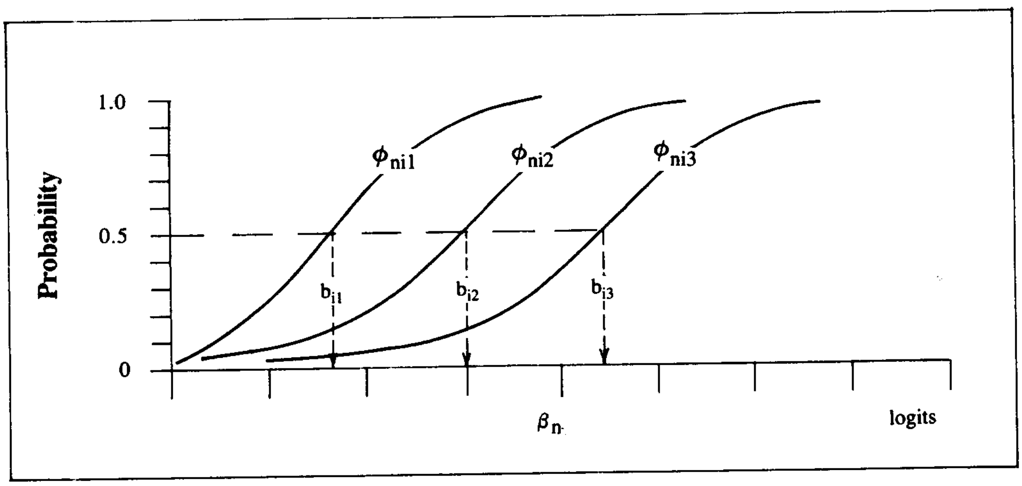

Figure 2. Item operating curves for a three-step item i.

From (2.1.4) and with the requirement that person n must complete one of the $m_i + 1$ possible scores on item i (i.e., $\sum_{k=0}^{m_i} \pi_{nik} = 1$), a general expression for the probability of person n scoring x on item i is

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta_n - b_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij})} \qquad x = 0, 1, \ldots, m_i \qquad (2.1.5)$$

where x is the count of the successfully completed item steps, $b_{i0} = 0$ and

$\sum_{j=0}^{0} (\beta_n - b_{ij}) = 0$, so that $\exp \sum_{j=0}^{0} (\beta_n - b_{ij}) = 1$. The numerator $\exp \sum_{j=0}^{x} (\beta_n - b_{ij})$ contains only the difficulties of these x completed steps, b_{i1}, b_{i2}, ..., b_{ix}. The denominator is the sum of all $m_i + 1$ possible numerators. Such an equation, (2.1.5), is called the partial credit model (PCM) (Masters, 1982; Wright & Masters, 1982).

The partial credit model asserts that the probability of person n scoring x on the m_i-step item i is a function of the person's ability β_n on the test and the difficulties of the m_i "steps" in item i. In like manner, we can also draw a trace line for every score probability, π_{nix}, $x = 0, 1, ..., m_i$, against its corresponding ability parameter expressed in the unit of logits, i.e., logit $= \beta_n - b_i$. Such a trace line is called category probability curve (Wright & Masters, 1982, p.44), item-option characteristic curve (Hambleton & Cook, 1977, p.80), or category characteristic response curve (CCC) (Jansen & Roskam, 1986, p.76). Therein, the item parameter, b_{ik}, is located exactly at the intersection of two category probability curves, $\pi_{i(k-1)}$ and π_{ik}, for level $k - 1$ and level k, $k = 1, 2, ..., m_i$. Due to the varied step difficulties, the location of such an intersection is not necessary to be ordered according to sequential steps in the drawn figure. An example of category probability curves for Figure 1 is shown in Figure 3.

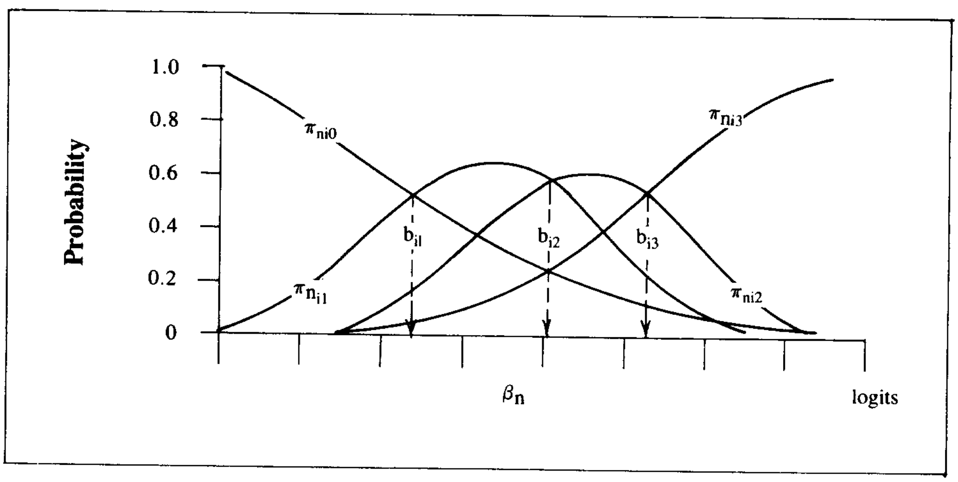

Figure 3. Category probability curves for a three-step item i.

The Estimation of the Model Parameters

Members of the family of the Rasch model share two properties : Parameter separability and specific objectivity (Fischer, 1973; Rasch, 1960, 1977; Whitely & Dawis, 1974; Wright, 1977). Andersen (1973c) showed that there exist sufficient statistics for the person and item parameters in the Rasch model. This means that the measures of person ability in the model can be conditioned out of the estimation equations for the items, i.e., the sample-free item calibration, and the item difficulty estimates in the model also can be conditioned out of the estimation equations for the sample, i.e., the item-free measures of ability (Jansen, & Roskam, 1986). In other words, "the possibility of separating two sets of parameters must be a fundamental property of a very important class of models", which makes "specifically objective" comparisions of persons and items possible (Rasch, 1977, p.66).

Thus, under the partial credit model, the conditional procedure for estimating the model parameters can be shown to be as follows (Douglas, 1978; Masters, 1982, pp.159-161).

The probability of a person n making any particular response vector (x_{ni}) on an L-item test is

$$P\{X_{ni} \mid \beta_n; b_{ij}\} = \prod_{i=1}^{L} \left[\frac{\exp \sum_{j=0}^{X_{ni}} (\beta_n - b_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij})} \right]$$

$$= \frac{\exp \sum_{i=0}^{L} \sum_{j=0}^{X_{ni}} (\beta_n - b_{ij})}{\Psi_n} \qquad (2.2.6)$$

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

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., $r_n = \sum_{i=0}^{L} X_{ni}$, then the probability of person n scoring the score r is

$$P\{r \mid \beta_n; b_{ij}\} = \frac{\sum_{(X_{ni})}^{r} \exp \sum_{i=0}^{L} \sum_{j=0}^{X_{ni}} (\beta_n - b_{ij})}{\Psi_n}$$

$$= \frac{\exp(r\beta_n)}{\Psi_n} \sum_{(X_{ni})}^{r} \exp\left(- \sum_{i=1}^{L} \sum_{j=0}^{X_{ni}} b_{ij}\right) \qquad (2.2.7)$$

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

The conditional probability of the response vector (x_{ni}), given the score r, is obtained by dividing (2.2.6) by (2.2.7)

$$P\{x_{ni} \mid r; \beta_n; b_{ij}\} = \frac{P\{x_{ni} \mid \beta_n; b_{ij}\}}{P\{r \mid \beta_n; b_{ij}\}}$$

$$= \frac{\exp(r\beta_n)\exp(-\sum_{i=1}^{L}\sum_{j=0}^{X_{ni}} b_{ij})}{\exp(r\beta_n)\sum_{(X_{ni})}^{r}\exp(-\sum_{i=1}^{L}\sum_{j=0}^{X_{ni}} b_{ij})}$$

$$= \frac{\exp(-\sum_{i=1}^{L}\sum_{j=0}^{X_{ni}} b_{ij})}{\sum_{(X_{ni})}^{r}\exp(-\sum_{i=1}^{L}\sum_{j=0}^{X_{ni}} b_{ij})} \qquad (2.2.8)$$

Obviously, 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 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 difficulities 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.

Then, the conditional probability of an entire matrix of response (**X**), given the vector of person test scores (r_n), is

$$P\{\mathbf{X} \mid r_n; \beta_n; b_{ij}\} = \prod_{n=1}^{N}\left[\frac{\exp(-\sum_{i=1}^{L}\sum_{j=0}^{X_{ni}} b_{ij})}{\sum_{(X_{ni})}^{r_n}\exp(-\sum_{i=1}^{L}\sum_{j=0}^{X_{ni}} b_{ij})}\right] \qquad (2.2.9)$$

The fact that the person parameter does not appear in (2.2.9) means that the step difficulties can be estimated independently of the abilities of the persons in the calibrating sample.

In like manner, a conditional probability expression containing only the person parameter can be obtained. The probability of observing a particular N-person vector of responses (x_{ni}) to an item i is

$$P\{x_{ni} \mid \beta_n; b_{ij}\} = \prod_{n=1}^{N} \left[\frac{\exp \sum_{j=0}^{x_{ni}} (\beta_n - b_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij})} \right]$$

$$= \frac{\left[\exp\left(\sum_{n=1}^{N} x_{ni}\beta_n\right)\right]\left[\exp\left(-\sum_{n=1}^{N}\sum_{j=0}^{x_{ni}} b_{ij}\right)\right]}{\Psi_i} \qquad (2.2.10)$$

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

If the "score" S_{im} of a particular N-person vector of responses is defined as the count of the total number of persons who have completed the m_ith step on item i, i.e., $S_{im} = \sum_{n=1}^{N} x_{ni}$, then the probability of observing some particular vector of item counts, $\mathbf{S} = (S_{i1}, S_{i2}, \ldots, S_{im})$, for item i is

$$P\{\mathbf{S} \mid \beta_n; b_{ij}\} = \frac{\left[\sum_{(x_{ni})}^{S} \exp\left(\sum_{n=1}^{N} x_{ni}\beta_n\right)\right]\left[\exp\left(-\sum_{n=1}^{N}\sum_{j=0}^{x_{ni}} b_{ij}\right)\right]}{\Psi_i} \qquad (2.2.11)$$

where $\sum_{(x_{ni})}^{S}$ denotes the sum over all those response vectors which produce the item count vector **S**.

Then, the conditional probability of the response vector (x_{ni}), given the vector **S**, can be obtained by dividing (2.2.10) by (2.2.11)

$$P\{x_{ni} \mid \mathbf{S}; \beta_n; b_{ij}\} = \frac{P\{x_{ni} \mid \beta_n; b_{ij}\}}{P\{\mathbf{S} \mid \beta_n; b_{ij}\}}$$

$$= \frac{\exp\left(\sum_{n=1}^{N} x_{ni}\beta_n\right)}{\sum_{(x_{ni})}^{S} \exp\left(\sum_{n=1}^{N} x_{ni}\beta_n\right)}$$

Also obviously, 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 who completed each step in an item. In other words, no further information about the step difficulties can be obtained by keeping track of any other aspect of the performance of individuals, and so, the item counts (S_{ij}) contain all the information available in the data matrix about the step difficulties (b_{ij}).

So, the conditional probability of an entire matrix of responses (**X**), given the vector of item counts (S_{im}), is

$$P\{\mathbf{X} \mid S_{im};\ \beta_n;\ b_{ij}\} = \prod_{i=1}^{L} \left[\frac{\exp\left(\sum_{n=1}^{N} x_{ni}\beta_n\right)}{\sum_{(X_{ni})}^{S_{im}} \exp\left(\sum_{n=1}^{N} x_{ni}\beta_n\right)} \right] \qquad (2.2.13)$$

The implication of (2.2.13) is that the person measures can be estimated independently of the particular of the items used.

From (2.2.8) and (2.2.12), the separability of the parameters in partial credit 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, i.e., r_n, and for an item's step difficulties the sufficient statistics are counts of the number of persons completing each step, i.e., S_{im}. 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 item difficulty.

There is another approach to estimate the model parameters of the partial credit model. This approach is based on Wright & Panchapakesan's (1969) unconditional maximum likelihood estimation for Rasch's dichotomous model. This procedure is simply a method of the evaluation of the maximum likelihood function with the partial credit model π_{nix} integrating across all person n and item i, that is,

$$\Lambda = \prod_{n=1}^{N} \prod_{i=1}^{L} \pi_{nix}$$

$$= \frac{\exp \sum_{n=1}^{N} \sum_{i=1}^{L} \sum_{j=0}^{X_{ni}} (\beta_n - b_{ij})}{\prod_{n=1}^{N} \prod_{i=1}^{\Lambda} \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij}) \right]} \qquad (2.2.14)$$

Taking logarithms,

$$\lambda = \log \Lambda$$
$$= \sum_{n=1}^{N} \sum_{i=1}^{L} x_{ni}\beta_n - \sum_{n=1}^{N} \sum_{i=1}^{L} \sum_{j=1}^{x_{ni}} b_{ij} - \sum_{n=1}^{N} \sum_{i=1}^{L} \left[\log \sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij}) \right] \quad (2.2.15)$$

in which which $\sum_{j=0}^{x_{ni}} b_{ij} = \sum_{j=1}^{x_{ni}} b_{ij}$ because b_{i0} is assumed to be zero.

Then, as specified above, define the person n's "score" on the L-item test as a count of the total number of completed item steps, i.e. $r_n = \sum_{i=1}^{L} x_{ni}$. And define the number of persons completing step j on item i as S_{ij}, so the sum across all N persons can be rewritten as $\sum_{n=1}^{N} \sum_{j=0}^{x_{ij}} b_{ij} = \sum_{j=1}^{m_i} S_{ij}b_{ij}$. This means that 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 weighting the step difficulties by these counts to form the sum $S_{i1}b_{i1} + S_{i2}b_{i2} + \ldots + S_{im}b_{im}$. With these simplifications the log likelihood function, (2.2.15), becomes

$$\lambda = \sum_{n=1}^{N} r_n\beta_n - \sum_{i=1}^{L} \sum_{j=1}^{m_i} S_{ij}b_{ij} - \sum_{n=1}^{N} \sum_{i=1}^{L} \log \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij}) \right] \quad (2.2.16)$$

For simplification, the latter part of (2.2.16) is taking the first derivatives with respect to β_n and b_{ij} as

$$\frac{\partial \log \left[\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij}) \right]}{\partial \beta_n} = \frac{\sum_{k=0}^{m_i} k \exp \sum_{j=0}^{k} (\beta_n - b_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - b_{ij})}$$
$$= \sum_{k=0}^{m_i} k\pi_{nik} = \sum_{k=1}^{m_i} k\pi_{nik}$$

and

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

$$= - \sum_{k=j}^{m_i} \pi_{nik}$$

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

Then, taking the first derivatives of (2.2.16) with respect to β_n and b_{ij} and setting those equations to be zero, (2.2.16) becomes

$$\frac{\partial \lambda}{\partial \beta_n} = r_n - \sum_{i=1}^{L} \sum_{i=1}^{m_i} k\pi_{nik} \qquad n=1, N \qquad (2.2.17)$$

and

$$\hat{\beta}_n = r_n = \sum_{i=1}^{L} \sum_{k=1}^{m_i} k\pi_{nik} \qquad (2.2.18)$$

where $\hat{\beta}_n$ is the maximum likelihood estimate for the person ability parameter, β_n, which 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).

$$\frac{\partial \lambda}{\partial b_{ij}} = - S_{ij} + \sum_{n=1}^{N} \sum_{k=j}^{m_i} \pi_{nik} \qquad i=1, L; j=1, m_i \qquad (2.2.19)$$

and

$$\hat{b}_{ij} = S_{ij} = \sum_{n=1}^{N} \sum_{k=j}^{m_i} \pi_{nik} \qquad (2.2.20)$$

where $\hat{b}_{ij}$ is the maximum likelihood estimate for the item difficulty parameter, b_{ij}. Meanwhile, the $\sum_{k=j}^{m_i} \pi_{nik}$ 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 (2.2.16) with respect to β_n and b_{ij} are as follows.

$$\frac{\partial^2\lambda}{\partial\beta_n^2} = -\sum_{i=1}^{L}\left[\sum_{k=1}^{m_i} k^2\pi_{nik} - \left(\sum_{k=1}^{m_i} k\pi_{nik}\right)^2\right] \quad (2.2.21)$$

$$\frac{\partial^2\lambda}{\partial b^2_{ij}} = -\sum_{n=1}^{N}\left[\sum_{k=j}^{m_i} \pi_{nik} - \left(\sum_{k=j}^{m_i} \pi_{nik}\right)^2\right] \quad (2.2.22)$$

(2.2.21) and (2.2.22) can be solved by the Newton-Raphson iterative procedure. The person and item parameters are estimated by

$$\hat{\delta}_n^{t+1} = \hat{\delta}_n^{t} - \frac{r_n - \sum_{i=1}^{L}\sum_{K=1}^{m_i} kP^t}{-\sum_{i=1}^{L}\left[\sum_{k=1}^{m_i} k^2P^t - \left(\sum_{k=1}^{m_i} kP^t\right)^2\right]} \qquad n=1,\ N \quad (2.2.23)$$

$$\hat{d}_{ij}^{t+1} = \hat{d}^t_{ij} - \frac{-S_{ij} + \sum_{i=1}^{N}\sum_{K=1}^{m_i} kP^t}{-\sum_{n=1}^{N}\left[\sum_{k=j}^{m_i} P^t - \left(\sum_{k=j}^{m_i} P^t\right)^2\right]} \qquad i=1,\ L;\ j=1,\ m_i \quad (2.2.24)$$

where $\hat{\delta}_n^t$ is the estimate of β_n after t iterations, $\hat{d}^t_{ij}$ is the estimate of b_{ij} after t iterations, and P^t is the estimated probability of person n responding in step k to item i after t iterations.

For convenience and avoiding the indeterminancy in the scale origin, the mean step difficulty d.. is set equal to zero. And the asymptotic estimates of the standard errors are calculated from the last iteration by

$$SE\ (\hat{\delta}_n) = \left[\sum_{i=1}^{L}\left(\sum_{k=1}^{m_i} k^2P - \left(\sum_{k=1}^{m_i} kP\right)^2\right)\right]^{-1/2} \quad (2.2.25)$$

$$SE\ (\hat{d}_{ij}) = \left[\sum_{n=1}^{N}\left(\sum_{k=j}^{m_i} P - \left(\sum_{k=j}^{m_i} P\right)^2\right)\right]^{-1/2} \quad (2.2.26)$$

Although the unconditional maximum likelihood (UML) procedure is said to be superior to the conditional maximum likelihood (CML) (Wright & Panchapakesan, 1969), the UML has been proven to be inconsistent (Andersen, 1973b, 1973c) and the slight bias can be corrected by multiplying the term (L − 1)/L, where L is the number of items. Thereafter, the simulation research of Wright & Douglas (1977a, 1977b) have corroborated the "correctness" of the correction term. Gustafsson (1980b) also had shown that the factor (L − 1)/L seemed to work satisfactorily for other numbers of items as well. Unfortunately, this problem is not so easy. Van den Wollenberg, Wierda, & Jansen (1988) and Jansen, van den Wollenberg, & Wierda (1988) replicate the "bias correction procedure" by a simulation study and find that this bias cannot be removed by the correction factor $(L - 1)/L$. The bias is dependent not only on the number of items, but also on the distribution of the item parameters, which makes correcting for bias practically impossible. But in the intial studies with data simulated to fit the partial credit model, it is suggested that this same correction may be appropriate for removing bias in item step estimates when $m_i > 1$ (Masters, 1982).

The Parameter Fit Statistics

When data fit the partial credit model, the fit index statistic of item i is calculated from

$$t_i = (v_i^{1/3} - 1)(3/q_i) + q_i/3 \tag{2.3.27}$$

where

$$v_i = \frac{\sum_{n=1}^{N} (x_{ni} - E_{ni})^2}{\sum_{n=1}^{N} W_{ni}} \tag{2.3.28}$$

is distributed as a mean square with expected value one and expected variance

$$q_i^2 = \frac{\sum_{n=1}^{N} (C_{ni} - W^2_{ni})}{(\sum_{n=1}^{N} W_{ni})^2} \tag{2.3.29}$$

and

$$E_{ni} = \sum_{k=0}^{m_i} kP_{nik} \tag{2.3.30}$$

$$W_{ni} = \sum_{k=0}^{m_i} (k - E_{ni})^2 P_{nik} \tag{2.3.31}$$

$$C_{ni} = \sum_{k=0}^{m_i} (k - E_{ni})^4 P_{nik} \tag{2.3.32}$$

which are the estimates of the expected value, the expected variance, and the expected kurtosis, respectively, of each examinee's score on item i. And, P_{nik} is the estimated probability of person n scoring k on item i.

The t_i has a mean near zero and a standard deviation near one. The larger the absolute value of t_i, the more serious the misfit of item i in a test. Thus, t_i is used as an index of item parameter fit.

In like manner, the person parameter fit can be calculated from

$$t_n = (v_n^{1/3} - 1)(3/q_n) + (q_n/3) \tag{2.3.33}$$

where

$$v_n = \frac{\sum_{i=1}^{L} (x_{ni} - E_{ni})^2}{\sum_{i=1}^{L} W_{ni}} \tag{2.3.34}$$

is the weighted mean square with expected value one and variance

$$q_n^2 = \frac{\sum_{i=1}^{L} (C_{ni} - W^2_{ni})}{(\sum_{i=1}^{L} W_{ni})^2} \tag{2.3.35}$$

and the expected mean mean, variance, and kurtosis, respectively, are

$$E_{ni} = \sum_{k=0}^{m_i} kP_{nik} \tag{2.3.36}$$

$$W_{ni} = \sum_{k=0}^{m_i} (k - E_{ni})^2 P_{nik} \tag{2.3.37}$$

$$C_{ni} = \sum_{k=0}^{m_i} (k - E_{ni})^4 P_{nik} \tag{2.3.38}$$

in which P_{nik} is the estimated probability of examinee n scoring k on item i.

The t_n has an expectation near zero and variance near one when the model holds. The larger the absolute value of t_n, the more serious the misfit of the estimate of person n's ability in a sample group. Thus t_n is also used as an index of person parameter fit.

The only difference between t_i and t_n is that squared residuals, i.e., $(x_{ni} - E_{ni})^2$, are summed over persons for item i and summed over items for person n. Item fit statistics play an important role in the construction and calibration of an instrument. Person fit statistics are useful for assessing the validity of measures made with instruments which have already been established (Wright & Masters, 1982).

III. An Empirical Example

To show how the person and item parameters are calculated, a pseudo data matrix is simulated for 40 persons' responses on 10 items which are scored on the three-step fromat like Figure 1.[1] Although the data matrix can be constructed by the CREDIT computer program (Wright, Masters, & Ludlow, 1982), the easier PROX procedure (Wright & Stone, 1979) is used in this example.

The PROX procedure assumes that person and item parameters are more or less normally distributed. Its advantages are that it can be done by a hand calculator and that it satisfies most of the principles underlying Rasch calibration and measurement. All that PROX needs for person- and item-parameter estimation are the person scores and item scores from Table 1. The calibrations for the 10 items are shown in Table 2. And the measures for the 40 persons are shown in Table 3. Because the calculations of the mean square v_i, error term q_i, and fit statistic t_i for item parameters, and v_n, q_n, and t_n for person parameters by a hand calculator are very tedious, time-consuming, and easily susceptible to rounding error, they are not shown in Table 2 and Table 3. If one can access the CREDIT program, one can calculate the fit statistics for item and person parameters and decide which parameters lack fit. Since the visual checking of the estimated parameters does not show large differences, the fit statistics may not indicate those that lack fit. For a detailed discussion of the calculations by the PROX procedure, it is suggested that the reader refers to Wright & Masters (1982, Ch. 4).

IV. Comments on the Partial Credit Model

The partial credit model is formulated as an alternative to Andrich's Rating Scale Model (Andrich, 1978a, 1978b, 1978c, 1978d, 1979) for situations in which ordered response choices are free to vary in number and difficulty from item to item. The primary goal of the partial credit model is to more precisely estimate an examinee's ability by assessing his/her partial knowledge on the wrong-responses pattern in a given test. Thus, the application of the partial credit model is restricted to tests or questionnaires that are constructed with an ordered response format. Two examples — one for building a "fear-of-crime" variable, the other for assessing the performance of pre-kindergarten children using the partial credit model — are shown

Table 1. A raw data matrix scored on the three-step format.

Person Number	Item	1 2 3 4 5 6 7 8 9 10	Person Scroe	Step freq.
	Step	0123012301230123012301230123012301230123		0123
1		1100111111101100110011001111111001100111	17	0613
2		1111111111001110100011001110100011001111	16	2323
3		1110100011001000100011001110111111111000	12	4222
4		1110110011101000100011001110111111001000	12	3331
5		1100110011111000100010001110111111001100	12	3412
6		1111111111111000100011111111111110001111	21	3007
7		1111111111101000100011111110111111111111	22	2026
8		1100110011111100110011001100111010001100	12	1711
9		1100111011111111100011101100111011101100	17	1342
10		11101111100011101100111111110100011101100	16	2242
11		1111111011001111100011001100111011001000	14	2422
12		1111100111111001000100010001100100011000	10	4402
13		1111100110011001100110010001111111111000	14	2503
14		1111111111001100110011001111100011001100	15	1603
15		1100110011101000100011001100100011001111	10	3511
16		1110111011111111111011101000111010001100	17	2152
17		1110100010001100111011111100111010001111	14	3232
18		1100110011111111100010001100111011101111	17	2323
19		1111111111111111110011001110111011101110	22	0244
20		1110100011001111110011001100111011101111	16	1432
21		1111100110011001100111011111111011111100	18	0523
22		1100100011111111111011101100111011001100	16	1432
23		1110100011101000110011101000111010001110	11	4150
24		1000110010001000111011001110111011111111	14	3232
25		1100110011001100110011001110110011111100	13	0811
26		1110111111001111111011001111111011111111	23	0235
27		1000111011001111100011101100111111001111	16	2332
28		1111111111001100110011001110100011001100	14	1612
29		1111111011111110110011111110110010001111	20	1234
30		1100110011001100110011001110110010001100	10	1810
31		1111111111101000100010001000100011001111	12	5113
32		1000110011001100100011001100100011111111	11	3502
33		1110111010001110111111001110100011001100	14	2341
34		1100111010001100110011001000110011001100	9	2710
35		1111110010001000100010001111111010001110	11	5122
36		1111000110011001110111111001000111011111	17	2323
37		11111100111011111110011001100111011001100	12	0622
38		1110111011101100110011001110111011101100	14	1450
39		1110110011001100110011001110111011101000	13	1540
40		1100100011001000110011001110111011111100	12	2521
Item Score		78 61 62 53 33 53 62 63 55 68		
Step Freq. 0		3 7 6 11 15 5 6 9 10 5		
1		11 15 16 15 18 23 12 6 14 17		
2		11 8 8 4 6 6 16 18 7 3		
3		15 10 10 10 1 6 6 7 9 15		

Table 2. The PROX calibrations for 10 items

$m_i=3$, $L=10$, $N=40$

Item Number	Item Scores	Step Scores			Proportions			Step Logits			Logit Squared			Intial Step Calibrations			Sample Spread Expansion	Final Step Estimates			Step Errors		
i	S_i	S_{ij}			P_{ij}			X_{ij}			X^2_{ij}			d°_{ij}			Y	$d_{ij}=Yd^\circ_{ij}$			$SE(\hat{d}_{ij})$		
5	33	18	6	1	.450	.150	.025	.20	1.73	3.66	.04	2.99	13.40	−.94	.59	2.52	1.07	−1.01	.63	2.70	.34	.47	1.08
4	53	15	4	10	.375	.100	.250	.51	2.20	1.10	.26	4.84	1.21	−.63	1.06	−.04	1.07	−.67	−.13	−.04	.35	.56	.39
6	53	23	6	6	.575	.150	.150	−.30	1.73	1.73	.09	2.99	2.99	−1.44	.59	.59	1.07	−.54	.63	.63	.34	.47	.47
9	55	14	7	9	.350	.175	.225	.62	1.55	1.24	.38	2.40	1.54	−.52	.41	.10	1.07	−.56	.44	.11	.35	.45	.47
2	61	15	8	10	.375	.200	.250	.51	1.39	1.10	.26	1.93	1.21	−.63	.25	−.04	1.07	−.67	.27	−.04	.35	.42	.39
3	62	16	8	10	.400	.200	.250	.41	1.39	1.10	.17	1.93	1.21	−.73	.25	−.04	1.07	−.78	.27	−.04	.35	.42	.39
7	62	12	16	6	.300	.400	.150	.85	.41	1.73	.72	.17	2.99	−.29	−.73	.59	1.07	−.31	−.78	.63	.37	.35	.47
8	63	6	18	7	.150	.450	.175	1.73	.20	1.55	2.99	.04	2.40	.59	−.94	.41	1.07	.63	−1.01	.44	.47	.34	.45
10	68	17	3	15	.425	.075	.375	.30	2.51	.51	.09	6.30	.26	−.84	1.37	−.63	1.07	−.90	1.47	−.67	.34	.64	.35
1	78	11	11	15	.275	.275	.375	.97	.97	.51	.94	.94	.26	−.17	−.17	−.63	1.07	−.18	−.18	−.67	.38	.38	.35

Sum = 34.11　Sum = 57.94
Mean = 1.14　Variance = 1.55

$$P_{ij} = S_{ij}/N$$

$$X_{ij} = \log [(1-P_{ij})/P_{ij}] = \log [(N - S_{ij})/S_{ij}]$$

$$X.. = \sum_i^L \sum_i^{m_i} X_{ij}/M$$

$$M = \sum_i^L m_i \ (=30)$$

$$X^2.. = 1.30$$

$$d^0_{ij} = x_{ij} - x..$$

$$U = \frac{\sum_i^L \sum_j^{m_i} (x_{ij} - x..)^2}{M - 1} = \frac{\sum_i^L \sum_j^{m_i} x^2_{ij} - Lx^2_{ij}}{M - 1} = 1.55$$

$$Y = [\frac{1 + V/2.89}{1 - UV/8.35}]^{1/2} = [\frac{1 + .25/2.89}{1 - (1.55)(.25)/8.35}]^{1/2} = 1.07$$

* V values are shown in Table 3.

$$SE(\hat{d}_{ij}) = Y[\frac{1}{NP_{ij}(1 - P_{ij})}]^{1/2} = Y[\frac{N}{(N - S_{ij})(S_{ij})}]^{1/2}$$

Table 3. The PROX measures for 40 persons

$m=3,\ L=10,\ N=40$

Person Score r	Person Count N_r	Score Logit Y_r	Counted Logit N_rY_r	Counted Logit Squared $N_rY_r^2$	Initial Measure $\delta_r^0=Y_r$	Test Spread Expansion X	Final Ability Estimate $\delta_r=X\delta_r^0$	Ability Error $SE(\hat{\delta}_r)$
23	1	1.19	1.19	1.42	1.19	1.27	1.51	.55
22	2	1.01	2.02	2.04	1.01	1.27	1.28	.52
21	1	.85	.85	.72	.85	1.27	1.08	.51
20	1	.69	.69	.48	.69	1.27	.88	.49
19		.55			.55	1.27	.70	.48
18	1	.41	.41	.17	.41	1.27	.52	.47
17	5	.27	1.35	.36	.27	1.27	.34	.47
16	5	.13	.65	.08	.13	1.27	.17	.46
15	1	.00	.00	.00	.00	1.27	.00	.46
14	7	−.13	−.91	.12	−.13	1.27	−.17	.46
13	2	−.27	−.54	.15	−.27	1.27	−.34	.47
12	7	−.41	−2.87	1.18	−.41	1.27	−.52	.47
11	3	−.55	−1.65	.91	−.55	1.27	−.70	.48
10	3	−.69	−2.07	1.43	−.69	1.27	−.88	.49
9	1	−.85	−.85	.72	−.85	1.27	−1.08	.51
		Sum =	−1.7	9.78				
		Mean =	−.043			Variance = .25		

$P_r = r/mL$

$Y_r = \log[P_r/(1 - P_r)] = \log[r/(M - r)]$

$Y. = \sum_r^{M-1} N_rY_r/N \qquad Y.^2 = (Y.)^2 = .002$

$\delta_r^0 = Y_r$

$$V = \frac{\sum_r^{M-1} N_r(Y_r - Y.)^2}{N - 1} = \frac{\sum_r^{M-1} N_rY_r^2 - NY.^2}{N - 1} = \frac{9.78 - 40\ (.002)}{40 - 1} = .25$$

$$X = \left[\frac{1 + U/2.89}{1 - UV/8.35}\right]^{1/2} = \left[\frac{1 + (1.55)/2.89}{1 - (1.55)(.25)/8.35}\right]^{1/2} = 1.27$$

* U values are shown in Table 2.

$SE(\hat{\delta}_r) = X\,[1/mLP_r\,(1 - P_r)]^{1/2} = X\,[M/r(M-r)]^{1/2}$

in Masters (1982), Masters & Wright (1982), and Wright & Masters (1982). Other examples for banking test items which use the partial credit scoring method to equate the test forms with ordered response choices are illustrated in Masters (1984) and Masters & Evans (1986). Besides, Smith (1987) shows that the results of assessing partial knowledge in vocabulary support O'Connor's theory of vocabulary acquisition. Dodd & Koch (1987) indicate that the usefulness of item and test information in the partial credit model is not restricted to item or test selection, but is also useful in actual construction of test items.

From such examples, we can induce three important features of the partial credit model as described below (Masters, 1982, p.172).

First, the category probability curve for every performance level is fixed by the difficulties of the item's subtasks, and so it can vary from item to item. Checking on the scoring strategies, every item's pattern can be easily understood.

Second, when a test item is constructed with more than two performance levels, a latent ability variable can be best understood in terms of the concept of "steps" between adjacent performance levels. Such a concept provides an opportunity to identify ill-constructed item steps which could be revised to make the item more informative.

Third, it allows the possibility of detecting an item's misfit to a particular step in that item. Such a detection can distinguish inappropriate item from inappropriate respondent groups.

Although there is no evidence showing weaknesses in using the partial credit model, there are some papers indicating that the Rasch model is not overall superior to other models. For example, Divgi (1986) strongly objects to several properties of the Rasch model, criticizes its inappropriateness under several conditions, and concludes that the Rasch model is not suitable for multiple-choice items.

In a comparision of model fits, Albanese & Forsyth (1984) show that the Rasch model fails to fit more items than does the two-parameter logistic model. Hambleton & Traub (1973) found that the two-parameter model predicted score distributions better than the Rasch model did. Goldman & Raju (1986), Waller (1981), and Yen (1981) reported that the two-parameter model fitted attitude surveys better than the Rasch model. Future applications may come to favor the use of the two-parameter model. Andersen (1973a) also found that the Rasch model did not fit the verbal part of SAT, and attributed the lack of fit to unequal item discriminations.

Since the partial credit model as currently formulated is based on the Rasch model, does it suffer from such weaknesses as criticized above? The answer to this question is still unknown to us. The present author tries to generalize the one-parameter partial credit model to a two-parameter partial credit model in his recent proposal (Yu, 1991). Perhaps, taking a "step discrimination" parameter into consideration in the partial credit model will improve the model fit and make it suitable for multiple-choice items. It may be useful in mastery testing to discriminate between mastery and nonmastery groups. So far, these kinds of research are under way. We hope the day of broad uses of the partial credit model in a variety of fields to assess and improve measurement problems more precisely will come soon.

V. Summary and Conclusion

The traditional number-right scoring method cannot provide a satisfactory estimate of an examiness's ability measures, because it cannot handle problems such as guessing and partial knowledge. Several remedies have been proposed to compensate for the defect of the traditional scoring method. Unfortunately, however, these remedial approaches cannot consistently and convincingly improve the estimates of examinees' ability measures. Thus, the pursuit of a theoretically rigorous and more precisely estimating method is necessary. Masters' partial credit model is apparently the one that we need.

The partial credit model was invented by Geoff N. Masters to score examinees' partial knowledge. His model uses the Rasch-type logistic latent trait model to score the ordered response items. It assumes that responses to the adjacent item steps can reflect examinees' different knowledge levels, and awards partial credits for examinees' partial knowledge on an item. Hence, it can improve the precision of estimation of examinees' ability measures.

An easier procedure is illustrated to show how to calibrate step difficulties and to measure person abilities. For a large sample and a longish test, it is suggested that the computer program CREDIT be used. Several applications in practical measurement problems are reported. The strengths of the partial credit model are also summarized. The weaknesses of the Rasch model are indicated and the possible direction of future research is briefly discussed too.

Acknowledgement

The author owes his thanks to Dr. Maurice M. Tatsuoka for helpful comments and suggestions on the first draft.

References

Albanese, M. A., & Forsyth, R. A. (1984). The one-, two- and modified two-parameter latent trait models: An empirical study of relative fit. *Educational and Psychological Measurement, 44,* 229-246.

Allen, M. J., & Yen, W. M. (1979). *Introduction to Measurement Theory.* Monterey, CA: Brooks/Cole.

Andersen, E. B. (1973a). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123-140.

Andersen, E. B. (1973b). Conditional inference for multiple-choice questionnaires. *British Journal of Mathematical and Statistical Psychology, 26,* 31-44.

Andersen, E. B. (1973c). *Conditional Inference and Models for Measuring.* Copenhagen: Mentalhygiejnisk Forlag.

Anderesen, E. B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications.* Amsterdam: North-Holland.

Andrich, D. (1978a). A binomial latent trait model for the study of Likert-style attitude questionnaires. *British Journal of Mathematical and Statistical Psychology, 31,* 84-98.

Andrich, D. (1978b). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1978c). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

Andrich, D. (1978d). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2,* 581-594.

Andrich, D. (1979). A model for contigency tables having an ordered response classification. *Biometrics, 35,* 403-415.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47,* 105-113.

Baker, F. B. (1985). *The Basis of Item Response Theory.* Portsmouth, N.H.: Heinemann.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick, *Statistical Theories of Mental Test Scores* (Part V, pp. 397-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

Commbs, C. H., Milholland, J. E., & Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16,* 13-37.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory.* New York: Holt, Rinehart and Winston.

Cureton, E. E. (1966). The correction for guessing. *The Journal of Experimental Education, 4,* 44-47.

Notes:

1. This data matrix is invented only for illustrative purposes, not for real measurement situations. For a real situation of item calibrations and ability measures, it is suggested that larger samples and more items be used.

Davis, F. B. (1959). Estimation and use of scoring weight for each choice in multiple-choice test items. *Educational and Pschological Measurement, 19,* 291-298.

Davis, F. B. (1967). A note on the correction for choice success. *The Journal of Experimental Education, 5,* 43-47.

Davis, F. B., & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement, 19,* 159-170.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology, 18,* 87-123.

Diamond, J., & Evans, W. (1973). The correction for guessing. *Review of Educational Ressearch, 43,* 181-191.

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement, 23,* 283-298.

Dodd, B. G., & Koch, W. R. (1987). Effects of variations in item step values on item and test information in the partial credit model. *Applied Psychological Measurement, 11,* 371-384.

Douglas, G. A. (1978). Conditional maximum-likelihood estimation for a multiplicative binomial response model. *British Journal of Mathematical and Statistical Psychology, 31,* 73-83.

Dressel, P. L., & Schmid, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13,* 574-595.

Edwards, A. L., & Thurstone, L. L. (1952). An internal consistency check for scale values determined by the method of successive intervals. *Psychometrika, 17,* 169-180

Fischer, G. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Forsyth, R., Saisangjan, U., & Gilmer, J. (1981). Some empirical results related to the robustness of the Rasch model. *Applied Psychological Measurement, 5,* 175-186.

Glass, G. V., & Wiley, D. E. (1964). Formula scoring and test reliability. *Journal of Educational Measurement, 1,* 43-47.

Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two- parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement, 46,* 11-21.

Gustafsson, J. E. (1980a). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40,* 377-385.

Gustafsson, J. E. (1980b). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33,* 205-233.

Hambleton, R. K. (Ed.) (1983). *Applications of Item Response Theory.* Vancouver, British Columbia: Educational Research Institute of British Columbia.

Hambleton, R. K., & Cook, L. L. (1977). Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement, 14,* 75-96.

Hambleton, R. K., Roberts, D. M., & Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement, 7,* 75-82.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications.* Boston, MA: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. (1978).

Developments in latent trait theory: Models, technical issues, and applications. *Review of Educational Research, 48,* 476-510.

Hambleton, R. K., & Traub, R. E. (1973). Analysis of empirical data using two logistic latent trait models. *British Journal of Mathematical and Statistical Psychology, 26,* 195-211.

Hambleton, R. K., & van der Linden, W. J. (1982). Advances in item response theory and applications: An introduction. *Applied Psychological Measurement, 6,* 373-378.

Hendrickson, G. F. (1971). The effect of differential option weighting on multiple-choice objective tests. *Journal of Educational Measurement, 8;* 291-296.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item Response Theory: Application to Psychological Measurement.* Homewood, IL: Dow Jones-Irwin.

Huynh, H., & Casteel, J. (1987). The useful of the Bock model for scoring with information from incorrect responses. *Journal of Experimental Education, 55,* 131-136.

Jackson, R. A. (1955). Guessing and test performance. *Educational and Psychological Measurement, 15,* 74-79.

Jacobs, P., & Vandeventer, M. (1970). Information in wrong responses. *Psychological Reports, 26,* 311-315.

Jansen, P. G. W., & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51,* 69-91.

Jansen, P. G. W., van der Wollenberg, A. L., & Wierda, F. W. (1988). Correction unconditional parameter estimates in the Rasch model for inconsistency. *Applied Psychological Measurement, 12,* 297-306.

Levine, M. V., & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43,* 675-685.

Linn, R. L. (Ed.) (1989). *Educational Measurement* (3rd ed.). New York: Macmillan.

Lord, F. M. (1963). Formula scoring and validity. *Educational and Psychological Measurement, 23,* 663-672.

Lord, F. M. (1964). The effect of random guessing on test validity. *Educational and Psychological Measurement, 24,* 745-747.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12,* 7-11.

Lord, F. M. (1980). *Application of Item Response Theory to Pratical Testing Problems.* Hillsdale, N J: Lawrence Erlbaum Associates.

Little, E. B. (1962). Overcorrection for guessing in multiple-choice test scoring. *Journal of Educational Measurement, 55,* 245-252.

Lyerly, S. B. (1951). A note on correcting for chance success in objective tests. *Psychometrika, 16,* 21-30.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

Masters, G. N. (1984). Constructing an item bank using partial credit scoring. *Journal of Educational Measurement, 21,* 19-32.

Masters, G. N., & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement, 10,* 355-367.

Masters, G. N., & Wright, B. D. (1982). Defining a 'fear-of-crime' variable: A comparison of two Rasch models. *Education Research and Perspectives, 9,* 18-31.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49,* 529-544.

Müller, H. (1987). A Rasch model for continuous ratings. *Psychometrika, 52,* 165-181.
Patnaik, D., & Traub, R. E. (1973). Differential weighting by judged degree of correctness. *Journal of Educational Measurement, 10,* 281-286.
Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: The University Chicago Press. (Reprinted in 1980).
Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy, 14,* 58-94.
Reilly, R. R., & Jackson, R. (1973). Effects of empirical option weighting on reliability and validity of an academic aptitude test. *Journal of Educational Measurement, 10,* 185-194.
Rippey, R. (1968). Probabilistic testing. *Journal of Educationl Measurement, 5,* 211-215.
Sabers, D. L., & White, G. W. (1969). The effect of differential weighting of individual item reponses on the predictive validity and reliability of an aptitude test. *Journal of Educational Measurement, 6,* 93-96.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, No. 17.*
Samejima, F. (1973a). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.
Samejima, F. (1973b). A comment on Birnbaum's three-parameter logistic model in the latent trait theory. *Psychometrika, 38,* 221-233.
Sax, G., & Collet, L. (1968). The effects of differing instructions and guessing formulas on reliability and validity. *Educational and Psychological Measurement, 28,* 1127-1136.
Shuford, E. H., Albert, A., & Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrika, 31,* 125-145.
Smith, R. M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement, 24,* 217-231.
Stanley, J. C., & Wang, M. D. (1968). *Differential Weighting: A Survery of Methods and Empirical Studies.* New York: College Entrance Examination Board.
Thissen, D. M. (1976). Information in wrong responses to Raven progressive matrices. *Journal of Educational Measurement, 13,* 201-214.
Traub, R. E., & Lam, Y. R. (1985). Latent structure and item sampling models for testing. *Annual Review of Psychology, 36,* 19-48.
van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1988). Consistency of Rasch model parameter estimation: A simulation study. *Applied Psychological Measurement, 12,* 307-313.
Wang, M. D., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40,* 663-705.
Waller, M. I. (1981). A procedure for comparing logistic latent trait models. *Jouranl of Educational Measurement, 18,* 119-125.
Whitely, S. E., & Dawis, R. V. (1974). The nature of objectivity with the Rasch model. *Journal of Educational Measurement, 11,* 163-178.
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14,* 97-116.
Wright, B. D., & Douglas, G. A. (1977a). Conditional versus unconditional procedures for sample-free analysis. *Educational and Psychological Measurement, 37,* 47-60.
Wright, B. D., & Douglas, G. A. (1977b). Best procedures for sample-free item analysis.

Applied Psychological Measurement, 1, 281-295.

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

Wright, B. D., Masters, G. N., & Ludlow, L. H. (1982). *CREDIT: A Computer Program for the Rasch Polychotomous Model.* Chicago: University of Chicago, Department of Education, Statistical Laboratory.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

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

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

Yu, M. (1991). A two-parameter partial credit model. *Doctoral dissertation of University of Illinois at Urbana-Champaign* (unpublished).