# A TWO-PARAMETER PARTIAL CREDIT MODEL FOR THE ORDERED-RESPONSE DATA

Min-Ning Yu

余 民 寧*

## 摘　要

本文旨在提出步驟鑑別度以改進Masters於1982年所提的「（一個參數）部份計分模式」，而成爲「兩個參數部份計分模式」，並探索它的存在可能性。本文作者自行設計的電腦程式，TPPCM，使用最大近似值估計法去估計模式參數與能力參數，並驗証各種參數之適合度，及提供訊息函數，並舉例說明與討論，獲四項結論：㈠兩個參數部份計分模式的存在性獲得肯定；㈡步驟鑑別度能提供區別受試者表現水準之參考；㈢步驟訊息函數獨由每個試題的步驟鑑別度所決定，對試題的選擇與試卷的編輯很有貢獻；㈣除了「明確客觀性」與「參數可分離性」外，兩個參數部份計分模式與一個參數部份計分模式分享共同的基本假設與特性。

## Abstracts

The purpose of this paper is to expand Masters' (1982) (one-parameter) partial credit model to be a two-parameter partial credit model. A FORTRAN 77 computer program, TPPCM, developed by present author, which uses maximum likelihood estimation solutions, is used to calibrate model parameters, test goodness-of-fit, and provide information functions of a dataset. From the analysis and discussion of findings of this research, four conclusions cam be drawn as follows: (a) The existence of the two-parameter partial credit model is confirmed. This model becomes an alternative model to score persons' partial knowledge or calibrate any questionnaire or test with ordered-response formats. (b) Step discriminations provide a good help in partitioning persons' performance levels. (c) Step information functions are uniquely and differently determined from step discriminations of each item. Hence it implies potentials for item and test design, selection, and construction. (d) The two-parameter partial credit model shares the same features of the one-parameter partial credit model, except that of specific objectivity and parameter separability.

*作者為本校教育系副教授兼任實驗小學校長

## I. Introduction

The number-right scoring method has traditionally been used to score an examinee's performance of teacher-made exams or standard achievement tests. Whatever types of answer keys (for example, dichotomous or multiple-choice item) are used, the number-right scoring method used to evaluate students' or examinees' achievement seems to be a natural thing. No one will doubt its appropriateness. Unforunately, this scoring method misses much information about examinees' real-ability estimates. One of its greatest defects is that it ignores the existence and importance of partial knowledge. That is, this scoring method cannot distinguish examinees who have no knowledge about an item from those who have partial knowledge.

According to learning theory, the learning process progresses little by little. The acquisition of knowledge is a cumulative and continuous, not an all-or-none, status (Gagné, 1962; Gagné & Paradise, 1961; Ludlow & Hillocks, 1985). Therefore, partial knowledge exists. It represents the partial results of teachers' instruction and students' learning. Measures of partial knowledge can provide a lot of information and improve the precision of estimates of examinees' real abilities.

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 precise estimating method to be used. Consequently, several alternative scoring methods have been proposed to compensate for drawbacks of the number-right scoring method.

The common use of remedies for the 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). This formula score 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 formula score cannot give us any information about examinees' ability measures which are intermediate in scoring correct and scoring wrong items.

Some alternatives to the formula score method are the use of differential weighting schemes (Davis & Fifer, 1959; Hambleton, Roberts, & Traub, 1970; Hendrickson, 1971; Patanik & Traub, 1973; Reilly & Jackson, 1973; Sabers &

White, 1969), and confidence testing (de Finetti, 1965; Hambleton et al., 1970; Rippey, 1968) for assessing examinees' partial knowledge. The findings of these alternatives are usually interpreted in terms of test validity and reliability. However, they do not provide ability estimates either with known statistical properties or with standard errors of estimate associated with the estimated ability. Hence, this problem, as well as the preceding problem, invokes the consideration of using theoretically rigorous scoring models which are based on modern test theory. Latent trait theory (LTT) or item response theory (IRT) developed in modern test theory is the tool that we need to use (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 (1973), 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). 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 Masters' partial credit model is briefly reviewed in Yu (1991c).

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 in Masters (1982), Masters and Wright (1982), and Wright and 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 and Evans (1986). Besides, Smith (1987) shows that results of assessing partial knowledge in vocabulary support O'Connor's theory of vocabulary acquisition. Dodd and Koch (1987) indicates 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.

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 objected to several properties of the Rasch model, criticized its inappropriateness under several conditions, and concluded that the Rasch model was not suitable for multiple-choice items.

In a comparison of model fits, Albanese and Forsyth (1984) showed that the Rasch model failed to fit more items than did the two-parameter logistic model. Hambleton and Traub (1973) found that the two-parameter model predicted score distributions better than the Rasch model did. Goldman and 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. But one thing completely inconvincible in using the partial credit model is that partial credit model treats person ability estimtes the same if persons' raw scores were the same. This methodology and its underlying estimating algorithm for model parameters are actually inadequate in real situations. It is almost impossible for two persons to have identical response patterns as the number of items increases in real testing situations, particularly in using partial-credit-scoring directed test formats, even if they have the same raw scores. Therefore, the treatment of same raw scores as having same estimates is not appropriate.

It is reasonable to be skeptical about such an assumption of equal discrimination in the partial credit model that might be responsible for these weaknesses as criticized above. If we take a "step discrimination" parameter into account in the partial credit model, it might improve the precision of person ability estimations and the model fit, and make it suitable for multiple-choice items. In addition, providing step, item, and test information, it is useful in mastery testing to discriminate between mastery and nonmastery groups. It is also useful for item selection and test construction. Even for item banking and computerized adaptive testing, it implies a lot of advantages and potentials. Among these advantages, only the estimation, fitness, and information function of model parameters are selected to be studied in this exploratory research.

## II. Formulation of the Two-Parameter Partial Credit Model

The main purpose of this section is to generalize the one-parameter partial credit model, that is, Masters' (1982) partial credit model, to a two-parameter partial credit model.

The proposed two-parameter partial credit model shares the same philosophy with the one-parameter partrial credit model (Masters, 1982). Taking the step-discrimination parameter, $a_{ij}$, into account, the one-parameter partial credit model can be expanded to

$$\pi_{nix} = \frac{\exp \sum_{j=0}^{x} [a_{ij}(\beta_n - b_{ij})]}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(\beta_n - b_{ij})]} \qquad x=0,\ 1,\ \ldots,\ m_i \qquad (2.1)$$

which means that the probability of person n scoring x on the $m_i$-step item i is a function of the person's ability, $\beta_n$, on the test, the discriminations, $a_{ij}$, and the difficulties, $b_{ij}$, of the $m_i$ "steps" on item i.

### *Estimation of model parameters*

For notational convenience, it is assumed that $a_{i0}0 \equiv b_{i0} \equiv 0$, so that

$$\sum_{j=0}^{0} [a_{ij}(\beta_n - b_{ij})] \equiv 0.$$

Because the two-parameter partial credit model cannot separate the step discriminations from the estimations of ability and diffculty parameters, the conditional maximum likelihood (CML) procedure is no longer appropriate for the estimation of the two-parameter partial credit model parameters. That is, sufficient statistics do not exist for the two-parameter partial credit model. Hence, the unconditional maximum likelihood (UML) procedure is used instead. The unconditional maximum likelihood procedure is based on Wright and Panchapakesan's (1969) estimation algorithm for Rasch's (1980) dichotomous model.

The likelihood function of the entire data matrix (X) is the continued product of the probability, $\pi_{nix}$, over all persons n and items i, that is,

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

$$= \frac{\exp \sum_{n=1}^{N} \sum_{i=1}^{L} \sum_{j=0}^{x_{ni}} [a_{ij}(\beta_n - b_{ij})]}{\prod_{n=1}^{N} \prod_{i=1}^{L} \{ \sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(\beta_n - b_{ij})] \}} \qquad (2.2)$$

Taking logarithms,

$$\lambda = \log\Lambda$$

$$= \sum_{n=1}^{N} \sum_{i=1}^{L} \sum_{j=0}^{x_{ni}} a_{ij}\beta_n - \sum_{n=1}^{N} \sum_{i=1}^{L} \sum_{j=0}^{x_{ni}} a_{ij}b_{ij} -$$

$$\sum_{n=1}^{N} \sum_{i=1}^{L} \{ \log \sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(\beta_n - b_{ij})] \} \qquad (2.3)$$

For simplicity, let

$$F = \log \sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(\beta_n - b_{ij})] \qquad (2.4)$$

Taking the first partial derivatives of F with respect to $\beta_n$, $a_{ij}$, and $b_{ij}$, respectively, we obtain

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

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

$$= \sum_{k=0}^{m_i} \sum_{j=0}^{k} a_{ij}\pi_{nik} = \sum_{k=1}^{m_i} \sum_{j=1}^{k} a_{ij}\pi_{nik} \tag{2.5}$$

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

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

$$= \sum_{k=j}^{m_i} \sum_{j=0}^{k} (\beta_n - b_{ij})\pi_{nik} = \sum_{k=j}^{m_i} \sum_{j=1}^{k} (\beta_n - b_{ij})\pi_{nik} \tag{2.6}$$

$$\frac{\partial F}{\partial b_{ij}} = \frac{\dfrac{\partial}{\partial b_{ij}} \left\{ \sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(ß_n - b_{ij})] \right\}}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(ß_n - b_{ij})]}$$

$$= \frac{\sum_{k=j}^{m_i} \left(- \sum_{j=0}^{k} a_{ij}\right) \exp \sum_{j=0}^{k} [a_{ij}(ß_n - b_{ij})]}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} [a_{ij}(ß_n - b_{ij})]}$$

$$= - \sum_{k=j}^{m_i} \sum_{j=0}^{k} a_{ij}\pi_{nik} = - \sum_{k=j}^{m_i} \sum_{j=1}^{k} a_{ij}\pi_{nik} \tag{2.7}$$

where the discrimination, $a_{ij}$, and the difficulty, $b_{ij}$, of step j appear only in those terms for which $k \geq j$, so that the derivatives of $\sum_{k=j}^{m_i} \sum_{j=0}^{k} (\bullet)$ with respect to $a_{ij}$ and $b_{ij}$ truncate to be $\sum_{k=j}^{m_i} \sum_{j=1}^{k} (\bullet)$. This means that the probability of person n completing at least j steps in item i. The derivations of step parameters share the same schemes of Masters' (1982, pp. 164-165) partial credit model.

Then the first partial derivatives of $\lambda$ with respect to $ß_n$, $a_{ij}$, and $b_{ij}$, respectively, are

$$\frac{\partial \lambda}{\partial ß_n} = \sum_{i=1}^{L} \sum_{j=1}^{x_{ni}} a_{ij} - \sum_{i=1}^{L} \sum_{k=1}^{m_i} \sum_{j=1}^{k} a_{ij}\pi_{nik} \qquad n=1, \ldots, N \tag{2.8}$$

$$\frac{\partial \lambda}{\partial a_{ij}} = \sum_{n=1}^{N} x_{ni}ß_n - S_{ij}b_{ij} - \sum_{n=1}^{N} \sum_{k=j}^{m_i} \sum_{j=1}^{k} (ß_n - b_{ij})\pi_{nik} \qquad i=1, \ldots, L;\ j=1, \ldots, m_i \tag{2.9}$$

$$\frac{\partial\lambda}{\partial b_{ij}} = -S_{ij}a_{ij} + \sum_{n=1}^{N} \sum_{k=j}^{m_i} \sum_{j=1}^{k} a_{ij}\pi_{nik} \qquad i=1, \ldots, j=1, \ldots, m_i \qquad (2.10)$$

where $x_{ni}$ is the count of steps completed by person n on item i (that is, the raw score of person n), and $S_{ij}$ is the number of persons completing step j on item i (that is, the raw score of step j on item i).

The second partial derivatives of $\lambda$ with respect to $\beta_n$, $a_{ij}$, and $b_{ij}$, respectively, are

$$\frac{\partial^2\lambda}{\partial\beta^2_n} = -\sum_{i=1}^{L} \Big[ \sum_{k=1}^{m_i} \Big( \sum_{j=1}^{k} a_{ij}\Big)^2 \pi_{nik} - \Big( \sum_{k=1}^{m_i} \sum_{j=1}^{k} a_{ij}\pi_{nik}\Big)^2 \Big] \qquad (2.11)$$

$$\frac{\partial^2\lambda}{\partial a^2_{ij}} = -\sum_{n=1}^{N} \Big\{ \sum_{k=j}^{m_i} \Big[ \sum_{j=1}^{k} (\beta_n - b_{ij}) \Big]^2 \pi_{nik} - \Big[ \sum_{k=j}^{m_i} \sum_{j=1}^{k} (\beta_n - b_{ij})\pi_{nik}\Big]^2 \Big\} \qquad (2.12)$$

$$\frac{\partial^2\lambda}{\partial b^2_{ij}} = -\sum_{n=1}^{N} \Big[ \sum_{k=j}^{m_i} \Big( \sum_{j=1}^{k} a_{ij}\Big)^2 \pi_{nik} - \Big( \sum_{k=j}^{m_i} \sum_{j=1}^{k} a_{ij}\pi_{nik}\Big)^2 \Big] \qquad (2.13)$$

Then person-and step-parameters can be approximately estimated by using the Newton-Raphson iterative procedure, that is,

$$\beta_n^{t+1} = \beta^t_n - \frac{\displaystyle\sum_{i=1}^{L} \sum_{j=1}^{x_{ni}} a_{ij} - \sum_{i=1}^{L} \sum_{k=1}^{m_i} \sum_{j=1}^{k} a_{ij}P^t_{nik}}{\displaystyle -\sum_{i=1}^{L} \Big[ \sum_{k=1}^{m_i} \Big( \sum_{j=1}^{k} a_{ij}\Big)^2 P^t_{nik} - \Big( \sum_{k=1}^{m_i} \sum_{j=1}^{k} a_{ij}P^t_{nik}\Big)^2 \Big]} \qquad n=1, \ldots, N \qquad (2.14)$$

$$\hat{a}_{ij}^{t+1}=\hat{a}^{t}_{ij} - \frac{\sum_{n=1}^{N} x_{ni}\beta_n - S_{ij}b_{ij} - \sum_{n=1}^{N}\sum_{k=j}^{m_i}\sum_{j=1}^{k}(\beta_n - b_{ij})P^{t}_{nik}}{-\sum_{n=1}^{N}\left\{\sum_{k=j}^{m_i}\left[\sum_{j=1}^{k}(\beta_n - b_{ij})\right]^{2}P^{t}_{nik} - \left[\sum_{k=j}^{m_i}\sum_{j=1}^{k}(\beta_n - b_{ij})P^{t}_{nik}\right]^{2}\right\}}$$

$$i=1, \ldots, L;\ j=1, \ldots, m_i \qquad (2.15)$$

$$\hat{b}_{ij}^{t+1}=\hat{b}^{t}_{ij} - \frac{-S_{ij}a_{ij} + \sum_{n=1}^{N}\sum_{k=j}^{m_i}\sum_{j=1}^{k} a_{ij}P^{t}_{nik}}{-\sum_{n=1}^{N}\left[\sum_{k=j}^{m_i}\left(\sum_{j=1}^{k} a_{ij}\right)^{2}P^{t}_{nik} - \left(\sum_{k=j}^{m_i}\sum_{j=1}^{k} a_{ij}P^{t}_{nik}\right)^{2}\right]}$$

$$i=1, \ldots, L;\ j=1, \ldots, m_i \qquad (2.16)$$

where $P^{t}_{nik}$ is the estimated probability of person n responding in step k to item i after t iterations, $\hat{\beta}_n^{\,t}$ is the estimate of $\beta_n$ after t iterations, $\hat{a}_{ij}^{\,t}$ is the estimate of $a_{ij}$ after t iterations, and $\hat{b}_{ij}^{\,t}$ is the estimate of $b_{ij}$ after t iterations.

To get rid of indeterminancy in the scale origin, the mean step difficulty, b.., is usually set equal to zero. And the asymptotic estimates of standard errors are given by square roots of the reciprocal of the negative second partial derivatives from the last iteration, that is

$$SE(\hat{\beta}_n)=\left\{\sum_{i=1}^{L}\left[\sum_{k=1}^{m_i}\left(\sum_{j=1}^{k} a_{ij}\right)^{2}P_{nik} - \left(\sum_{k=1}^{m_i}\sum_{j=1}^{k} a_{ij}P_{nik}\right)^{2}\right]\right\}^{-1/2} \qquad (2.17)$$

$$SE(\hat{a}_{ij})=\left\{\sum_{n=1}^{N}\left[\sum_{k=j}^{m_i}\left(\sum_{j=1}^{k} (\beta_n - b_{ij})\right)^{2}P_{nik} - \left(\sum_{k=j}^{m_i}\sum_{j=1}^{k} (\beta_n - b_{ij})P_{nik}\right)^{2}\right]\right\}^{-1/2}$$

$$(2.18)$$

$$SE(\hat{b}_{ij})=\left\{\sum_{n=1}^{N}\left[\sum_{k=j}^{m_i}\left(\sum_{j=1}^{k} a_{ij}\right)^{2}P_{nik} - \left(\sum_{k=j}^{m_i}\sum_{j=1}^{k} a_{ij}P_{nik}\right)^{2}\right]\right\}^{-1/2} \qquad (2.19)$$

Due to the use of UML procedure containing a slight bias (Andersen, 1973c), a correction factor, (L-1)/L, is suggested to removing such a bias in the dichotomous case (Wright, 1988; Wright & Douglas, 1977a, 1977b). In the present case, it is suggested that the same correction may be appropriate for removing bias in parameters when $m_i > 1$ (Masters, 1982).

### *Test of goodness-of-fit*

Once $\beta_n$, $a_{ij}$, $b_{ij}$ are estimated, they are used to compute the expected scores for every person on each item. Expected scores are compare to the observed scores; their differences are residuals. The fit analysis of item i and person n is based on such residual values (Ludlow, 1985, 1986; Ludlow & Hillocks, 1985).

The expected score for person n on item i is obtained from the following equation

$$E_{ni} = \sum_{k=0}^{m_i} kP_{nik} \tag{2.20}$$

where $P_{nik}$ is the estimated probability of person n passing the k^th^ step on item i, and k is the step number, $k=0,1,2, \ldots, m_i$ ($m_i=3$ for all i in this example). Then a residual is defined by

$$R_{ni}=x_{ni}-E_{ni} \tag{2.21}$$

Such residuals can provide a check on the degree of fit of item and person estimates.

The expected variance for (2.20) can be computed as follows:

$$V_{ni}= \sum_{k=0}^{m_i} (k-E_{ni})^2P_{nik} \tag{2.22}$$

Then residuals can be expressed in standard form as:

$$Z_{ni}=\frac{x_{ni}-E_{ni}}{(V_{ni})^{1/2}} \tag{2.23}$$

There residuals have expected value zero and variance one. Thus, the Wald test (1943) can be used as an index of fit analysis for item i, person n, and the whole model.

For item i, the index is expressed as:

$$W_i = \sum_{n=1}^{N} (Z_{ni})^2 \qquad (2.24)$$

which follows an approximate chi-square distribution with (N−1) degrees of freedom. For person n, the index is expressed as:

$$W_n = \sum_{i=1}^{L} (Z_{ni})^2 \qquad (2.25)$$

which follows an approximate chi-square distribution with (L−1) degrees of freedom.

If person n answers in a manner consistent with his/her estimated ability and step parameters within each item (for example, no guessing even if he/she is unable to answer it), then the residuals will be small. Unexpected incorrect responses result in large negative residuals, and unexpected correct responses result in large positive residuals. Therefore, a large fit statistic, $W_i$, results when either unexpected failures or unexpected successes, or both, have occurred. Under such an index, a bad item is spotted. In like manner, a large fit statistic, $W_n$, results when person n does not consistently answer test items (due, for example, guessing, cheating, sleeping, fumbling, plodding, or cultural bias (Wright, 1977) ). Consequently, the unusual responses can be detected for person n.

For testing the goodness-of-fit of the whole model, the Wald test of (2.24) across item i or (2.25) across person n can be used as an index of model fit, that is,

$$W_m = \sum_{n=1}^{N} \sum_{i=1}^{L} (Z_{ni})^2 \qquad (2.26)$$

which follows an approximate chi-square distribution with (N−1) x (L−1) degrees of freedom. A large fit statistic, $W_m$, indicates that the model does not adequately fit the data. Under such a circumstance, other models are suggested.

### *Information functions*

According to Birnbaum (1968), the step information function can be written as

$$I_{ik}(\beta) = \frac{(P'_{ik})^2}{P_{ik}Q_{ik}} \tag{2.27}$$

where $P_k$ is the probability of passing the k<sup>th</sup> step on item i by an examinee with ability level ß, $P'_{ik}$ is the slope of the step characteristic curve at ability level ß, and $Q_{ik}=1-P_{ik}$. Equation (2.27) can be also precisely rewritten as

$$I_{ik}(\beta)= \frac{a^2_{ij}}{\exp[a_{ij}(\beta-b_{ij})] \;\{\; 1+\exp[-a_{ij}(\beta-b_{ij})] \;\}^2} \tag{2.28}$$

Due to the additive feature of information functions, the item information function is given by summing the step information functions; that is,

$$I_i(\beta) = \sum_{k=1}^{m_i} I_{ik}(\beta) \tag{2.29}$$

And the test information function is given by summing the item information functions; that is,

$$I(\beta)= \sum_{i=1}^{L} \sum_{k=1}^{m_i} I_{ik}(\beta) \tag{2.30}$$

The standard error associated with a maximum likelihood estimate of ability ß is given by the square root of the reciprocal of the value of the information function at ß. It is for this reason that information functions are important. Not only can they be used to assess the precision of ability estimates, but they can also be used in the design of tests. The step or item information function provides a measure of the usefulness of that step or item when used for a particular ability level. Therefore, the steeper the slope of the step or item characteristic curve, the higher will be the step or item information function at that particular ability level, and the estimate of that particular ability level will be more precisely determined.

## III. Dataset and Computer Program

### *Dataset*

The dataset examined in this study consists of item data for freshman students on an algebra test. Forty-eight algebraic items were administrated to 572 Algebra 1 students in a metropolitan school district in the midwest to assess achievement in learning algebra. Three items are assumed to measure a single cognitive component. Hence, by summing three successive items to form a component or subtest, we created 16 component or subtest scores. These 16 component scores are used in this dataset for illustration. Therefore, the dataset becomes a 572 × 16 raw data matrix, where each of the 16 component scores ranged from 0 to 3. Hereafter, we will treat these 16 component scores as 16 item scores.

The item analysis shows that the mean and standard deviation of 572 persons' scores on these 16 items are 27.83 and 8.31. The inter-item correlations range from .132 to .424 and the average is .275. The coefficient alpha is .857 indicating that these 16 items are highly consistent. Besides, the result from factor analysis shows that one dominant factor accounts for 32.3% of variance. Other factors contributed about equally to the variance. This feature fits Lord's (1980, p.21) two suggestions about a rough test of unidimensionality for a given test. Two such conditions are (a) the first eigenvalue is large compared to the second and (b) the second eigenvalue is not so much larger than any of the others. Hence, the current dataset is approximately unidimensional. The assumption of unidimensionality of item response theory holds for partial credit scoring in this case.

### *Computer program*

A FORTRAN 77 computer program, called TPPCM (Yu, 1991b), is developed by the present author to estimate, test goodness-of-fit, and provide information functions for the two-parameter partial credit model parameters. This program adopts a revised two-stage "back and forth" iterative procedure from Birnbaum (1968) and Wingersky (1983) for the unconditional maximum likelihood estimation, or called joint maximum likelihood estimation (JMLE), of these parameters. The intent is to obtain a global maximum for the overall likelihood function. So the stopping rule of iterations is set as the absolute value of increment in the log-likelihood function for two successive iterative stages or cycles being less than 0.01.

Because discrimination estimates for one or more steps may become very large in an ill-structured dataset, and ability and difficulty estimates are arbitrarily decided with no scale origins. Some constraints are used to handle such identification problems. To decide the scale of ability estimates, $ß_n$s are scale to be on the same scale as difficulty estimates with a mean zero and a variance one. The average of difficulty estimates, b.., is set to be zero, and the maximum value of discrimination estimates is set to be 10.

This program is used to run such a dataset and has eleven major subroutines and a top-down main program. It adopts some programming skills from numerical analysis (Atkinson, Harley, & Hudson, 1989) and is designed for ease of understanding (Koffman & Friedman, 1990).

This program as currently developed is to handle datasets with three-step items. So the input and output formats are treated as fixed. For future use of dataset with more, or less, than three-step items, the fixed format should be free and redesigned.

This program is also designed for easy use. Users only provide the numbers of persons, items, and steps in the main program, as well as the chi-square values for L degrees of freedom (L being the number of items) at $p$=.99 and $p$=.95 levels. Prepare the input data (that is, preson by item matrix) with first five columns reserved for identification numbers. Then put them together into this program. This program will automatically print the following results: (a) initial step parameters, (b) maximum likelihood functions for each stage and cycle, (c) estimates of step parameters (that is, discriminations and difficulties and their standard errors), (d) fit statistics for step parameters, (e) persons' response patterns, raw scores, ability estimates, standard errors of estimates, true scores, and fit statistics, (f) fit statistic for the whole model, and (g) plots of category characteristic curves, step characteristic curves, and three kinds of information curves.

For the detailed algorithm, readers are referred to Yu (1991a, 1991b).

## IV. Results

The two-parameter partial credit model has been used to analyze the dataset. Estimates of step parameters, their calibration errors, and statistics summarizing the fit of item i and the whole model are given in Table 1. Category characteristic curves (CCCs) are shown in Figure 1 and step characteristic curves (SCCs) are shown in Figure 2 as tools for interpreting the estimation of step parameters.

Table 1 shows that there are many difficulty and discrimination patterns in this test. The fit statistics show that Items 10, 15, and 16 are probably bad items. They may not be consistently constructed to partially scoring persons' knowledge levels. The overall fit statistic for the test is not significant. This indicates that the two-parameter partial credit model fits the 16-item test.

Since some step difficulties or discriminations are small and some are large in a given item, it is hard to say which item is more difficult or discriminating than others. We will choose three quite different items to illustrate and discuss their properties.

### *Estimation of step difficulty parameters*

From Table 1 and Figure 1, three sets of parameter estimates for each item determine a unique set of category characteristic curves for the four performance levels in that item. The estimates $b_{i1}$, $b_{i2}$, and $b_{i3}$ are located at the intersections of curves 0 and 1, 1 and 2, and 2 and 3. These three illustrative items have quite different difficulty patterns. They represent different types of test items. Their features are analyzed and discussed as follows.

For person n with ability estimate less than 0.017 logits the most probable score on Item 4 is 0. Persons with ability estimates greater than 0.017 logits but less than 0.178 logits will probably score 1 on Item 4. Persons with ability estimates greater than 0.178 logits but less than 3.749 logits will score 2 on Item 4. To score 3 on Item 4, a person needs an ability estimate greater than 3.749 logits. Because both difficulites of the first and the second steps on Item 4 are very close, this feature makes Item 4 like a two-step item. Besides, the difficulty of the third step is much higher than the first two. Consequently, most persons whose abilities are greater than 0.02 logits may probably score 2 rather than 1 on Item 4. Although Item 4 loses one step function, it may be retained in order to give partial credit to persons with partial knowledge. Table 1 also shows that the lack-of-fit stastistic for Item 4 is not significant at $\alpha=0.05$. Hence it is still a useful item.

Item 10 has a little different difficulty pattern from Item 4. The difficulites of the first two steps are almost the same. Besides, the difficulty of the third step is not far from the first two. Therefore, Item 10 is a near-binary item indeed. Persons with ability estimates less than $-0.047$ logits will never achieve event the first step of Item 10. Those whose ability estimates greater than 0.039 logits but less than 0.513 logits may have little chance to score 2 rather than 1 on Item 10. Only persons whose ability estimates greater than 0.513 logits may score 3 on Item 10. Item

Table 1

*Difficulty and Discrimination Estimates from the Two-Parameter Partial Credit Model*

| Item | $b_{i1}$ | $b_{i2}$ | $b_{i3}$ | $a_{i1}$ | $a_{i2}$ | $a_{i3}$ | $\chi^2$ |
|---|---|---|---|---|---|---|---|
| 1 | -0.841 | -0.239 | 0.618 | 1.016 | 1.020 | 1.059 | 531.857 |
| | (.020) | (.019) | (.019) | (.012) | (.012) | (.013) | |
| 2 | -0.744 | 0.739 | 1.040 | 1.070 | 0.998 | 1.148 | 347.527 |
| | (.027) | (.023) | (.023) | (.016) | (.016) | (.017) | |
| 3 | -0.065 | 0.197 | 1.632 | 1.149 | 1.135 | 1.288 | 482.327 |
| | (.024) | (.022) | (.023) | (.018) | (.018) | (.020) | |
| 4 | 0.017 | 0.178 | 3.749 | 0.911 | 0.918 | 1.176 | 364.092 |
| | (.035) | (.029) | (.031) | (.021) | (.020) | (.024) | |
| 5 | -1.754 | -0.172 | 0.254 | 0.965 | 0.959 | 1.002 | 483.054 |
| | (.020) | (.019) | (.019) | (.009) | (.009) | (.010) | |
| 6 | -1.389 | -0.317 | -0.125 | 0.978 | 0.979 | 1.015 | 536.448 |
| | (.018) | (.018) | (.018) | (.009) | (.009) | (.010) | |
| 7 | -0.651 | 0.068 | 0.928 | 1.048 | 1.032 | 1.105 | 554.507 |
| | (.022) | (.020) | (.021) | (.014) | (.014) | (.015) | |
| 8 | -0.332 | -0.302 | 0.402 | 1.086 | 1.101 | 1.126 | 456.487 |
| | (.018) | (.018) | (.018) | (.013) | (.013) | (.013) | |
| 9 | -0.608 | 0.340 | 0.998 | 1.080 | 1.041 | 1.145 | 476.223 |
| | (.023) | (.021) | (.021) | (.015) | (.015) | (.016) | |
| 10 | -0.047 | 0.039 | 0.513 | 1.152 | 1.149 | 1.187 | 765.096** |
| | (.019) | (.018) | (.019) | (.015) | (.015) | (.016) | |
| 11 | -0.631 | -0.015 | 0.514 | 1.064 | 1.054 | 1.103 | 542.014 |
| | (.020) | (.019) | (.019) | (.013) | (.013) | (.014) | |
| 12 | -0.836 | 0.014 | 0.351 | 1.050 | 1.036 | 1.086 | 444.002 |
| | (.020) | (.019) | (.019) | (.012) | (.012) | (.013) | |
| 13 | -1.201 | -0.786 | -0.399 | 0.966 | 0.983 | 1.009 | 502.251 |
| | (.017) | (.017) | (.017) | (.009) | (.009) | (.009) | |
| 14 | -0.574 | -0.696 | -0.430 | 0.999 | 1.018 | 1.047 | 391.619 |
| | (.017) | (.016) | (.017) | (.010) | (.010) | (.010) | |
| 15 | 0.730 | 0.026 | -0.277 | 1.195 | 1.214 | 1.242 | 1114.380** |
| | (.017) | (.016) | (.017) | (.016) | (.016) | (.016) | |
| 16 | -0.472 | 0.067 | 0.487 | 1.086 | 1.071 | 1.121 | 665.798** |
| | (.020) | (.019) | (.019) | (.014) | (.014) | (.014) | |
| | | | | | | | Total : 8649.156 |

Note. Standard errors are shown in parentheses.
** $p < .01$. * $p < .05$.
$r_{ab}=.520$ ($p < .001$).

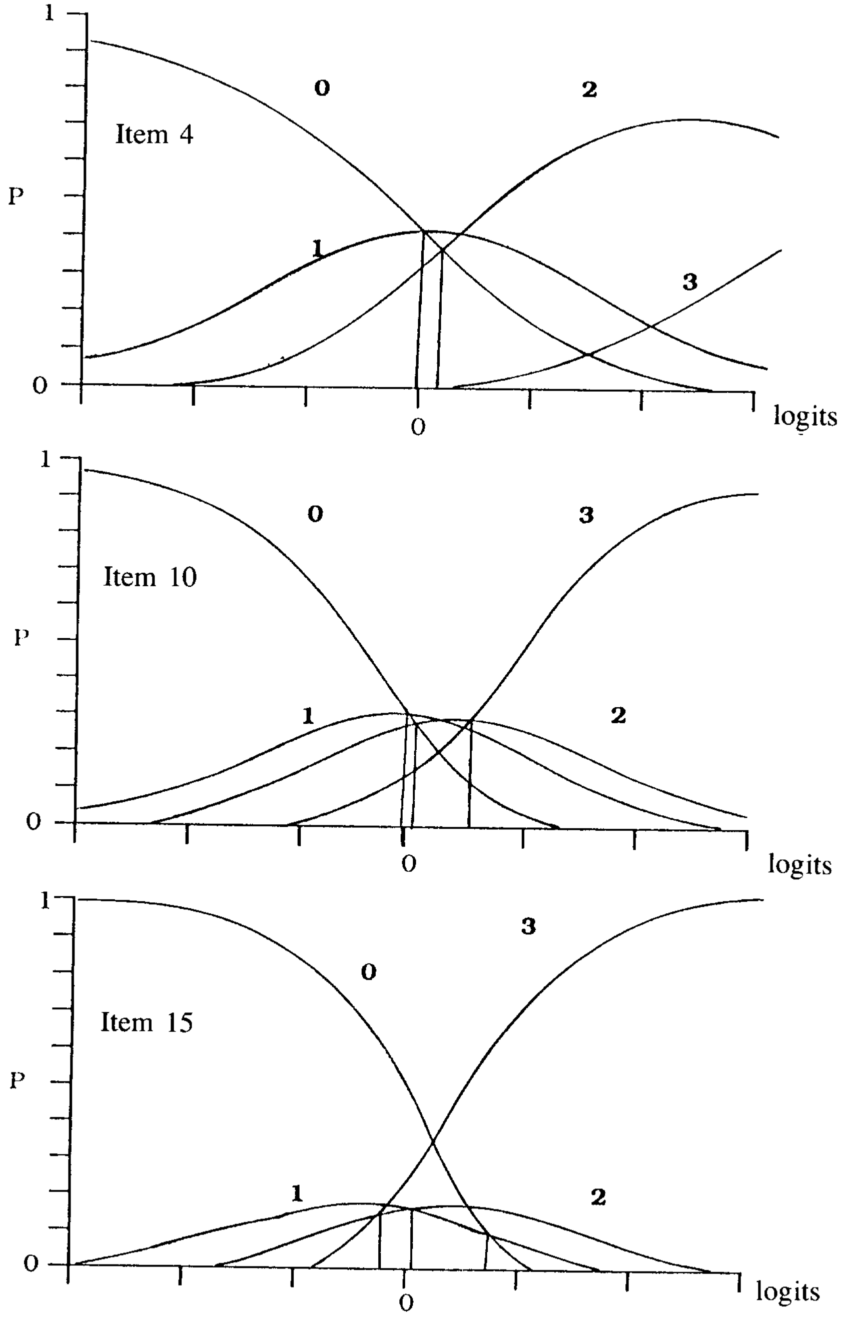

Figure 1. Category characteristic curves for three illustrative items.

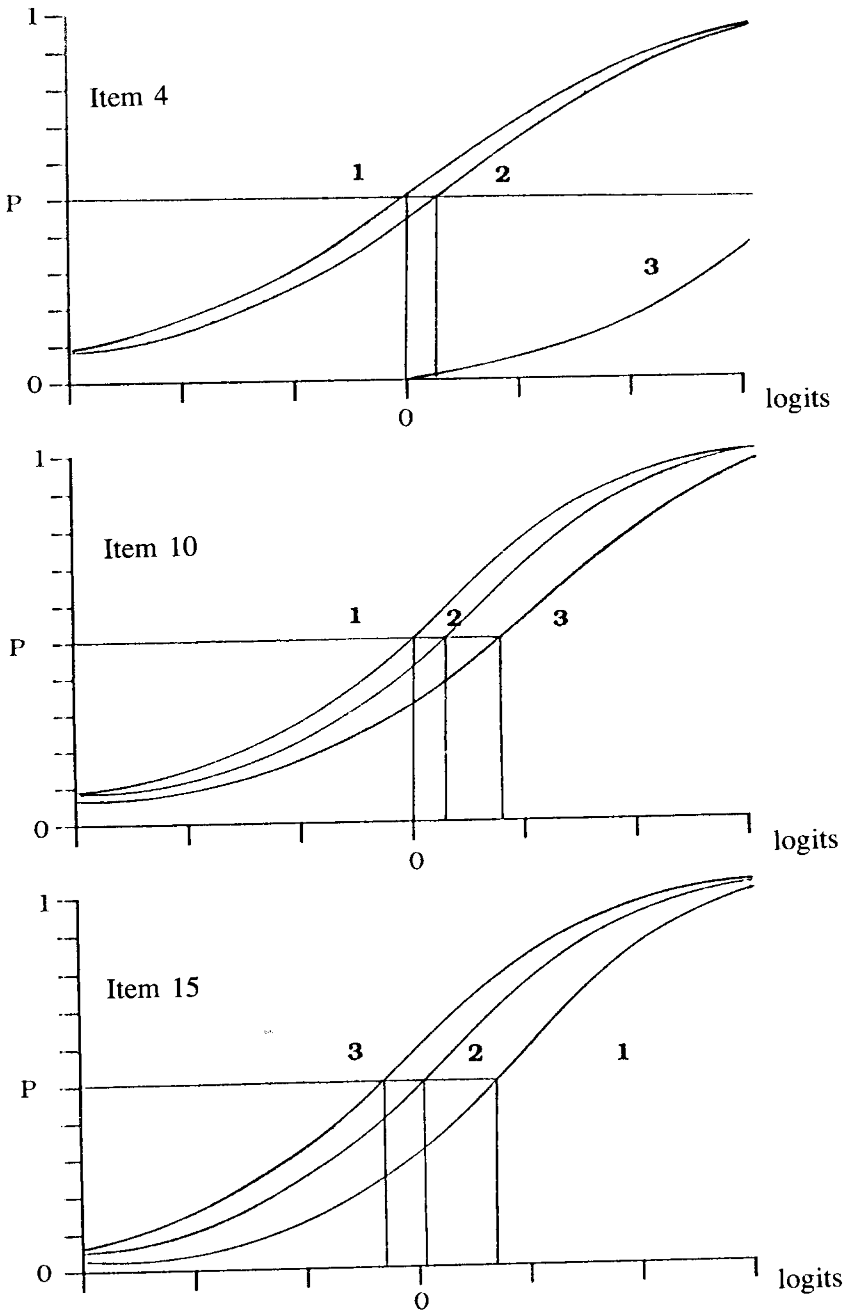

Figure 2. Step characteristic curves for three illustrative items.

10 loses its original function made up to give partial credit to persons with partial knowledge. This implies that the most likely score on Item 10 is either 0 or 3. This binary feature makes Item 10 a bad item from a partial credit model perspective.

Table 1 shows that Item 10 is not a good item because of a significant lack-of-fit statistic at $\alpha=0.01$. The significant lack-of-fit statistic reveals that large positive residuals occur and large unexpected successes exist on Item 10. This may be due mainly to its different step difficulties. Since the difficulties of the first two steps are the same and the third step is not every difficult either, Item 10 is, in fact, a near-binary item. Besides, each step is almost equivalently discriminating. Item 10 loses its function of differentially discriminating different ability levels. Consequently, Item 10 becomes indistinguishable and useless in identifying partial knowledge (see analysis below too). This item may not be well constructed to partially score different knowledge levels. It needs to be remodified or redesigned.

Item 15 is a quite different item from the above two. The difficulties of three steps are in reverse order. That is, step 1 is more difficult than step 2, and step 2 is more difficult than step 3. From the point of view of difficulty in problem solving, persons who pass the first step (that is, the most difficult step) will certainly pass the second and the third steps (that is, the easier steps) too. There is no reason for person n, whose ability is higher than 0.730 logits, to fail the easier step subtaks (that is, the second and the third step subtasks). So the most probable score on Item 15 should be 3 or 0. Since the number of persons who achieve this item is not dichotomously distributed, Item 15 may not be consisently designed to measure the same (or the same directional) latent trait as other items do. This item needs to be totally revised or redsigned.

Table 1 shows that the lack-of-fit statistic of Item 15 is significant at $\alpha=0.01$. Large positive residuals also show that a large number of unexpected successes occurs on Item 15. This means that Item 15 is really a bad item. This item might not have been developed intentionally to score for partial knowledge. Since persons who pass the first step will always pass the latter two steps, and the step discriminations are not so different from each other, the major determinatnt of persons' performance on Item 15 completely depends on the step difficulties only. Because of the reversed order in difficulties, Item 15 may lose two step functions served as partially grading. This may or may not cause serious misfit problems. If step discriminations were significantly different from each other, the item function might still fit its constructed purposes, and Item 15 become a good item or, at least, not bad an item. If step discriminations were nearly the same as each other as in the current case, an item that lost its function might lose its fitness to the test purposes. But this reality still needs to be proved by checking the item content, test instruction, and step discriminating features (see analysis below too).

### *Estimation of step discrimination parameters*

As shown in Table 1 and Figure 2, three sets of parameter estimates for each item define a specific set of step characteristic curves for the four performance levels in that item. The discrimination estimates $a_{i1}$, $a_{i2}$, and $a_{i3}$ and difficulty estimates $b_{i1}$, $b_{i2}$, and $b_{i3}$ jointly mark be the spots whose corresponding probabilities are exactly equal to .5, and whose difficulty locations are the same as in CCCs (that is, Figure 1), through the intersection of pairs of successive curves. These three curves separate the performance levels on the whole item into four specific areas. Each area bounded or separated by any two successive curves represents each probability of person n's various scores x (ranging from 0 to 3) on item i.

The step characteristic curves should be interpreted in the same way we did the category characteristic curves, but slightly differently from our usual concept of item characteristic curves (ICCs) for a binary item. That is, for any given ability estimate, the sum of ordinate which vertically cross the four areas should be equal to one. Figure 2 indicates that the most probable scores for middle ability estimates on Item 4 are either 0 or 2, because the probability of areas bounded by step 1 and step 2 is very small. It also shows that the probable score for higher ability estimates is either 2 or 3, depending on how large ability estimate a person has, and lower ability estimates 0, because their areas are larger than the middle one.

In Item 10, persons with middle ability levels will have some chance to score 2 rather than 1. Because the area bounded by curves 2 and 3 is larger than that bounded by curves 1 and 2. The probable scores for a higher ability person are 3s and a lower ability person Os. This case is similar to Item 4.

Similar interpretations can be applied to Item 15 too, but two things are different Firstly, the interpreation should be in the opposite direction. Secondly, middle ability persons will have more chance to score 2 than to score 1 on Item 15. Higher ability persons will have a lot of chance to score 3 (that is, right-hand side area), and lower ability persons will have a lot of chance to score 0 (that is, the left-hand side area), because the area bounded by curves 1 and 2 is larger than that bounded by curves 2 and 3. Other areas are similar to those of Item 10.

From Figure 2 we know that step discriminations will make the difference between each subtask performance of persons at any ability level possible. Step characteristic curves with different discriminations will jointly determine the probability of partial scores of persons at any possible ability level, when the model holds. This is the potential of the two-parameter partial credit model on interpretation of partial scores.

For any item with closer step difficulties, step discriminations may confusedly discriminating different ability persons. This makes the areas partitioned by step characteristic curves become mixed and entangled with each other. It also makes the interpretation of performance out of the intertwined areas ambiguous and unreliable. Consequently, such an item loses its partially scoring function. Therefore, a purposely clear and well-functioned item should not have intertwining step characteristic curves. Other well-constructed and normally functioning items will have salient step characteristic curves and make the partitions of performance space more obvious.

### *Features of step parameters*

From Table 1, Figure 1, and Figure 2 we know that these three illustrative items reveal much of the real situation in educational testing. The item steps may 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. This phenomenon shares the same features of the one-parameter partial credit model.

Besides, step discriminations really exist. They may not be same for each step of each item. Their existence can explain what ability levels really can be measured by a given item (see interpretation below too) and show how subtasks discriminate different ability levels. Consequently, the estimation of person abilities can be improved by the use of the two-parameter partial credit model.

Both Figures 1 and 2 help us interpret how the items behave and how persons' performance levels are identified. For any given ability level, the sum of probabilities for scoring x (x ranges from 0 to 3 in this test) is always equal to one. This basic assumption coupled with step characteristic curves, as well as step discriminations, make the partitions of performance space clearer. From the performance space on each item, educators (or psychologists) will more easily evaluate students' (or subjects') partial knowledge levels (or partial latent traits) and identify at which step their problem solving may go wrong. This feature will make the design of cognitive-component test and the evaluation of learning results more efficient and powerful.

### *Estimation of ability parameters*

Some estimates of ability parameters, their calibration errors, response patterns, raw scores, true scores, and indexes of fit of person n are selected to be discussed and are shown in Table 2. Partitions of different ability levels by different step difficulties and peaked step information are drawn in Figure 3. Item and test information functions which more precisely measure some ability levels are illustrated in Figure 4. Both figures are used to help interpret the estimation of ability parameters.

There is a negative relation between ability estimates and their standard error estimates: that is, the higher the ability estimates, the smaller the standard error estimates will be. This means that the estimates of higher abilities are quite precise, but the estimates of lower abilities may be imprecise. Hence, the fit statistics of lower ability estimates may be significant and show their responses to be unusual. Because true scores are monotonically realted to ability estimates, the higher the ability estimates, the larger the true scores will be.

The estimates of ability parameters of some selected persons ordered in raw scores ranging from 4 to 47 (the minimum score is 0 and the maximum is 48) are shown in Table 2. Form Table 2 we know that persons with the same raw scores will not necessarily have the same ability estimates, although the differences among them are slight. This is due to their slightly different response patterns and different step discriminations. This phenomenon indicates that the one-parameter partial credit model might not be appropriate in estimating person ability by assuming that persons with same raw scores will have same ability estimates. In addition, standard errors are decreasing in their magnitudes with increasing ability estimates. This means that the higher ability estimate is more precise and reliable when measured by the computer program, TPPCM (Yu, 1991b). Because true scores are monotonically related to ability estimates, so their magnitudes are increasing with ability estimates.

Table 2 aslo shows that lower ability and some higher ability persons may not consistently answer the test items. Since large, significant fit statistics occurred in lower ability persons, large positive residuals reveal that their responses are inconsistent. Large, significant fit statistics occur in some higher ability persons, indicating that their negative residuals are large. This means that their responses

Table 2

*Some Selected Ability Estimates from the Two-Parameter Partial Credit Model*

| Id # | Response patterns | Raw scores | $\beta_n$ | Se | True scores | $\chi^2$ |
|---|---|---|---|---|---|---|
| 64 | 0000110000001001 | 4 | -3.260 | .426 | 1.19 | 27.746** |
| 20 | 0012101100011000 | 8 | -2.324 | .301 | 3.04 | 44.116** |
| 306 | 2100000012110100 | 9 | -2.103 | .281 | 3.79 | 48.202** |
| 33 | 0020002001002121 | 11 | -1.741 | .247 | 5.43 | 77.289** |
| 107 | 1210010112101001 | 12 | -1.590 | .232 | 6.29 | 27.229** |
| 160 | 2100102210002200 | 13 | -1.444 | .219 | 7.25 | 17.350 |
| 193 | 0111211012101200 | 14 | -1.293 | .206 | 8.40 | 14.652 |
| 283 | 2010011001222002 | 14 | -1.285 | .206 | 8.46 | 16.946 |
| 280 | 0101222111101300 | 16 | -1.009 | .185 | 10.96 | 8.806 |
| 398 | 1302112100103021 | 18 | -0.736 | .166 | 13.99 | 17.379 |
| 383 | 0102120012112221 | 18 | -0.733 | .166 | 14.02 | 11.437 |
| 291 | 1201120112023301 | 20 | -0.486 | .151 | 17.24 | 6.945 |
| 149 | 1221120012222202 | 22 | -0.244 | .138 | 20.74 | 6.125 |
| 327 | 1211231200023202 | 22 | -0.248 | .138 | 20.67 | 5.227 |
| 349 | 2112122110232202 | 24 | 0.015 | .128 | 24.23 | 4.935 |
| 50 | 2101121212221332 | 26 | 0.223 | .119 | 27.88 | 5.613 |
| 394 | 0111231112213232 | 26 | 0.222 | .119 | 27.86 | 5.634 |
| 409 | 2122122222311221 | 28 | 0.447 | .112 | 31.12 | 8.315 |
| 128 | 1321322320223301 | 30 | 0.673 | .106 | 34.04 | 11.662 |
| 441 | 2212132322222220 | 30 | 0.675 | .106 | 34.07 | 9.049 |
| 449 | 1111233223323302 | 32 | 0.913 | .101 | 36.69 | 12.467 |
| 166 | 3222322322113330 | 34 | 1.164 | .097 | 38.97 | 19.774 |
| 500 | 2111333312323321 | 34 | 1.165 | .097 | 38.98 | 11.944 |
| 464 | 2232331223113332 | 36 | 1.434 | .094 | 40.94 | 21.325 |
| 536 | 3322333111333313 | 38 | 1.728 | .091 | 42.59 | 50.058** |
| 487 | 2322232223233333 | 40 | 2.073 | .089 | 44.02 | 19.571 |
| 95 | 3332322223332333 | 42 | 2.488 | .087 | 45.22 | 42.836** |
| 563 | 2232333323333333 | 44 | 3.037 | .085 | 46.24 | 26.535* |
| 496 | 3321333323333333 | 44 | 3.037 | .085 | 46.24 | 18.409 |
| 475 | 3332333333333333 | 47 | 3.772 | .085 | 47.62 | 3.227 |

** $p < .01$. * $p < .05$.

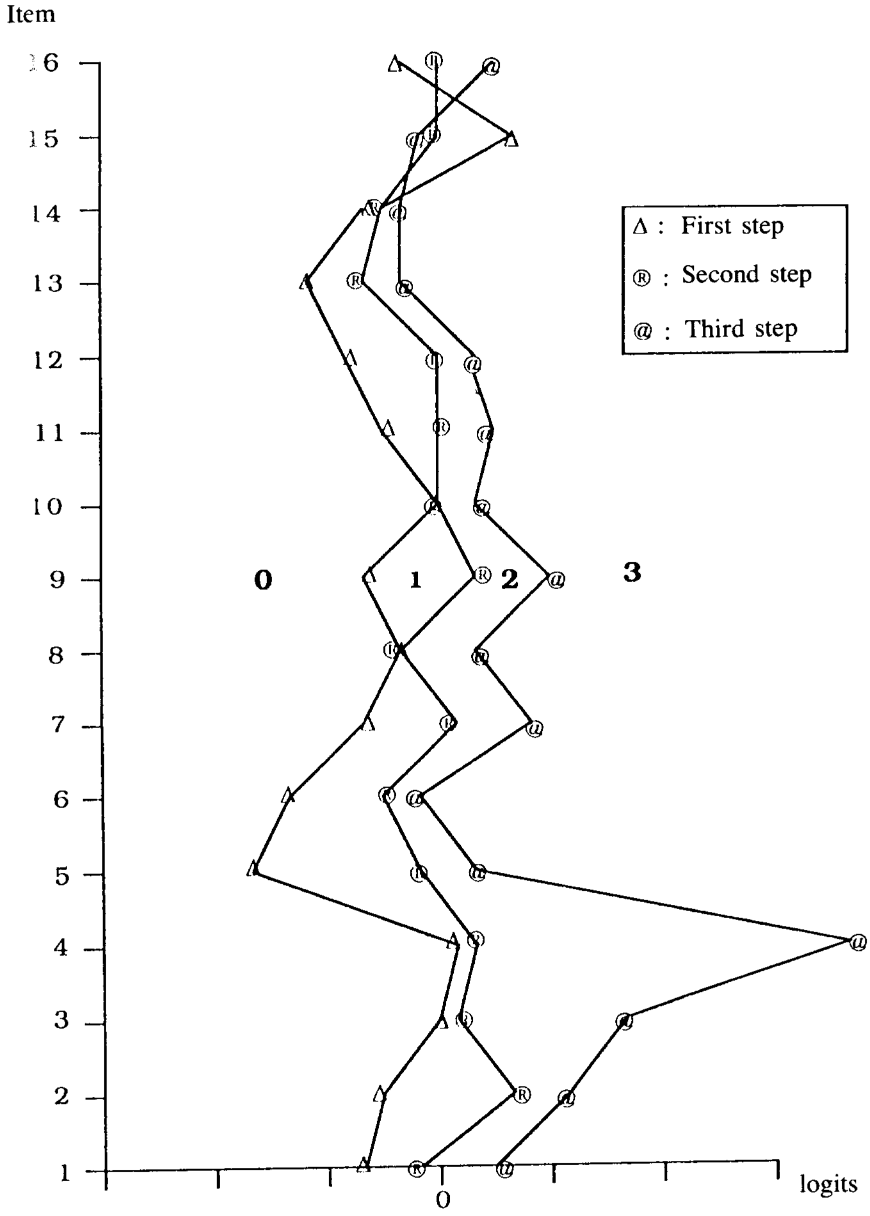

Figure 3. Partitions of performance levels with different step difficulties and peaked step information which correspond to the same locations of step difficulties.

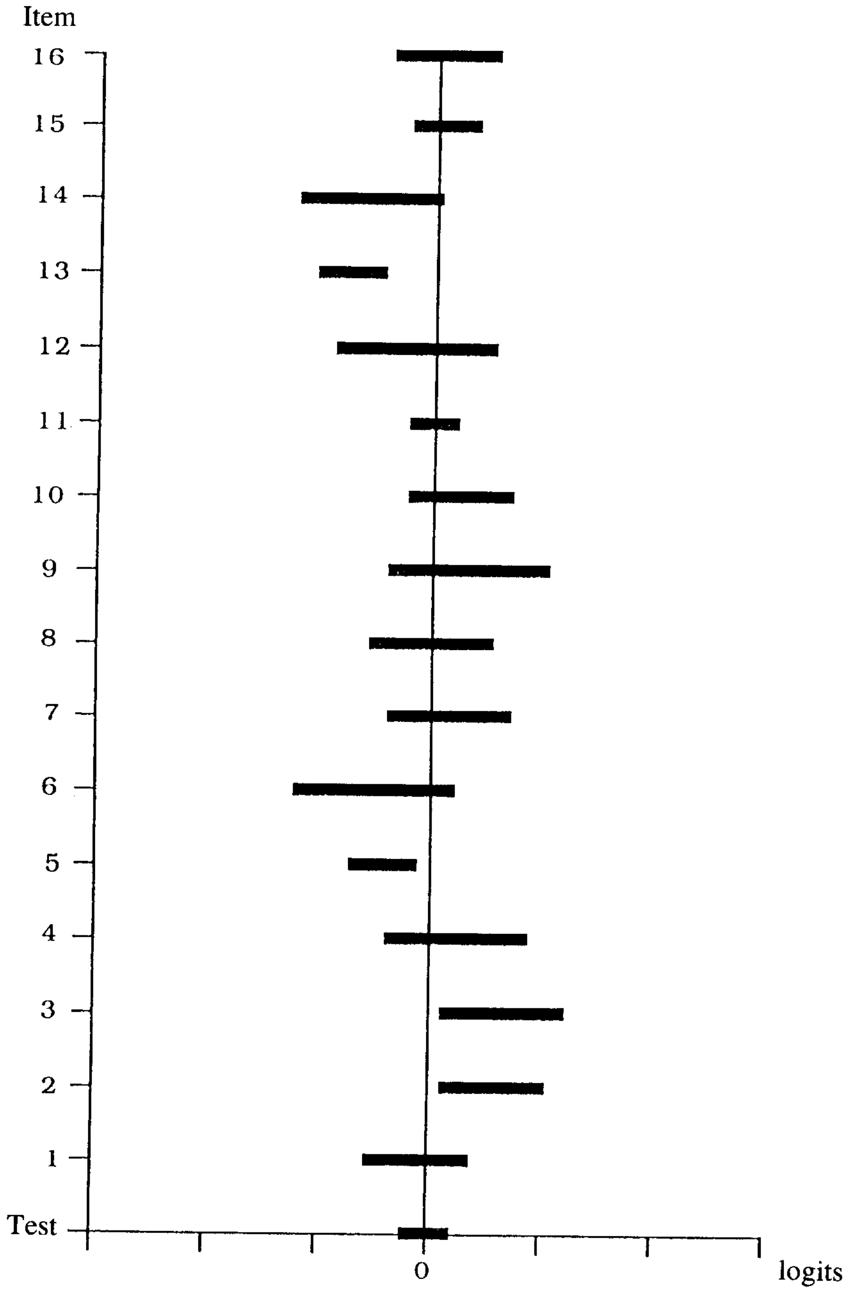

Figure 4. Ranges of ability estimates which correspond to peaked item and test information.

are unusual. Their unexpectedly successful and unsuccessful responses may be due to other reasons, for example, guessing, cheating, sleeping, fumbling, plodding, or cultural bias (Wright, 1977). This table also shows that most persons with middle to high ability estimates answer the test items consistently. Because their residuals are small, the corresponding fit statistics are not significant. This means that most middle to higher ability persons consistently respond to the current test items.

From Figure 3 we know that some particular ability estimates which correspond to the peaked step information function have the same corresponding locations of step difficulties. Therefore, performance levels are partitioned into four areas as shown in Figure 3. Most of lower ability persons are not expected to achieve any score on most items. But they get several 1s and 2s on their response patterns. Such unexpected successes result in large positive residuals. Consequently, fit statistics indicate that their responses may not be consistent. On the other hand, some higher ability persons are expected to achieve every item. But they get some 2s and 1s on some easier items. Such unexpected failures result in large negative residuals. Hence, fit statistics show that their responses are not consistent either. Other persons are expected to achieve items which correspond to their ability estimates. So that residuals are very small and no fit statistics signal their unusual responses. Their responses are more consistent than those of lower ability and some higher ability persons.

Figure 3 also shows that different steps discriminate different ability levels. Such a fact is shown on the same locations of step diffculties. Each step measures the same ability estimates as the step difficulties to which it corresponds. Therefore, Figure 3 is a map that shows how different ability areas (or performance levels) are partitioned by different step discriminations. Thus, residual analyses and fit statistics can be rechecked by visualizing persons' response patterns and their corresponding locations of ability estimates. If some responses occur in unexpected areas, such responses may be questionable. Their residuals may be large, and hence their fit statistics may be significant. For example, person #33 is supposed to achieve only the first step on Item 5 and score 1, as can be seen by checking the location of his/her ability estimate on this map. But he/she gets 2s on four items and 1s on three items whose difficulties are beyond his/her ability to handle. Hence, his/her unexpectedly successful responses on these items must be questionable. Person #536 is supposed to achieve every step, except the third step of Item 4, according to the location of his/her ability estimate on this map. But he/she gest 1s on four items and 2 on two items whose difficulties should be under his/her control. Hence, his/her unexpectedly failed responses on these items have to be doubtful. From both examples we know that Figure 3 provides a good help of interpreting ability estimations.

### *Information functions*

The particular ability levels that correspond to peaked item and test information functions are shown in Figure 4. From Figure 4 we know that the test information function precisely measures ability levels ranging from −.20 to .20 logits. Although different items may precisely measure different ability levels, they all center on the average difficulty estimate. That is, this test is more suitable for use in measuring middle ability persons. The estimation of ability parameters in middle ranges will be more precise than those at the extremes of the ability continuum.

The black bars shown in Figure 4 are the ranges of ability levels which correspond to peaked item and test information. They mean that the range of ability levels inside such a bar is precisely measured by a given item. Becuase the item information is summed from the step information, the ability levels which each item precisely measured are different from item to item. Taken as a whole, since the test information is the sum of the item information functions, the middle ability levels are most precisely measured by this 16-item test. The evidence is also shown in Figure 4.

Generally speaking, the two-parameter partial credit model fits this test. That is to say, each item can be used for screening different ability levels, although three items did not function well and needed to be revised. Especially, this test is suitable for use in scoring partial knowledge of middle ability persons. For lower ability persons, this test may not function well. Because most items are too difficult for them, their unexpected successes on some items may be due to other factors (as noted by Wright, 1977) that the test is not intended to measure. For some higher ability persons, this test may not work well either. Because most items are easy for them, unexpected failures on some items may be due to some personal problems (for example, fatigue, boredom, mindlessness, anxiety, stereotype, etc.) that the test cannot measure. Hence, persons' partial knowledge can be more precisely assessed by the use of the two-parameter partial credit model.

## V. Discussion

From the analysis of an empirical example shown in the preceding section, we know that step discriminations really exist. That is, the discrimination for each step of each item will not be same as the one-parameter partial credit model assumed. This finding not only proves the possibility of the two-parameter partial credit model, but also answers Masters' (1982, p. 172) interpretation that item steps [that is, difficulties| which interact with ability level to become "poorly or too discriminating" result in a misfit item. In fact, it is step discriminations, not step difficulties, interacting with person ability which highlights an item misfit or a person's response to be unusual or inconsistent as expected. This interactive and discriminating feature shown in Figure 3 makes person performance different and cause residual analyses to be examined. These deviations from expected values are summarized in the interpretation of fit statistics.

When step difficulties are very close to each other for an item, step discriminations may easily interact with different ability persons. Step characteristic curves may cross each other for such an item. This makes the interpretation of performance levels partitioned by step characteristic curves difficult. Persons whose ability estimates go beyond the intersection of two step characteristic curves may score more or less points on such an item. The reversed situation can be applied to someone whose ability estimate is below the intersection of two step characteristic curves. Hence, the performance levels bounded by intertwined step characteristic curves become very unstable. Several factors may contribute to this interaction of step discriminations and person abilities, for example, the content of test item, the instructions for test administration, and some nuisance variables described in Wright (1977). Such an interaction may result in a misfit item. A misfit item not only loses its original function made up to partially score persons' partial knowledge, but also contaminates the test of goodness-of-fit of the whole model. Therefore, a misfit item deserves to be remodified or redesigned for future use.

When step difficulties are not close to each other for an item, step characteristic curves may also intertwine with each other, depending on how step discriminations are distributed. But this situation will not make the interpretation of performance levels partitioned by step characteristic curves ambiguous and unreliable. Because partitions of ability areas are more obvious than those described above, the interpretation of goodness-of-fit of items will be easier than before. In addition, step difficulties may not be increasingly ordered in step sequence. This is basic feature of the partial credit model. But we have to indicate that a reversed order in step difficulties may make the item lose its constructional purpose and easily result in a misfit item, simply cooperating with different step discriminations.

From the above analysis and discussion, we may induce an important feature about a good item. An item with intertwined step characteristic curves may not necessarily be a bad item. But a good item with a clear and well-constructed purpose should not have intertwining step characteristic curves. A good item should have salient step characteristic curves and make the partition of performance space more obvious.

Although writing test-items is more of an art than a science, several suggestions may be made when creating such an art in partial scoring items. First, identify salient subfactors, major steps, or conscepts (totally, call them "components" for short) of an item purported to measure. Second, locate the underlying hierarchical relationships or orders among these components of an item. Third, design any possible format that randomly mixes these components. So far as the persent author knows, the multiple-choice item (including the Likert-type scaled questionnaire) with partial credits assigned to each choice and the open-ended eassay or test with expers' judgments or scoring are the most suitable formats. This also depends on designers' writing skills. Fourth, conduct a preliminary test to collect values of step parameters and test goodness-of-fit for such items. Fifth, revise bad items and save good items for future use. These suggestions may help to create good partial scoring items.

The existence of step discriminations can both improve the estimation of person abilities and information functions for item selection. Persons with the same raw scores will not necessarily get the same ability estimates. This is due to their different response patterns and different step discriminations and difficulities. Since step discriminations are provided, the estimation of ability parameters can be more precisely determined than the one-parameter partial credit model does. In addition, not as the one-parameter partial credit model does — the peaked step information always corresponds to the zero logit on the ability continuum, the two-parameter partial credit model has different locations on the ability continuum which correspond to the peaked step informations. Scuh locations are the same as the step difficulities. This finding can help interpret how person performance is discriminated by item steps. For example, by mapping persons' abilities in Figure 3 and checking their response patterns, the unusual responses are easily spotted by eye. And how many scores they should have are also easily understood by merely watching how many steps they have passed.

The Wald test (1943) used in testing of goodness-of-fit of the whole model may not be the most appropriate approach. Although this test provides an easy way to analyze response residuals, an extreme misfit item or several inconsistently answering persons will inflate this statistic, because the fit statistic of the whole model is the sum of fit statistics of all items or all persons. One item wiht an extremely large chi-square value will easily explode such a summation and make the fit statistic of the whole model significant. Consequently, a significant lack-of-fit statistic showing that the two-parameter partial credit model does not fit the dataset may be quite misleading.

If we modify such an extremely bad item and readminister such a test to the same sample, the fit statistic may show that this model still fits the dataset. Therefore, when we use this fit statistic to check the model-fitness, we have to consider the contribution of fit statistics from other items to the whole at the same time If the lack-of-fit statistic for the whole model is significant and some items also have significant lack-of-fit statistics, or many persons with unusual responses and significant lack-of-fit statistics, then we may say that this model does not fit the dataset. Otherwise, we have to trace back which items have large lack-of-fit statistics, delete or revise them and readminister them to the same persons and see what happen to the whole model. If the lack-of-fit statistic is still significant, we may say that this model really does not fit the dataset. Revising the bad items and readministering such a test should be suggested in fitting the model to the dataset.

Finally, one warning that should be mentioned here regards to the method used for creating the 16 items. The component scoring method, often used in modern analysis of human cognitive abilities with a latent trait model (for example, Andrich, 1985), treating combined items as subtests or components may suffer from two weaknesses that violate the basic assumptions of item response theory. One is that there are several ways to combine items into a subtest or a component. No way is certain to be better than others. The other is that, unless theories or reasons are provided, no ways guarantee which combination will not violate the independence assumption. Since several items are combined together to be a subtest or a component, each component score will depend on how many original successful items there are. Hence, any possible component score (ranging from 0 to $m_i$, where $m_i$ is the number of steps) may not be exclusively and independently determined for each item.

Fortunately, each of the original 48 items were equally weighted to measure a person's proficiency in algebra. Hence, three successive items summed to be a "big item" that purport to measure the amount of a latent trait (that is, component) are reasonable for creating the current dataset. Factor analysis also shows that this test is dominated by a major factor. Thus, the basic assumptions of item response theory hold in the application of the two-parameter parrtial credit model to the current dataset. However, we still have to be aware of the basic assumptions of item response theory, no matter how this model fits the dataset or not.

## VI. Summary and Conclusion

The assessment of partial credit has a long-term history in the literature of psychometrics. Many solutions based on classical test theory have been suggested and proposed. Thereafter, due to their failure in satisfying rigorous theory background and precise estimation, new methodologies based on modern test theory are currently being examined.

Since Masters (1982) proposed a Rasch-type partial credit model, it became the best-known model to score persons' partial knowledge. Unfortunately, several weaknesses criticized above may be occurring in real testing situations too. Hence, the present author takes the step discrimination into account in Masters' partial credit model and expands it to the two-parameter partial credit model as follows:

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

The maximum likelihood estimation solutions and a FORTRAN 77 computer program, TPPCM, are described and used to estimate such model parameters. An illustrative example is also shown and analyzed to show how model parameters are calibrated, goodnesses-of-fit are tested, and information functions are provided. Hence the major purposes of this research are accomplished.

From the analysis and discussion of this illustrative example, four conclusions can be drawn from the findings of this research as follows.

1. The existence of the two-parameter partial credit model is confirmed. This model becomes an alternative model to score persons' parrtial knowledge or calibrate any questionnaire or test with ordered-response formats.

2. Step discriminations provide a good help in partitioning persons' performance levels, hence the fitnesses of person ability estimates on the map of performance space are easily spotted.

3. Step information functions are uniquely and differently determined from step discriminations of each item. Hence it implies potentials for item and test design, selection, and construction.

4. The two-parameter partial credit model shares the same features of the one-parameter partial credit model, except that of specific objectivity and parameter separability.

## 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.

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 extention of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika*, *47*, 105-113.

Andrich, D. (1985). A latent-trait model for items with response dependencies: Implications for test construction and analysis. In S.E. Embretson (Ed.), *Test design: Development in psychology and psychometrics* (pp. 245-275). Orlando, FL: Academic Press.

Atkinson, L.V., Harley, P.J., & Hudson, J.D. (1989). *Numerical methods with FORTRAN 77: A practical introduction*. Wokingham, England: Addison-Wesley.

Baker, F.B. (1985). *The basic of item response theory*. Portsmouth, NH: Heinemann.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord., & M.R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

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

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

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

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

Davis, F.B. (1967). A note on the correction for chance success. *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 Research, 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.

Gagné, R.M. (1962). The acquisition of knowledge. *Psychological Review, 69,* 335-365.

Gagné, R.M., & Paradise, N.E. (1961). Abilities and learning sets in knowledge acquisition. *Psychological Monograph, 75,* 1-23.

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.

Hambleton, R.K. (Ed.). (1983). *Applications of item response theory.* Vancouver, British Columbia: Educational Research Institute of British Columbia.

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

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 usefulness 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.

Koffman, E.B., & Friedman, F.L. (1990). *Problem solving and structured programming in FORTRAN 77* (4th ed.). Reading, MA: Addison-Wesley.

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.

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

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.

Loard, F.M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Ludlow, L.H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45,* 851-859.

Ludlow, L.H. (1986). Graphical analysis of item response theory residuals. *Applied Psychological Measurement, 10,* 217-229.

Ludlow, L.H., & Hillocks, G. Jr. (1985). Psychometric considerations in the analysis of reading skill hierarchies. *Journal of Experimental Education, 54,* 15-21,

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 comparision 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. (1980). *Probability models for some intelligence and attainment tests.* Chicago: The University of Chicago Press. (Original edition published in 1960).

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 Educational Measurement, 5,* 211-215.

Sabers, D.L., & White, G.W. (1969). The effect of differential weighting of individual item responses 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. *Psychological Monograph Supplement, No. 17.*

Samejima, F. (1973). Homogenerous case of the continuous response model. *Psychometrika, 38*, 203-219.

Sax, G., & Collect, L. (1968). The effects of differing instructions and guessing formulas on reliability and validity. *Educational and Psychological Measurement, 28*, 1127-1136.

Smith, R.M. (1987). Assessing partial knowledge in vocabulary. *Journal of Educational Measurement, 24*, 217-231.

Stanley, J.C., & Wang, M.D. (1986). *Differential weighting: A survey of methods and empirical studies.* New York: College Entrance Examination Board.

Thissen, D. (1976). Information in wrong responses to Raven progressive matrices. *Journal of Educational Measurement, 13*, 201-214.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transcations of the American Mathematical Society, 54*, 426-482.

Wang, M.D., & Stanley, J.C. (1970). Differential weighting: A review of methods and empricial studies. *Review of Educational Research, 40*, 663-705.

Waller, M.I. (1981). A procedure for comparing logistic latent trait models. *Journal of Educational Measurement, 18*, 119-125.

Wingersky, M.S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R.K. Hambleton (Ed.), *Applications of item response theory* (pp. 45-56). Vancouver, British Columbia: Educational Research Institute of British Columbia.

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

Wright, B.D. (1988). The efficacy of unconditional maximum likelihood bias correction: Comment on Jansen, van den Wollenberg, and Wierda. *Applied Psychological Measurement, 12*, 315-318.

Wright, B.D., & Douglas, G.A. (1977a). Conditional versus unconditional procedures for sample-free item 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., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

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

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

Yu, M. (1991b). *TPPCM: A FORTAN 77 computer program for the two-parameter partial credit model* (unpublished).

Yu, M. (1991c). The assessment of partial knowledge. *Journal of National Chengchi University, 63*, 401-428.