SCORING TESTS WITH DICHOTOMOUS AND POLYTOMOUS ITEMS

by

CIGDEM ALAGOZ

(Under the Direction of Seock-Ho Kim)

ABSTRACT

This study applies item response theory methods to the tests combining multiple-choice (MC) and constructed response (CR) item types. Issues discussed include the following: 1) the selection of the best fitting model from the most widely used three combinations of item response models; 2) the estimation of ability and item parameters; 3) the potential loss of information from both simultaneous and separate calibration runs. Empirical results are presented from a mathematics achievement test that includes both item types. Both two-parameter logistic (2PL) and three-parameter logistic (3PL) models fit to the data better than the one-parameter logistic (1PL) model for the MC items. Both graded response (GR) and generalized partial credit (GPC) models fit better to the CR items than the partial credit (PC) model. The 2PL&GR and 3PL&GPC model combinations provided better fit than did the 1PL&PC. Item and ability parameter estimates from separate and simultaneous calibration runs across various models were highly consistent. Calibrating the MC and CR items together or separately did not cause information loss. Use of the CR items in the test increased reliability. Simultaneous calibration of the MC and CR items provided consistent estimates and an implicitly weighted ability measure.

INDEX WORDS: Item response theory, Test scoring, Mixed item types, Dichotomous and polytomous items.

SCORING TESTS WITH DICHOTOMOUS AND POLYTHOMOUS ITEMS

by

CIGDEM ALAGOZ

B.A., Gazi University, Turkey 2000

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment

of the Requirements for the Degree

MASTER OF ARTS

ATHENS, GEORGIA

2005

SCORING TESTS WITH DICHOTOMOUS AND POLYTHOMOUS ITEMS

by

CIGDEM ALAGOZ

Approved:

Major Professor: Seock-Ho Kim

Committee: Deborah Bandalos

Stephen Olejnik

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2005

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION AND ITEM RESPONSE THEORY MODELS

Introduction

This study applies item response theory methods to the tests combining multiple-choice (MC) and constructed-response (CR) item types. Issues discussed are the selection of the best fitting model from the most widely used three combinations of item response models and the potential information loss from the simultaneous and separate calibration runs. Empirical results are presented from a mathematics achievement test that includes both item types.

The MC and CR items are being used in many testing situations to complement each other and enhance the reliability and validity of test scores. The MC items are preferred to reduce the cost and time of the measurement. Besides, they may enhance the reliability and validity of the test. On the other hand, the CR items are thought to be more appropriate for measuring certain skills that require different levels of cognitive process. The CR items in that case can be used to establish possibly better construct validity.

In this study, the MC and CR items measure the same overall construct of mathematics ability, but somewhat different levels of the ability. Earlier studies presented whether it is better to combine these two types of items to create a common scale (Ercikan, Schwarz, Julian, Burket, Weber, & Link, 1998) and if it is better to weight them to create a scale (Wainer & Thissen, 1993).

When the unidimensionality assumption holds, item response theory solves the problem of assigning weights to these two item types. The inquiry of the unidimensionality can be performed by a factor analytic method in item response theory. As long as the MC and CR items are believed to measure the same overall construct, the question of giving more weight to one type of item depends on judgment. An explicit weighting procedure may not be a requirement in item response theory. Calibrating the MC and CR items together may yield the ability scale that reflects the implicit weights of these two parts. In this study the ability scale from the simultaneous calibration is compared with that from the separate calibration of the respective MC and CR items. Item, ability parameters, and information functions are compared. It has been reported that loss of information of CR items might occur from simultaneous calibration. This study investigates such loss of information from simultaneous calibration. The rest of this chapter presents definitions of the relevant concepts and models under item response theory and classical test theory with an emphasis on item response theory models for both dichotomously scored items and polytomously scored items.

Measurement

Measurement is defined as the act of assigning numbers or symbols to characteristics of objects according to rules (Lord & Novick, 1968). In measurement settings in education, there may exist unobservable, latent variables that we are particularly interested in, such as achievement, reading ability, mathematics ability, intelligence, and aptitude. Such variables cannot be measured directly since they are constructs rather than physical quantities. However, their attributes can be described and listed with the guidance of theories of the relevant domains. Educational and psychological measurement is concerned with assigning numbers to these latent traits. Classical test theory and item response theories assign numbers to characteristics of examinees using different procedures.

Classical Test Theory

Concepts from classical test theory and item response theory are generally comparable. Each theory tries to explain the latent variable with distinctive models. Classical test theory uses the model

$$X = T + E, \qquad (1)$$

where X is the observed score, E is the random error component and T is the true score. The true score for examinee j is defined as

$$T_j = \varepsilon\,(X_j) = \mu_{Xj}\,. \qquad (2)$$

Total variance in a distribution of observed scores is equal to the sum of the true variance plus the error variance:

$$\sigma^2{}_X = \sigma^2{}_T + \sigma^2{}_E\,, \qquad (3)$$

where $\sigma^2{}_X$ is the variance of the observed score, $\sigma^2{}_T$ is the variance of the true score, and $\sigma^2{}_E$ is the variance of the error scores.

Reliability refers to the accuracy, dependability, consistency, or repeatability of test results. In other words, it refers to the degree to which test scores are free of measurement errors. The reliability index can be expressed as the ratio of the standard deviation of the true score to the standard deviation of the observed score:

$$\rho\,(XT) = \frac{\sigma_T}{\sigma_X}\,. \qquad (4)$$

The reliability coefficient is defined to be the square of the reliability index. The coefficient is then the ratio of the variance of the true score to the variance of the observed score. Since the true score is not observable, one way to estimate the reliability coefficient is to obtain the Pearson product moment correlation between the observed scores from two parallel tests, X and X', where $T_j = T_j'$ for all; and $\sigma^2{}_X = \sigma^2{}_{X'}$.

Other basic concepts applicable to most test administration situations are item difficulty and item discrimination. The classical item difficulty for a dichotomously scored item is defined as the proportion of examinees who answered the item correctly (Crocker & Algina, 1986). Both biserial and point-biserial correlations between a given item score and the total score can be used as the classical item discrimination indices. For a polytomously-scored item, the average item score can be viewed as the item difficulty and the correlation between the item score and the total score can be used as an item discrimination index.

Item Response Theory Models

Having mentioned some basic measurement concepts from the classical test theory point of view, their counterparts from the item response theory are now presented below, first for the dichotomous items and second for the polytomous items.

The outcome of measurement under item response theory is a scale to which examinees as well as items are placed. In that sense, it is necessary to have a scale of measurement. Since we do not have the exact image of the latent variable, scaling is a difficult task. To overcome this problem, it is generally assumed that the ability scale has a midpoint zero, a unit of measurement of one, and a range from negative infinity to positive infinity(Baker, 2001).

The item characteristic curve (ICC) is the one central concept of item response theory. For a dichotomous item, plotted function corresponds ability levels to their probability of responding that item correctly. Each item in a test has its own ICC. Two widely used forms of these ICCs are the normal ogive model and the logistic ogive model. Ogive curves are any continuous cumulative frequency curves. In item response theory literature, these are the cumulative forms of the normal and logistic functions. The normal ogive model has the probability function of:

$$P_i(\theta_j) = P_i(\mu_i, \sigma_i, \theta_j) = \int_{-Z_{ij}=-(\theta_j-\mu_j)/\sigma_i}^{\infty} \frac{1}{\sqrt{2\pi}} e^{(-1/2)z^2} dz \,, \tag{5}$$

$$\text{where } Z_{ij} = \left(\frac{\theta_j - \mu_i}{\sigma_i} \right) = \alpha_i(\theta_j - \beta_i) \,. \tag{6}$$

β_i is defined as the point on the ability scale at which the probability of correct response is .5; this corresponds to the mean of the normal ogive, $\mu_i = \beta_i$. The parameter α_i is the steepness of the item characteristic curve. The σ_i is a measure of the spread of normal distribution. When σ_i is large, normal ogive is not very steep but more flat near β_i (Baker, 1992).

The second model for the ICC is the logistic ogive model, which has the probability function of

$$P_i(\theta_j) = P_i(\alpha_i{*}, \beta_i, \theta_j) = \Psi(Z_{ij}) = \frac{e^{Z_{ij}}}{1+e^{Z_{ij}}} = \frac{1}{1+e^{-Z_{ij}}} \;, \tag{7}$$

where $Z_{ij} = \alpha_i{*}(\theta - \beta_i)$. β_i is the location parameter and is the point on the ability scale at which $P_i(\theta_j) = .5$. $\alpha_i{*}$ is the discrimination parameter and is the reciprocal of the standard deviation of the logistic function. The form of the logistic ogive is very similar to the normal ogive. If the logistic deviate $Z_{ij}=\alpha_i^*(\theta_j-\beta_i)$, is multiplied by 1.702 and entered in the probability function of logistic ogive model, the absolute difference between $P_i(\theta_j)$ of the normal ogive and $P_i(\theta_j)$ of the logistic ogive is less than .01 over the full range of θ (Haley, 1952).

Note that the two models for the ICC can be further extended to more restricted and more general cases. The three special cases of the logistic ogive model for the dichotomous items will be presented in detail below, under the heading “ Dichotomous Scoring”.

Dichotomous Scoring

Under the logistic ogive model, probability of correct response is plotted as a function of ability. Three widely used dichotomous scoring models can be constructed with three parameters, namely: difficulty, discrimination, and guessing. The difficulty of an item describes where the item functions along the ability scale. An easy item functions among the low ability levels, whereas a difficult item functions among the high ability examinees. Discrimination describes how well the item can differentiate between examinees having abilities below and above the item difficulty. Discrimination reflects the steepness of the item characteristic curve where item functions. The steeper the curve the more the item discriminates, whereas the flatter the curve, the less the item discriminates. When these two parameters are employed, the ICC is asymptotic to the $P_j(\theta) = 0$ and $P_j(\theta) = 1$ lines. In some cases, it is observed that the lower tail of the ICC is asymptotic to a value greater than zero. This can be interpreted as the guessing parameter. It is noted that this parameter is lower than the chance level $1/m$, where m is the number of response alternatives in a multiple-choice item. When the guessing parameter is introduced to the probability function, the equation obtained is:

$P_i(\theta_j)= c_i + (1- c_j)\, \Psi[\alpha_i(\theta_j-\beta_i)]$, where c_i is the asymptotic probability of correct response. The above equation is the three-parameter model.

Rasch model, which is the one-parameter logistic model, uses only item difficulty to define the ICC:

$$P_i(\theta_j)= \frac{e^{(\theta_j-\beta_i)}}{1+e^{(\theta_j-\beta_i)}} = \frac{1}{1+e^{-(\theta_j-\beta_i)}}, \qquad (8)$$

where β_i is the difficulty parameter, and θ_j is the ability parameter.

The two-parameter logistic model employs the difficulty and discrimination parameters to define the ICC:

$$P_i(\theta_j) = \frac{e^{\alpha_i(\theta_j-\beta_i)}}{1+e^{\alpha_i(\theta_j-\beta_i)}} = \frac{1}{1+e^{-\alpha_i(\theta_j-\beta_i)}}, \quad (9)$$

where α_i is the discrimination parameter, β_i is the difficulty parameter, and θ_j is the ability parameter.

Again the three-parameter logistic model adds the guessing parameter to the definition of the ICC:

$$P_i(\theta_j) = c_i + (1-c_i)\frac{1}{1+e^{-\alpha_i(\theta_j-\beta_i)}}, \quad (10)$$

where c_i is the guessing parameters, α_i is the discrimination parameter, β_i is the difficulty parameter, and θ_j is the ability parameter.

Estimation of Parameters

A brief description of how item and ability parameters under dichotomous models are estimated is described herein. Although the estimation implemented in the computer program used in the empirical comparisons is the method of marginal maximum likelihood, separate estimation of item and ability parameters under the two-parameter logistic model will be briefly presented in this section using the group data instead of per individual person for ease of understanding. First, item parameter estimation is presented using the maximum likelihood and assuming the ability is known. Then, assuming item parameters are known, ability parameters are estimated.

Using the maximum likelihood procedure, the item characteristics of difficulty and discrimination are estimated. The multiplicative law of probability, if A and B are two independent events, $P(A \cap B) = P(A) * P(B)$ is used. Use of this theorem requires the assumption of the two events to be independent--namely an examinee responding to one item is independent from his/her responding to another item. $R=(r_1, r_2, \ldots, r_k)$ -- the vector of the observed number of correct responses at ability θ_j $(j=1,2,\ldots,k)$, binomially distributed with parameters f_j, and P_j, where P_j is the true probability of correct response and f_j is the frequency of subjects having ability θ_j, r_j is the number of examinees who answered the item correctly and f_j-r_j is the frequency of incorrect responses.

The observed proportion of correct response at ability θ_j is

$$p(\theta_j) = p_j = \frac{r_j}{f_j}. \tag{12}$$

Then, given the ability groups, the probability of R is given with the likelihood function:

$$\text{Prob}(R) = \prod_{j=1}^{k} \frac{f_j!}{r_j!(f_j - r_j)!} \quad P_j^{r_j} Q_j^{f_j - r_j} \tag{13}$$

Using the log (Prob(R)) = L is computationally advantageous because the parameter values maximizing Prob(R) will also maximize log (Prob(R)). Thus, log likelihood function is preferred over the likelihood function. Derivatives of L with respect to parameters will be zero for the parameter values, which will maximize L.

Solutions to these equations, which equal derivations with respect to parameters to zero, will give the item parameters. These equations cannot be solved directly. An iterative procedure based on Taylor series can be employed. Iterations are repeated until the difference of parameters estimated from $(t+1)$th and tth iterations is small enough:

$$\hat{\zeta}_{t+1} = \hat{\zeta}_t + \Delta\hat{\zeta}_t \tag{14}$$

$$\hat{\lambda}_{t+1} = \hat{\lambda}_t + \Delta\hat{\lambda}_t,$$

where $\Delta\hat{\zeta}_t$ and $\Delta\hat{\lambda}_t$ are the increments which we want to make as small as possible.

From the method of maximum likelihood, resulting estimates are generally unbiased (e.g. $E(\hat{\zeta}) = \zeta$); consistent (i.e. larger sample size corresponds with a better estimate), efficient (i.e. a small variance for the sampling distribution); and sufficient (in that it uses all the sample information).

It was assumed that the ability parameters are known when item parameters are obtained. The next part is to estimate ability parameters assuming that the item parameter estimates are the true values.

Maximum likelihood estimation of ability parameters makes the assumption that examinees are independent objects. Probability of the vector of item responses U_j for examinee j is given by the likelihood function,

$$\text{Prob}[U_j|\theta] = \prod P_i^{u_{ij}}(\theta_j) Q_i^{1-u_{ij}}(\theta_j),$$

where u_{ij} is the observed response examinee j to item i. An ability estimate, which minimizes this function, is obtained by an iterative solution procedure that is similar to the item parameter estimation.

<u>A Chi-Square Fit Statistic</u>

Pearson χ^2 is defined as

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} , \qquad (15)$$

where k is the number of categories, O_j the observed frequency, and E_j the expected frequency. We can use the same chi-square to assess the fit of an item.

Using the expected and observed correct response frequencies at ability level θ_j, χ^2 can be obtained as

$$\chi^2 = \sum_{j=1}^{k} \frac{f_j}{P_j Q_j} (p_j - P_j)^2 \qquad (16)$$

with *df = (k-2)* where *k* is the number of score groups, f_j is the frequency of response for ability level θ_j, $p_j = \frac{r_j}{f_j}$, r_j is the frequency of correct response for ability level θ_j, and P_j is obtained from the ICC.

Information Functions

Information functions reflect how well the individual items and the test as a whole estimates the ability over the scale. Since the variance is the measure of precision of measurement, the test information function is considered as the reliability coefficient in the item response theory context. Maximum likelihood estimator, $\hat{\theta}$ has normal asymptotic distribution with mean θ and variance $\sigma^2 = \frac{1}{I(\theta)}$, where $I(\theta)$ is the amount of information. When the variance of an estimator is large, the estimate of ability is less precise and the available information about an examinee's ability will also be less.

The information function for the test with *n* items is defined as

$$I(\theta) = \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad , \tag{18}$$

where $P_i(\theta)$ is obtained by evaluating the item characteristic curve at θ and $P_i'(\theta) = \frac{\partial P_i}{\partial \theta}$.

The item information is the decomposition of test information into each item and given as

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \tag{19}$$

where $P_i'(\theta) = \frac{\partial P_i}{\partial \theta}$.

The reliability coefficient and the item reliability in classical test theory can be seen as the counterparts of the test and item information functions.

Likelihood Ratio Tests for Goodness of Fit

A maximum of the likelihood function is computed and used as an index of fit. The test statistic is

$$Q = \quad -2\log(LR)\,, \tag{20}$$

where the *LR* stands for the likelihood function. The test statistic can be used to compare the relative fit of the model to data.

Polytomous Scoring

The previous sections presented models and important concepts for the data that consist of dichotomously scored items. In this section three widely used IRT models for polytomous scoring are going to be described which are the graded response, the partial credit, and the generalized partial credit models.

Graded Response Model

The hypothetical item variable scale is divided into categories under the graded scoring procedure. The lowest category contributes the lowest and the highest category contributes the highest to the test score. For item i, k= 1,2,…, m_i, where m_i is the number of response categories for item i. m_i can take on different values for different items, and also dichotomous response model is included in the graded response model when m_i = 2. U_{jik} representing the response to *ith* item, when *jth* examinee gives a response to *ith* item and category k, 1 is assigned to u_{jik} and 0 otherwise. The sum of all probabilities is

$$\sum_{k=1}^{m_i} P_k(\theta) = 1\,, \tag{21}$$

where $P_k(\theta)$ is the probability of an examinee's response falling into category k.

The probability of an examinee of ability θ selecting the item response category k is $P_k(\theta)$ and $\sum_{k=1}^{m_i} P_k(\theta) = 1$. Therefore,

$$P_{m_i}(\theta) = 1 - \sum_{k=1}^{m-1} P_k(\theta). \tag{22}$$

This restriction of the sum of probabilities results in fewer degrees of freedom than the item location parameters. This problem is solved with the introduction of the boundary curves. The probability of selecting a response category is defined by the boundaries on the probabilities:

$$P*_{m_i}(\theta) = 0$$

$$P*_{m_i-1}(\theta) = 0 + P_m(\theta)$$

$$P*_{m_i-2}(\theta) = P_{m_i-1}(\theta) + P_{m_i}(\theta)$$

$$\vdots$$

$$P_k^*(\theta) = \sum_{v=k+1}^{m_i} P_v(\theta) \tag{23}$$

$$\vdots$$

$$P_1^*(\theta) = \sum_{k=2}^{m_i} P_k(\theta)$$

$$P_0^*(\theta) = 1$$

Then $P_k\ (\theta) = P_{k-1}^*(\theta) - P_k^*(\theta)$. (24)

Given four response categories it assumes three non-trivial parallel boundary response functions. The first is the probability of choosing categories 2, 3, or 4 over category 1. The second is the probability of choosing categories 3 or 4 over 1 or 2. The third is the probability of choosing category 4 over 1, 2, or 3. Figure 1 shows the boundary response functions. The exact formula for the boundary response function is presented below under the logistic model.

Figure1. Boundary response functions

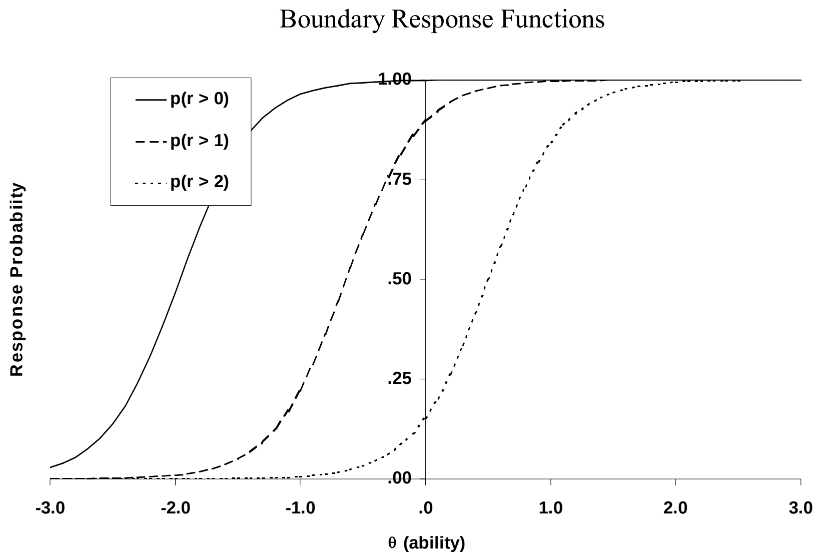
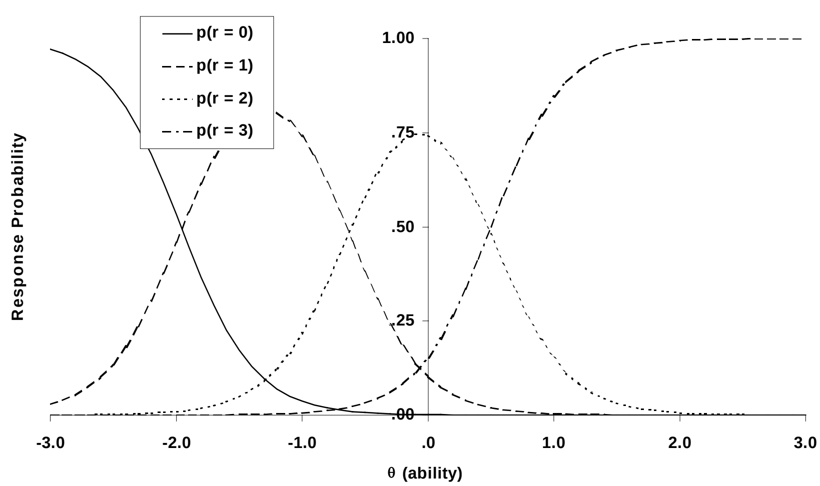
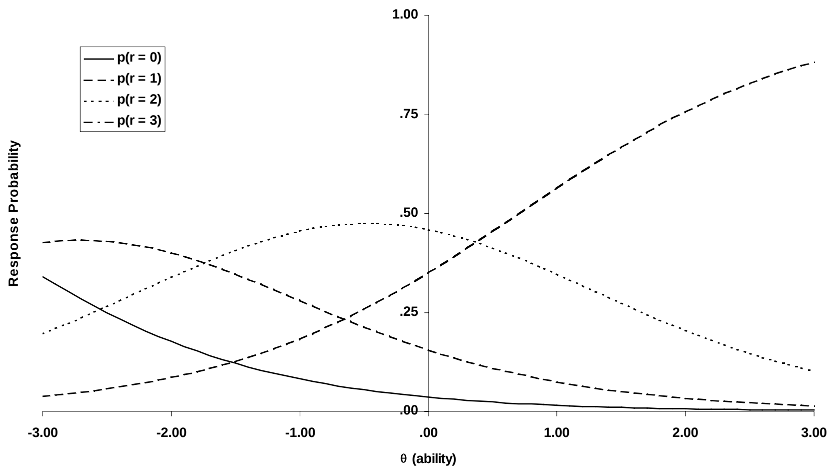


Figure 2. Item response characteristic curve



The relationship among the response categories can be seen from Figure 2. At the lowest ability levels, the lowest category has the highest probability. Moving from the lowest ability to higher ability levels, the higher categories show higher probabilities whereas the lower categories show lower probabilities. From Figures 2 and 3, a visual comparison of the high and low discrimination of ability can be made.

Figure3. Item response characteristic curves for a less discriminating domain



The boundary response curves are given with the equations

$$P_{i,k-1}^{\ *} = \frac{1}{1+e^{-(\zeta_{i,k-1}+\lambda_i\theta)}} \tag{25}$$

and

$$P_{i,k}^{\ *} = \frac{1}{1+e^{-(\zeta_{i,k}+\lambda_i\theta)}}\ , \tag{26}$$

where the linearized logit is $a_i(\theta-\beta_{ik})=\lambda_i\theta+\zeta_{ik}$. Item parameters are to be estimated with a likelihood function of the whole observed response matrix in a similar way for the dichotomous items.

Likelihood for ability with the response pattern $U_l = (u_{l1}, u_{l2}, \ldots, u_{ln})$ is given as

$$L_l(\theta) = P(U_l|\theta) = \prod_{i=1}^{n} P(U_{li} = k|\theta).$$

It can be approximated with the Fisher-scoring equation;

$$\hat{\theta}_{j(t+1)} = \hat{\theta}_{j(t)} - \left[\frac{\partial L / \partial \theta}{\partial^2 L / \partial \theta_j^2}\right]_{(t)}, \tag{27}$$

where $L=L_l(\theta)$and t denotes the iteration. For each examinee there will be such an equation. The whole process would be repeated until the convergence criterion is met, assuming the item parameters are known values.

The Nominal Response Model

The nominal response model, with some restriction, will give the partial credit model and the generalized partial credit models. Using these restrictions to the nominal response model, the computer program MULTILOG (Thissen, 1991) estimates parameters with the partial credit model and the generalized partial credit models. That is why first the nominal response model will be presented next. Then, the other two models that are derived from this model will be explained.

Under nominal scoring, possible responses are allocated to m non-ordered categories. As in the graded response case, item response characteristic curves are obtained for each response category with the probability function

$$P_{jk}(\theta) = \frac{\exp[Da_{jk}(\theta - b_{jk})]}{\sum_{u=1}^{m_j} \exp[Da_{ju}(\theta - b_{ju})]}, \tag{28}$$

where D is the scaling constant 1.7, $a_{jk}(\theta - b_{jk}) = \zeta_k + \lambda_k(\theta_j)$. $P_{jk}(\theta)$ is the probability of an examinee of ability θ_j choosing item response category k.

The Generalized Partial Credit Model

The generalized partial credit model (GPCM) assumes that the probability of obtaining score k over the probability of obtaining score k or $k\text{-}1$ fits a dichotomous model. This can be expressed as

$$\frac{P_{jk}(\theta)}{P_{j(k-1)} + P_{jk}(\theta)} = \frac{1}{1+\exp[-Da_j(\theta - b_{jk})]}, \tag{29}$$

Resulting event is that given that the score is $k\text{-}1$ or k, 2 –parameter dichotomous model is fitted to obtaining a score of k.

$$\sum_{k=1}^{m} P_{jk}(\theta) = 1,$$

where m is the number of response or score categories.

Nominal model representation of probability equation becomes GPCM when a_{jk} is replaced by $T_{jk}a_j$;

$$P_{jk}(\theta) = \frac{\exp[DT_{jk}a_{jk}(\theta - b_{jk})]}{\sum_{u=1}^{m_j}\exp[DT_{ju}a_{ju}(\theta - b_{ju})]}, \tag{30}$$

with the requirement that T_j must be a linear vector. T_j is called the scoring function.

Partial Credit Model

Partial credit model is given by

$$P_{jk}(\ \theta) = \frac{\exp(DT_{jk}\theta - \sum_{r=1}^{s} b_{ir})}{\sum_{u=1}^{m}\exp(DT_{jk}\theta - \sum_{r=1}^{u} b_{ir})}, \tag{31}$$

Generalized partial credit with the restriction of $a_{jk}=1$ gives the partial credit model. If the number of response categories is two, then the partial credit model is equivalent to 1PL model.

Partial Credit and Generalized Partial Credit Models with MULTILOG

In MULTILOG, GPCM is expressed as

$$P_{jk}(\theta) = \frac{\exp[Da_{jk}\theta + c_{jk}]}{\sum_{u=1}^{m_j} \exp[Da_{ju}\theta + c_{ju})]}. \tag{32}$$

This parameterization is similar to Equation 30, but c_{jk}=-$a_{jk}b_{jk}$, and the scoring function, T_j is handled by using contrasts. Estimation of item parameters requires estimation of contrasts among the parameters. TMATRIX and FIX commands must be included in the command file to specify these contrasts. The linear scoring function for the a_{jk} parameters is achieved by specifying a polynomial T matrix for those parameters, with the quadratic and higher contrasts fixed at zero. The linear contrasts that are left then serve as the scoring function. Contrasts for c_{jk} parameters must also be specified. Any contrasts can be used for these parameters. For the the generalized partial credit model, MULTILOG forces D=1, which means it uses the logistic instead of the normal ogive scale (Childs & Chen, 1999).

To get the estimations under the partial credit model with MULTILOG, using the nominal response model slopes are constrained to be equal across items with the EQUAL command and polynomial contrasts that are used for the a_k parameters.

Contrasts are $a' = \alpha' T_{a'}$, $c' = \gamma' T_{c'}$ and $d^{*\prime} = \delta' T_d$ in the MULTILOG where T matrices consists of the deviation contrasts. These matrices are presented in the MULTILOG manual (see Thissen, 1991 & chapter1, pp16-20).

Scoring the Combination of Multiple Choice and Open-Ended Items

Employing the item response theory models, simultaneous scoring of the combination of dichotomous and polytomous response items does not involve an explicit weighting of the two parts of the test. Assuming the unidimensional scale for ability, item response theory models estimate item and ability parameters with information from two sets of items. The one-parameter item response theory model for multiple-choice items and Master's partial credit model for constructed response items utilize the unweighted sum of the item responses as the basis of ability estimate (Sykes & Hou, 2003). Item response theory models, in this sense, do not weight two parts explicitly. Sykes and Hou (2003) compared test and conditional score reliabilities from implicit weighting and three types of explicit weighting. The empirical result from writing in grade 8 data showed that the implicitly weighted scale scores had the smallest standard errors compared to any explicitly weighted scale scores.

Billeaud et al (1997), however, proposed an explicit weighting method to combine multiple-choice and constructed response item scores, which involves a hybridization of summed-score and response-pattern computation of scaled scores. First, all of the items are calibrated together with the appropriate item response theory model for each item. Then the likelihood for summed score x for multiple-choice section , $L_x^{MC}(\theta)$, and the likelihood for summed score x' for the open-ended section, $L_{x'}^{OE}(\theta)$, are computed. Subsequently, the likelihood for each combination of a given summed score x on the multiple-choice section with any summed score x' on the open-ended section, $L_{xx'}(\theta) = L_x^{MC}(\theta)L_{x'}^{OE}(\theta)$, is computed. Then with the equation,

$$P_{xx'} = \int L_{xx'}(\theta)\phi(\theta)d\theta\,, \tag{33}$$

where $\phi(\theta)$ is the ability population density, the modeled probability of the response pattern of summed scores is computed. Given the response pattern of summed scores, the expected value of θ (i.e., the expected a posteriori, EAP, estimate) is computed as

$$EAP(\theta | x, x') = \frac{\int \theta L_{xx'}(\theta)\phi(\theta)d\theta}{P_{xx'}}. \tag{34}$$

Billeaud et al (1997) showed the use of IRT scale scores for patterns of summed scores. They calibrated items under the combination of the three-parameter model and the graded response model.

CHAPTER2

LITERATURE REVIEW

Use of Polytomous Items

It is claimed that the use of open-ended items enables test designers to measure skills that cannot be measured by multiple-choice items (Davis, 1992). Given that the use of open-ended items is getting increased demand by the test constructers, and the tests combining these items with multiple choice items are being administered more often, problems and issues regarding this testing situation are studied from different aspects by researchers.

Unidimensionality assumption of the IRT models, which brings the mathematical complexity of the model within reasonable bounds is investigated by many researchers. The assumption of unidimensionality is that only one ability or trait is necessary to explain or account for an examinee's test performance (Hambleton & Swaminathan, 1985). Unidimensionality does not have to be violated with only the addition of polytomous items. In educational practice, tests do not always satisfy the unidimensionality assumption of item response theory models. Studies have shown that when the unidimensionality assumption of dichotomous item response theory models is violated, the results from those analyses might not be valid (Folk & Green, 1989; Tuerlinckx & Boeck, 2001). Even when the unidimensionality assumption is violated, the test scoring under item response theory could be applied under some constraints for both dichotomous and polytomous items. One study with polytomous items showed that when the ability estimate was assumed to measure the average ability of two equally important abilities and the major ability of two unequally important abilities (75 percent of the total number of items), the procedure was generally robust to the violation. But, when the ability estimate was assumed to measure one of the two equally important abilities (dimensional strength is 50/50) as well as the minor ability of two unequally important abilities (25 percent of the total number of items), the estimation procedure was not robust (Dawadi, 1999).

Tests with polytomous items could be a more valid form of testing due to the richer format even though there is no evidence for. However, when it is the case that different item types measure different dimensions of proficiency, combining the scores from the separate parts of the test might have a negative impact, resulting in similar scores for examinees having different ability levels. However, when different item types are used to measure similar aspects of proficiency, combined scores are efficient and sensible.

Polytomous items are preferred because they are believed to increase the validity of the test. However, performance assessment with constructed response items is costly in testing time and scoring. Researchers have questioned if it is worth using such items. In one simulation study, their classification accuracy in computerized testing situation resulted in higher accuracy for polytomous items than the dichotomous items. Lower false negative and false positive classification error and the total error rates were reported for polytomous items than the dichotomous ones. The impact of test length constraint was smaller for polytomous items than dichotomous items (Lau & Wang, 1998).

Test score reporting of polytomous items is discussed by Samejima (1996) and using response pattern is suggested over the summed score. An advantage of the use of polytomous response items over the dichotomous ones is the increased test information. However, loss of test information by using the test score and in return the loss of accuracy in ability estimation is greater when responses are graded polytomously. Moreover, Samejima (1996) showed that the amount of test information would be increased when polytomous response categories are more finely classified, for example when using 7-point scale versus a 3-point scale. When the test score is an aggregation of response patterns, the amount of test information will be decreased if the test score is used as the basis of ability estimation, instead of the response pattern itself. Another empirical study reported that more information was observed for polytomous items than dichotomous ones (Donoghue, 1993).

In the computer adaptive testing context, the benefits of polytomous item response theory models are reported by several studies. These studies have shown that item pools smaller than those used with dichotomous model-based computer adaptive tests have resulted in satisfactory estimation (De Ayala, 1989; Dodd, Koch, & De Ayala, 1989).

Another issue raised by the use of polytomous items is the lower reliability. Desirable high reliability could be achieved by combining these items with the multiple-choice items. When the combinations of the summed score are used, there is a risk that the combination is less reliable than one of its components (Lukhele, Thissen, & Wainer, 1994). It is not a generalized result that the weighted combinations will have less reliability, but it is a possibility when the weighting is not well chosen.

Scoring of a Test with Dichotomous and Polytomous Items

Once it is decided that the use of polytomous items is desirable, magnitude is changed into the scoring procedures. Approaches to weight two components of the tests are discussed by Wainer and Thissen (1993). Under the subtitles of reliability weighting, item response theory weighting and validity weighting, weighting methods are defined. The problems with reliability weighting include the score scale of the polytomous items and the instability of estimated regression weights. It is shown that equal weights are often superior to estimated multiple regression weights, unless the sample is very large (Wilks, 1938). Estimation of the parameters of the item response models simultaneously solves the weighting and scaling problems because it places each item response to the latent variable scale. The problem with the use of item response theory to solve the weighting problem is the assumption of the item response theory models, namely, the unidimensionality issue. When a single, well-established observable validity criterion is available, weights that maximize the predictive validity of the test are chosen, called by Wainer and Thissen (1993) as criterion weighting. Criterion weighting solves the scaling problem simultaneously with the weighting problem. Item response theory provides validity weighting at the item level when a criterion is available (Wainer & Thissen, 1993).

While Wainer and Thissen (1993) did not choose one weighting method over another, Rudner (2001) evaluated alternative methods and presented formulas for composite reliability and validity as a function of component weights and suggested a process to determine weights. It is suggested to use judgment to optimize the weighting and determine the importance of reliability and validity.

Grima and Weichun (2002) scored a mathematics test with mixed item types and evaluated six different scaling methods. They calibrated dichotomous items with the three-parameter model and polytomous items with the generalized partial credit model. The methods they employed included calibrating all items simultaneously, or calibrating components, which were defined on some basis, such as the item type, judgment of experts, or factor analyses result. They reported reliabilities resulting from these methods and also they reported correlation analyses to compare score results from different methods. Grima and Weichun concluded that calibrating all items together resulted in the best fit to the model and it is the preferred approach.

Ercikan et al (1998) addressed the question of whether the multiple-choice items and the constructed-response items can be calibrated together. They assessed the appropriateness of calibrating those two types of items by examining the residuals of the test response data from the model. They pointed that the residuals would reflect the violation of assumptions because the deviations from the unidimensionality, local item dependence and fit introduce systematic variation in residuals. The loss of information due to the simultaneous calibration is discussed. It is noted that the constructed response items provide unique information about the examinees' abilities. Therefore the simultaneous calibration may cause the loss of information. Comparison of the results of the item, ability parameters and scores from separate and simultaneous calibrations is reported to assess the magnitude of the loss of information. Ercikan et al (1998) calibrated dichotomous items employing the three-parameter item response model, and the two-parameter partial credit model (i.e., the generalized partial credit model). For the reading, language, mathematics, and science domains, the mean item difficulty parameters and reliabilities from separate and simultaneous calibrations are reported. Item parameters from different methods are equated to make their comparison viable. They concluded that the calibration of items together did not result in model fit problems. Besides, investigation of local item dependence revealed that the dependence among items disappeared when items are calibrated together. The scores from various calibration methods are compared and correlations and mean score differences are also reported. Correlations among the homogenous components of the test and the whole test are also investigated and the correlation between the multiple choice part and the combination of items was higher for all domains as expected, because the number of multiple choice items are high and also common to the whole test.

Another issue in practice is the model selection, which is not only a concern to mixed tests, but to all item types and tests. The choice of models employed to calibrate items depends on the items and tests under consideration. After employing the model, available statistics can be used to test the goodness of fit of the model to the data. One of these tests, the likelihood ratio test, is a chi-square statistic and it is calculated as $-2\log L$, where L is the likelihood (Baker, 1992). The model, which gives the best fit, can be chosen based on the choice of fit statistic. Different methods for model selection are available under various estimation procedures with their corresponding statistics.

Comparisons of the models might be of interest because the purpose is to choose the best model and, consequently, obtain the best information about the examinees. In computer adaptive testing context, the partial credit and graded response models are compared in terms of their accuracy of the ability estimates (De Ayala, Dodd, & Koch, 1992). In the same study, robustness of the partial credit and graded response model-based ability estimation to the use of items, which did not fit these models, is also investigated. The authors used the likelihood ratio statistic for the model-fit investigation. Results showed that the partial credit computer adaptive test provided reasonably accurate ability estimation despite adaptive tests, which on the average contained up to 45% misfitting items. Graded response computer adaptive test provided slightly more accurate ability estimates than partial credit computer adaptive test (De Ayala, Dodd, & Koch, 1992).

CHAPTER 3

PROCEDURE

Instrumentation

This study used the 10th grade mathematics test of the Florida Comprehensive Assessment Test (FCAT). The test contains both multiple choice and constructed response items.

The purpose of the FCAT is to assess student achievement of high-order cognitive skills represented in the Sunshine State Standards in reading, mathematics, writing, and science. The Sunshine State Standards portion of the FCAT is a criterion-referenced test. The FCAT mathematics content scores are reported for five areas: (1) Number Sense and Operations, (2) Measurement, (3) Geometry and Spatial Sense, (4) Algebraic Thinking, and (5) Data Analysis and Probability. The three types of questions that included on the FCAT Mathematics are: multiple-choice questions, graded response questions, and performance tasks. Both multiple-choice and graded response questions are machine scored. Each answer to a performance task is scored holistically by at least two trained readers (http://www.firn.edu/doe/sas/fcat.htm). Each multiple-choice item had four response options. The maximum score range for constructed response items ranged between one and five. There were 26 multiple choice items, 15 short-answer items with two categories, four constructed response items with three categories, and two constructed response items with 6 categories on the test.

Sample

The data for the 10th grade FCAT mathematics consisted of 148,123 students from various ability levels. From the total group 1000 cases were randomly selected using SPSS and analyzed subsequently due to the limitation of the item response theory software. Several classical test theory statistics were investigated for sample as well as the total group. Students with limited English proficiency were removed from the data set before the random selection procedure. The total group included only the non-accommodated group of students for test.

The mean total score of the test for the sample group of 1000 was 29.98 (12.66), the parentheses contain the standard deviation, and that for the whole population mean was 29.39 (12.54). The mean total score of the constructed response items only yielded the similar means and standard deviations; 13.30 (7.83) for the population and 12.88 (7.76) for the sample. The mean total score of the multiple-choice item was 16.51(5.28) for the population and 16.68 (5.31) for the sample of 1000 students. Reliabilities for the whole group from the constructed-response (CR) items, the multiple-choice (MC) items and the combination of them were .878, .83, and .921 respectively. For the sample, they were .878, .834, and .922, respectively. The classical item difficulty values for the MC items were ranged from .36 to .84 for the population of students; they were ranged from .37 to .84 for the sample. The average values of the CR items ranges from .22 to 1,84 for the population; they range from .23 to 1.89 for the sample. The correlation between the total scores of the MC items and CR items was .842 for population; the correlation was .851 for the sample. In terms of the item statistics and other statistical characteristics, the sample seems to be similar to the population.

Computer Program

MULTILOG (Thissen, 1991) is a widely available item response theory software package that provides item and ability parameter estimates for polytomous models, as well as for dichotomous models. In MULTILOG it is possible to employ Samejima's (1969) graded response model, Master's (1982) partial credit model, and the generalized partial credit model for polytomous items and the one, two, and three-parameter models for dichotomous items. MULTILOG can be used for the nominal response model (Bock, 1972) and the multiple-choice model (Thissen & Steinberg, 1984). It employs the marginal maximum likelihood, using quadrature points and weights that approximate the density of the population ability distribution. In MULTILOG the normal distribution is used for the population ability distribution unless the user specifies the use of the Johnson curves to estimate the population distribution. For the ability estimation phase, the method of maximum likelihood is used, as well as two Bayesian methods known as the expected a posteriori (EAP) and the maximum a posteriori (MAP) methods. MULTILOG is compared with PARSCALE, another widely available software package for polytomous models (DeMars, 2002). Both programs yielded very similar item and trait parameter estimates, under the graded response model and the generalized partial credit model (DeMars, 2002).

CHAPTER 4

RESULTS

Dimensionality Investigation

Before analyzing data with MULTILOG, unidimensionality assumption was tested by performing factor analysis of data. The results from factor analyses indicated that data seems to be reasonably unidimensional because the first factor explained most of the total variance. Principal axis factoring extracted 10 factors with the eigenvalues greater than 1. The first factor explained 22.5 % of the total variance, the second factor added only 1.8% to explained variance, and the third explained 1%, and the other factors explained variances by the amounts ranging from .87 to .57 %. A confirmatory factor analysis was used to test the fit of the unidimensional structure to the data. The LISREL 8.54 software (Jöreskog & Sörbom, 2003) run indicated that most items loaded on a single factor, possibly called the general mathematics ability. Model fit indices indicated that the model fit was satisfactory. $\chi^2(1034) = 1603.5$ ($P < .05$) was significant, but because it is very powerful and the sample size is large, an investigation of model fit indices has been undertaken. RMSEA= .026 is smaller than the recommended cut-off value by Hu and Bentler (1999), which is less than .06. CFI= .99 index and also indicates the model with one factor fits to the data satisfactorily. The recommended cut off value for CFI by Hu and Bentler (1999) is .95 and the values closer to 1 are the indication of better fit.

First, MC and CR items were calibrated separately, with the one- (1PL), two- (2PL), and three-parameter logistic (3PL) models. The 27 MC items were calibrated with 1PL, 2PL and 3PL models with the computer program MULTILOG. The fit statistic was reported as 15579.0 (i.e. the negative twice the log likelihood). The 2PL produced a lower likelihood: (-2logL = 15298.7).

Fitting the 3PL to items provided the lowest likelihood (-2logL = 15211.7). The reliabilities available from MULTILOG were .82, .83, and .84 for the 1PL, 2PL and 3PL, respectively.

The second step was calibrating the CR items separately from the rest of the test. The partial credit (PC) model, graded response (GR) and generalized partial credit (GPC) models were employed. As a model fit index, the negative twice the log-likelihood values were obtained: 14380.9 for the PC, 14112.5 for the GR and 14106.8 for the GPC.

The third step was simultaneously calibrating the two types of items together. Combinations were made taking into account the number of parameters they estimate for the item response and the shapes of the boundary response curves. Three combinations were used: the 1PL and PC combination (1PL&PC), the 2PL and GR combination (2PL&GR), and the 3PL and GPC combination (3PL&GPC). Basically, the 1PL&PC fits the item response curve on the basis of the location (i.e. difficulty) parameters, the 2PL&GR combination estimates both location and discrimination parameters, and the 3PL&GPC combination estimates all three parameters of location, discrimination and possibly guessing. The values of the goodness of fit index, negative twice the log likelihood, were 42549.7, 41877.1, and 41708.7 for the 1PL&PC, 2PL&GR, and 3PL&GPC, respectively. Resulting reliabilities were all equal and .93 from the three combinations.

MC item parameter estimates are presented in the Table 1 from a separate calibration procedures. As can be seen from the Table 1, 23 of the items out of 26 functioned at the negative points of the ability scale when MC items were calibrated with 1PL and 2PL, whereas 16 of the items out of 26 functioned at the negative points of the ability scale when MC items were calibrated with 3PL.

Table 1.

Multiple-choice Item Parameter Estimates of the 1PL, 2PL and 3PL from Separate Calibrations

	Model				
	1PL&PC	2PL&GR		3PL&GPC		
Item	b_i [1]	a_i	b_i	a_i	b_i	c_i
1	-1.69	.48	-3.08	.30	-2.13	.26
2	-1.93	1.45	-1.54	.90	-1.27	.20
3	-1.28	.96	-1.32	.90	-0.23	.41
4	-1.27	1.97	-.89	1.50	-0.59	.20
5	-.38	.82	-.43	.70	0.30	.26
6	-.08	.94	-.07	1.00	0.52	.25
7	-.91	1.03	-.89	.70	-0.38	.23
8	-.06	.66	-.05	1.20	0.89	.35
9	-.63	2.00	-.45	1.40	-0.25	.13
10	.03	.58	.08	.40	0.73	.18
11	.00	1.42	.00	1.20	0.29	.15
12	-.57	.93	-.60	.90	0.23	.31
13	.62	.99	.65	1.30	0.90	.18
14	-1.37	1.14	-1.26	.80	-0.57	.32
15	-.77	.94	-.80	.60	-0.36	.19
16	-.53	1.06	-.50	.70	-0.06	.19
17	-1.92	1.25	-1.67	.70	-1.38	.21
18	-1.04	1.40	-.86	.90	-0.51	.19
19	-1.89	.73	-2.43	.40	-1.90	.23
20	-.60	.80	-.70	.50	-0.18	.19
21	-.43	1.42	-.35	1.20	0.04	.20
22	-.08	.91	-.07	.70	0.35	.17
23	-.12	.95	-.11	1.10	0.55	.28
24	-1.31	1.31	-1.11	.80	-0.83	.17
25	-.58	.72	-.72	.50	-0.13	.20
26	-1.34	1.10	-1.26	.70	-0.92	.19

[1] The average item discrimination parameter estimate under the method of marginal maximum likelihood was 1.02.

Next, MC item parameter estimates from the simultaneous calibrations are presented in Table 2.

Table 2

Multiple-choice Item Parameter Estimates of the 1PL, 2PL, and 3PL from Simultaneous Calibrations

	Model					
	1PL&PC	2PL&GR		3PL&GPC		
Item	b_i [1]	a_i	b_i	a_i	b_i	c_i
1	-1.54	.42	-3.47	.28	-2.24	.27
2	-1.77	1.37	-1.57	.87	-1.24	.23
3	-1.17	.92	-1.34	.91	-.19	.42
4	-1.16	1.92	-.88	1.44	-.54	.21
5	-.33	.84	-.40	.88	.36	.28
6	-.05	.96	-.05	1.16	.54	.26
7	-.82	1.07	-.84	.80	-.34	.22
8	-.03	.68	-.04	1.39	.88	.35
9	-.56	1.88	-.43	1.40	-.19	.14
10	.05	.61	.09	.48	.63	.17
11	.03	1.47	.03	1.38	.32	.16
12	-.51	.95	-.56	1.15	.30	.34
13	.60	1.01	.65	1.36	.87	.17
14	-1.25	1.15	-1.23	.85	-.65	.27
15	-.69	.96	-.77	.65	-.39	.17
16	-.47	1.06	-.48	.84	-.02	.20
17	-1.76	1.22	-1.68	.75	-1.36	.22
18	-.95	1.39	-.83	1.02	-.45	.20
19	-1.73	.74	-2.39	.47	-1.79	.24
20	-.54	.79	-.69	.63	.01	.24
21	-.37	1.43	-.32	1.36	.10	.21
22	-.05	.91	-.05	.78	.39	.18
23	-.09	.93	-.09	1.16	.55	.27
24	-1.20	1.20	-1.15	.82	-.76	.20
25	-.51	.75	-.68	.54	-.15	.18
26	-1.22	1.15	-1.20	.76	-.85	.18

[1] The average item discrimination parameter estimate under the method of marginal maximum likelihood was 1.13.

Functioning points of the items had wider range of values for 3PL than both 1PL and 2PL from simultaneous calibration.

Note that the CR items having two categories, which are correct and incorrect responses under the PC, GR, and GPC, were analyzed, in fact, using the 1PL, 2PL, and 2PL respectively.

The 3PL was not applied because the items do not allow guessing. Table 3 nevertheless presents the item parameter estimates of the dichotomous CR items from the separate calibrations. Table 4 presents the item parameter estimates from the simultaneous calibrations.

Table 3

Dichotomous Constructed Response Item Parameter Estimates of the PC, GR, and GPC from Separate Calibrations

	Model				
	PC	GR		GPC	
Items	b_i[1]	a_i	b_i	a_i	b_i
27	-.70	.84	-.88	.84	-.86
28	-.45	1.60	-.36	1.59	-.38
29	.47	1.73	.45	1.73	.36
30	1.04	1.29	1.06	1.30	.90
31	.30	.81	.43	.81	.32
32	.02	1.46	.07	1.46	.01
33	.48	1.65	.46	1.66	.36
34	-.51	1.17	-.49	1.16	-.5
35	.67	1.55	.65	1.55	.53
36	-.77	.71	-1.10	.71	-1.05
37	.20	.79	.30	.80	.21
38	-.36	1.15	-.33	1.15	-.35
39	1.79	1.16	1.91	1.16	1.63
40	.56	1.88	.51	1.88	.40
41	1.27	1.64	1.15	1.66	.95

[1] The average item discrimination parameter estimate under the method of marginal maximum likelihood was 1.24.

Estimates reported above under GR and GPC headings are quite similar because they are from the same model fit, which was 2PL. Small differences are due to the inclusion of the three and five category response items to the estimation procedure and the employment of different models to those items. As can be seen from Table 3, 9 of the items resulted in the same discrimination parameters from GR and GPC. The rest of the items had very similar discrimination parameters.

Table 4

Dichotomous Constructed Response Item Parameter Estimates of the PC, GR, and GPC from Simultaneous Calibrations

	Model				
	PC	GR		GPC	
Items	b_i[1]	a_i	b_i	a_i	b_i
27	-.74	.92	-.84	.92	-.89
28	-.48	1.65	-.39	1.68	-.36
29	.47	1.81	.37	1.91	.44
30	1.07	1.38	.95	1.47	1.05
31	.29	.90	.35	.94	.42
32	.01	1.60	.01	1.65	.06
33	.48	1.79	.38	1.90	.45
34	-.54	1.25	-.50	1.25	-.50
35	.68	1.69	.55	1.79	.64
36	-.80	.77	-1.05	.79	-1.11
37	.19	.89	.23	.93	.29
38	-.38	1.26	-.35	1.27	-.34
39	1.85	1.24	1.75	1.33	1.90
40	.56	2.13	.42	2.28	.50
41	1.31	1.85	1.01	1.99	1.14

[1] The average item discrimination parameter estimate under the method of marginal maximum likelihood was 1.13.

Correlations of the item parameter estimates from the separate and simultaneous calibrations are presented in Tables 5-10. In Tables 5-10, T_{MC+CR} represents the whole test and the item parameter estimates are from simultaneous calibration runs. T_{MC} represents the MC items and the item parameter estimates are from separate calibration runs. T_{CR} represents the CR items and the item parameter estimates are from separate calibration runs.

Table 5

Correlation of the MC Item Parameter Estimates from 1PL and 1PL&PC

		Correlations
T_{MC+CR}, T_{MC}	Difficulty	1 (*n*=26)

Table 6

Correlation of the MC Item Parameter Estimates from 2PL and 2PL&GR

		Correlations
T_{MC+CR}, T_{MC}	Difficulty	.997 (*n*=26)
	Discrimination	.994 (*n*=26)

Table 7

Correlation of the MC Item Parameter Estimates from 3PL and 3PL&GPC

		Correlations
T_{MC+CR}, T_{MC}	Difficulty	.997 (*n*=26)
	Discrimination	.973 (*n*=26)
	Guessing	.954 (*n*=26)

From Tables 8-10, it can be seen that correlations between the item parameter estimates for two category CR items from separate and simultaneous calibrations ranged from .99 to 1.

Table 8

Correlation of the Constructed Response Item Parameter Estimates from PC and 1PL&PC

		Correlations
T_{MC+CR}, T_{CR}	Difficulty	1 (*n* = 15)

Table 9

Correlation of the Constructed Response Item Parameter Estimates from GR and 2PL&GR

		Correlations
T_{MC+CR}, T_{CR}	Difficulty	1 (*n* = 15)
	Discrimination	.994 (*n* = 15)

Table 10

Correlation of the Constructed Response Item Parameter Estimates from GPC and 3PL&GPC

		Correlations
T_{MC+CR}, T_{CR}	Difficulty	1 (*n* = 15)
	Discrimination	.99 (*n* = 15)

Correlation between the difficulty parameters from the separate and simultaneous calibrations was 1 with 1PL and PC models, it was .997 with 2PL and GR, and it was .997 with 3PL and GPC. Correlation between the discrimination parameters from the separate and simultaneous calibrations is .994 with 2PL and GR, and it was .973 with 3PL and GPC.

Table 11

Three Category Constructed Response Parameter Estimates of the PC, GR, and GPC from Separate Calibrations

	PC item parameter estimates					
Item	a_1	a_2	a_3	c_1	c_2	c_3
1	-1.02	0	1.02	-.13	.15	-.02
2	-1.02	0	1.02	-.68	.41	.27
3	-1.02	0	1.02	1.22	-.66	-.55
4	-1.02	0	1.02	1.19	-.09	-1.11
	GR item parameter estimates					
Item	a_1	b_1	b_2			
1	1.15	-.69	.71			
2	1.17	-1.36	.47			
3	2.49	.59	.98			
4	1.82	.66	1.55			
	GPC item parameter estimates					
Item	a_1	a_2	a_3	c_1	c_2	c_3
1	-.82	0	.82	-.06	.08	-.02
2	-.91	0	.91	-.61	.37	.24
3	-1.65	0	1.65	1.54	-.46	-1.08
4	-1.36	0	1.36	1.42	.04	-1.45

Table 12

Three Category Constructed Response Parameter Estimates of the PC, GR, and GPC from Simultaneous Calibrations

	PC item parameter estimates					
Item	a_1	a_2	a_3	c_1	c_2	c_3
1	-1.07	0	1.07	-.14	.15	-.01
2	-1.07	0	1.07	-.68	.40	.28
3	-1.07	0	1.07	1.20	-.66	-.54
4	-1.07	0	1.07	1.18	-.08	-1.1
	GR item parameter estimates					
Item	a_1	b_1	b_2			
1	1.20	-.70	.63			
2	1.27	-1.29	.39			
3	2.59	.50	.87			
4	1.90	.57	1.43			
	GPC item parameter estimates					
Item	a_1	a_2	a_3	c_1	c_2	c_3
1	-.87	0	.87	-.10	.07	.02
2	-1.01	0	1.01	-.66	.38	.28
3	-1.85	0	1.85	1.48	-.47	-1.02
4	-1.53	0	1.53	1.36	.03	-1.39

Table 13

Five Category Constructed Response Parameter Estimates of the PC, GR, and GPC from Separate Calibrations

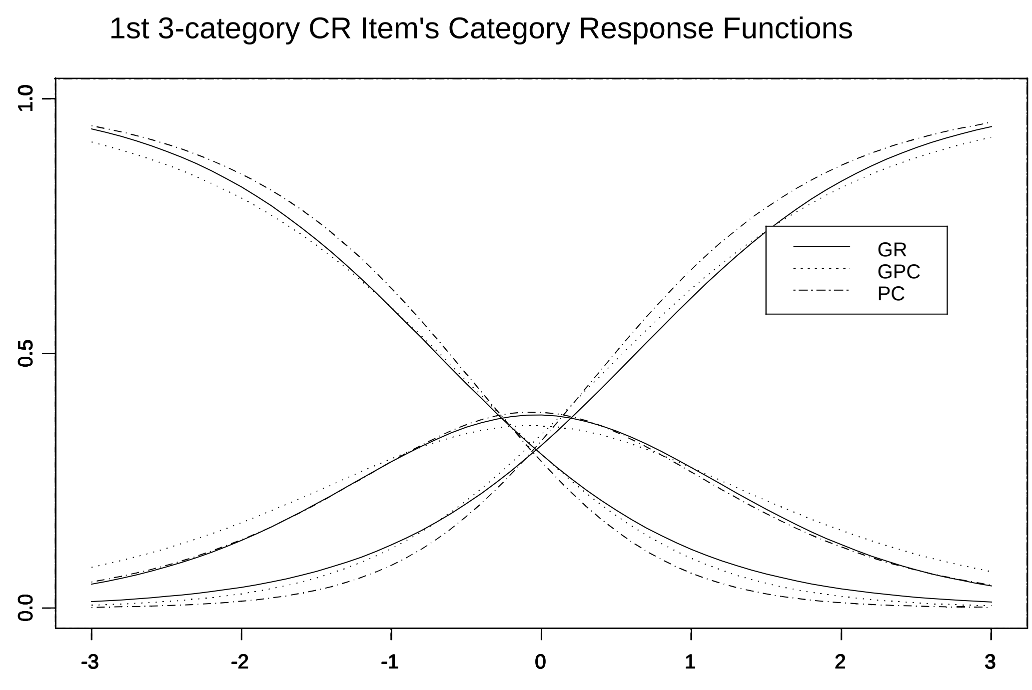
	PC parameter estimates									
Item	a_1	a_2	a_3	a_4	a_5	c_1	c_2	c_3	c_4	c_5
5	-2.03	-1.02	0	1.02	2.03	1.01	.60	.81	-1.09	-1.33
6	-2.03	-1.02	0	1.02	2.03	-.17	.10	.70	.18	-.81
	GR parameter estimates									
Item	a_1	b_1	b_2	b_3	b_4					
5	1.82	-.35	.27	1.21	1.55					
6	1.77	-.96	-.33	.53	1.41					
	GPC parameter estimates									
Item	a_1	a_2	a_3	a_4	a_5	c_1	c_2	c_3	c_4	c_5
5	-1.80	-.90	0	.90	1.80	1.05	.54	.74	-1.11	-1.22
6	-1.81	-.91	0	.91	1.81	-.06	.09	.63	.11	-.78

Table 14

Five Category Constructed Response Parameter Estimates of the PC, GR, and GPC from Simultaneous Calibrations

	PC parameter estimates									
Item	a_1	a_2	a_3	a_4	a_5	c_1	c_2	c_3	c_4	c_5
5	-2.15	-1.07	0	1.07	2.15	.98	.58	.83	-1.06	-1.32
6	-2.15	-1.07	0	1.07	2.15	-.20	.07	.70	.21	-.78
	GR parameter estimates									
Item	a_1	b_1	b_2	b_3	b_4					
5	1.93	-.38	.19	1.09	1.41					
6	1.95	-.92	-.35	.44	1.27					
	GPC parameter estimates									
Item	a_1	a_2	a_3	a_4	a_5	c_1	c_2	c_3	c_4	c_5
5	-2.05	-1.02	0	1.02	2.05	.97	.52	.74	-1.08	-1.16
6	-2.11	-1.05	0	1.05	2.11	-.16	.09	.65	.15	-.73

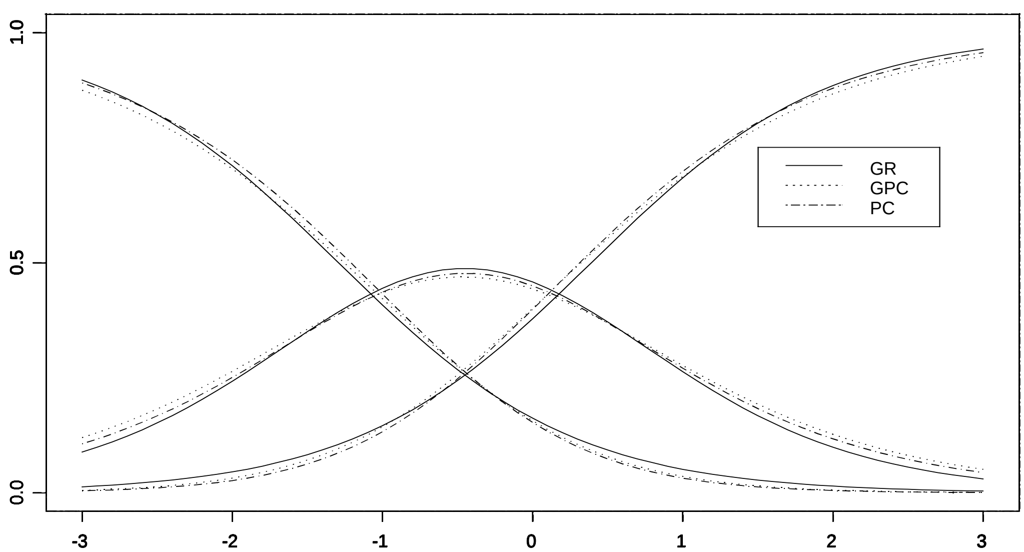
Figure 4. Category response functions of three category items.



(a)

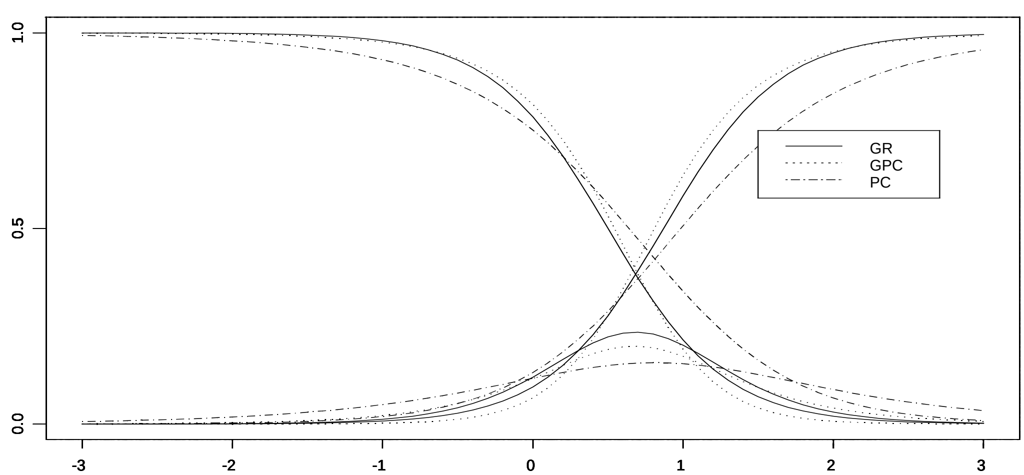
For the CR items, item response functions can be graphed and visually inspected (see Figure 4). The graphs, which are presented in Figure 4 and Figure 5, are from simultaneous calibration. All response functions seemed to be pretty similar. The points that items functioned on the ability scale were almost the same for categories across the different models. The slopes of the category response functions are close to each other.

2nd 3-category CR Item's Category Response Functions

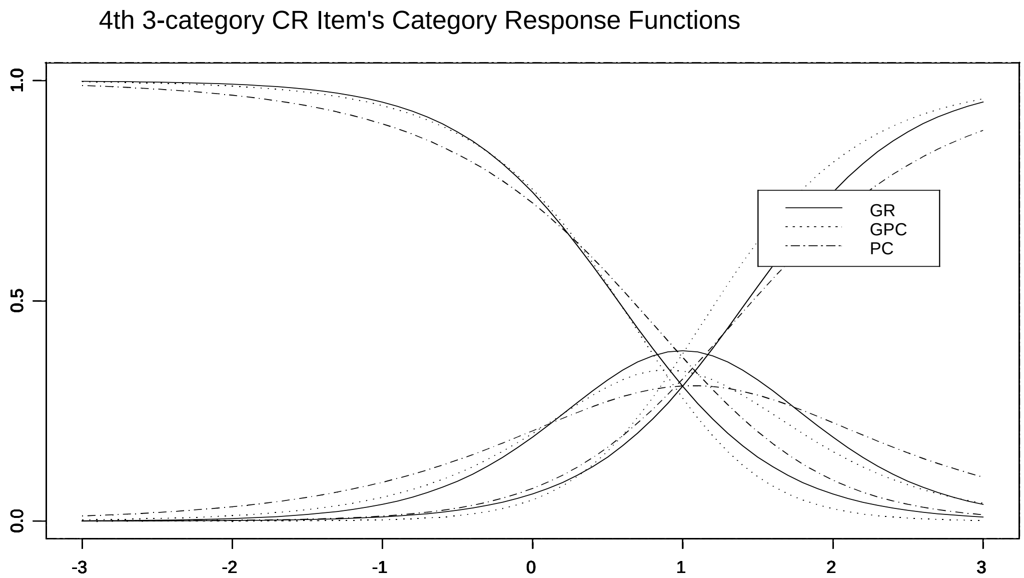


(b)

3rd 3-category CR Item's Category Response Functions

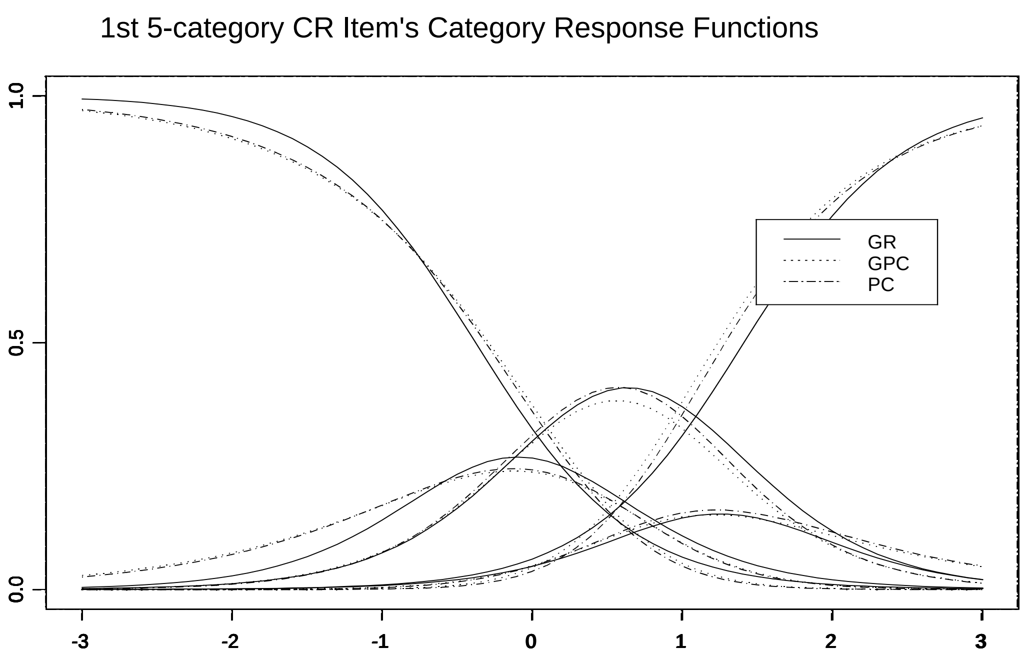


(c)

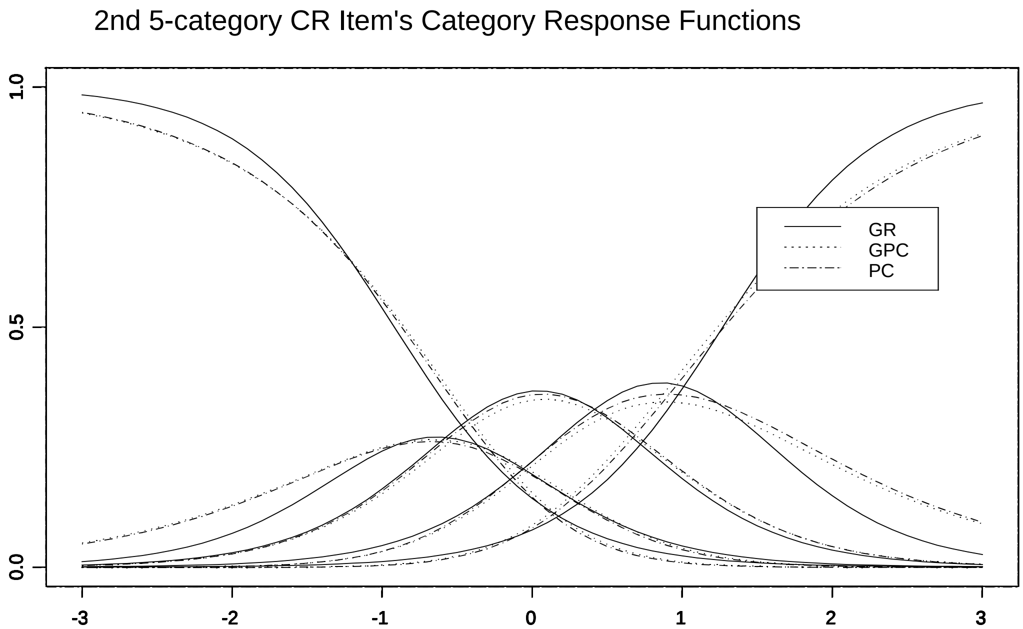


(d)

Figure 5. Category response functions of five category items.



(a)



(b)

Table 15

Correlation of Item Parameters from 1PL&PC, 2PL&GR, 3PL&GPC Combinations

		1PL&PC	2PL&GR
Item difficulty	1PL&PC		
	2PL&GR	.888 (*n*=26)	
	3PL&GPC	.924 (*n*=26)	.952 (*n*=26)
Discrimination	3PL&GPC		.639 (*n*=26)

<u>Ability Estimates</u>

A comparison was performed between the scores (i.e. ability estimates) to examine if separate or simultaneous calibrations led to different scores and if various combinations of models. Table 16 shows the correlations of the expected a posteriori ability estimates from the 1PL, 2PL, 3PL, PC, GR, GPC, 1PL&PC, 2PL&GR, 3PL&GPC estimation procedures.

Correlations of the ability estimates from the tests using same sets of items are higher than the tests with different types of items. For instance, the correlations among the MC part of the test are higher and the correlations among the CR part of the test are higher than the correlations that involve the whole test.

Table 16

Correlation of Expected A Posteriori Scores Resulted from Various Models

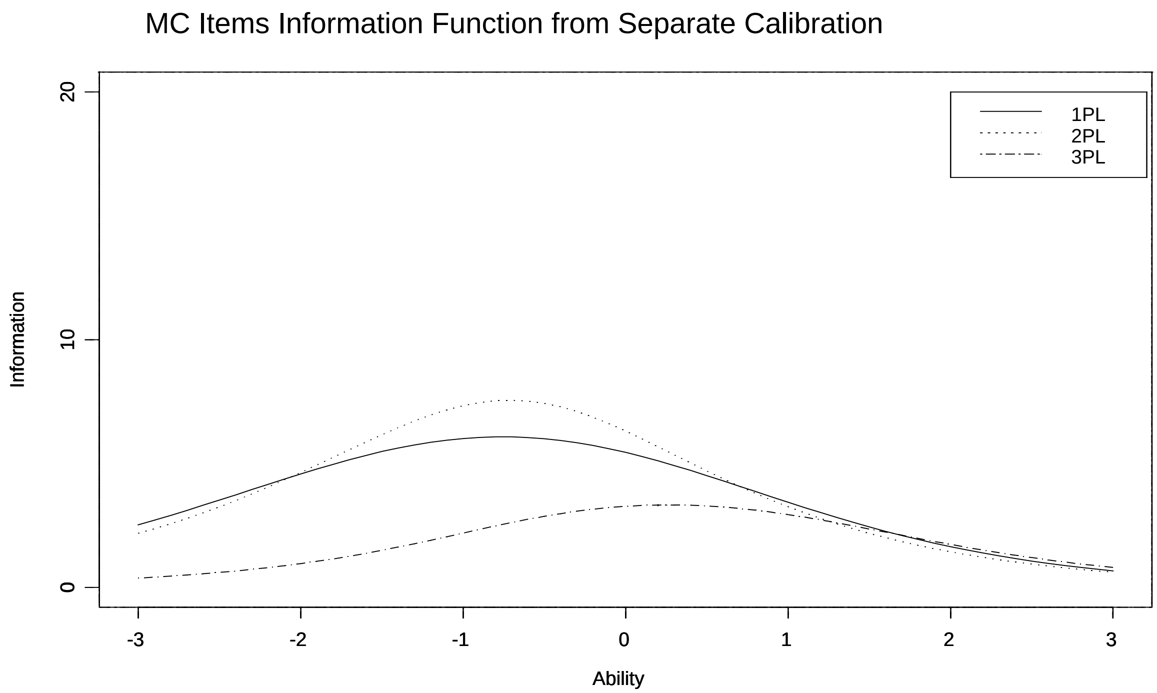
	1PL	2PL	3PL	PC	GR	GPC	1PL&PC	2PL&GR	3PL&GPC
1PL	—	0.992	0.985	0.847	0.848	0.850	0.945	0.935	0.931
2PL		—	0.994	0.850	0.850	0.852	0.944	0.939	0.935
3PL			—	0.850	0.850	0.852	0.940	0.937	0.938
PC				—	0.993	0.995	0.967	0.968	0.968
GR					—	0.998	0.963	0.970	0.969
GPC						—	0.964	0.970	0.971
1PL&PC							—	0.996	0.990
2PL&GR								—	0.994
3PL&GPC									—

Test information functions are investigated from the different calibrations.

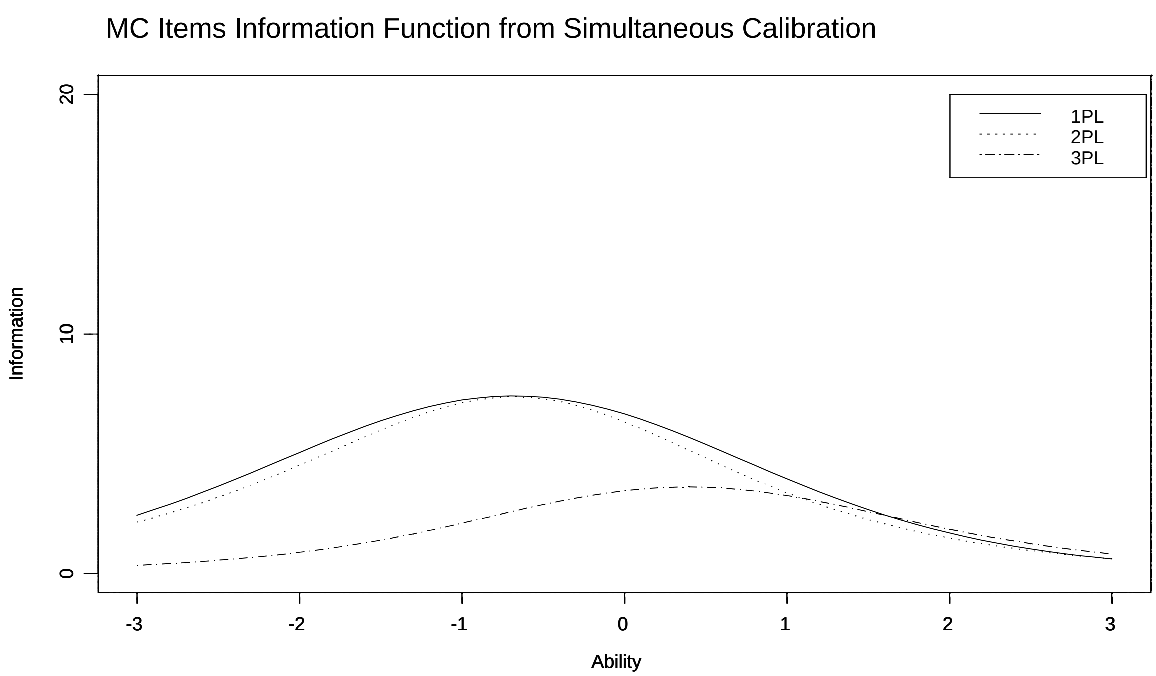
Test Information

The test information functions obtained from the MC items are presented in Figure 6 for the separate and simultaneous calibrations. The 3PL consistently yielded lower information function for the lower ability levels. The gap between the information functions from 2PL and 1PL was smaller from simultaneous calibration than the separate calibration.

Figure 6. Information functions for MC items.



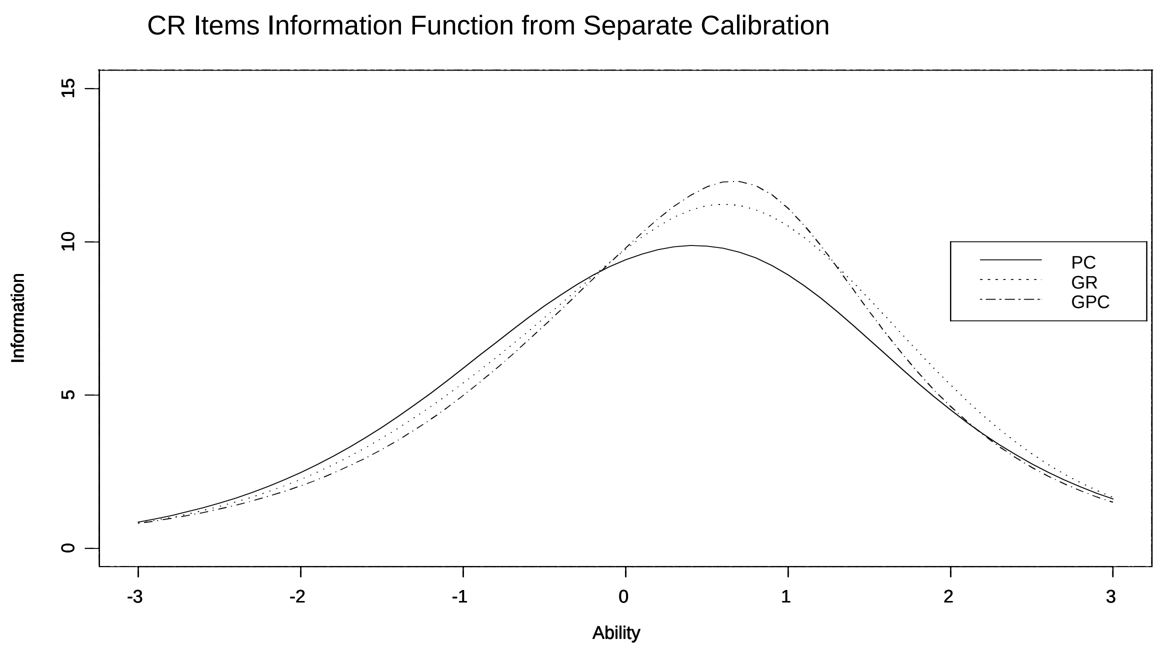
(a)



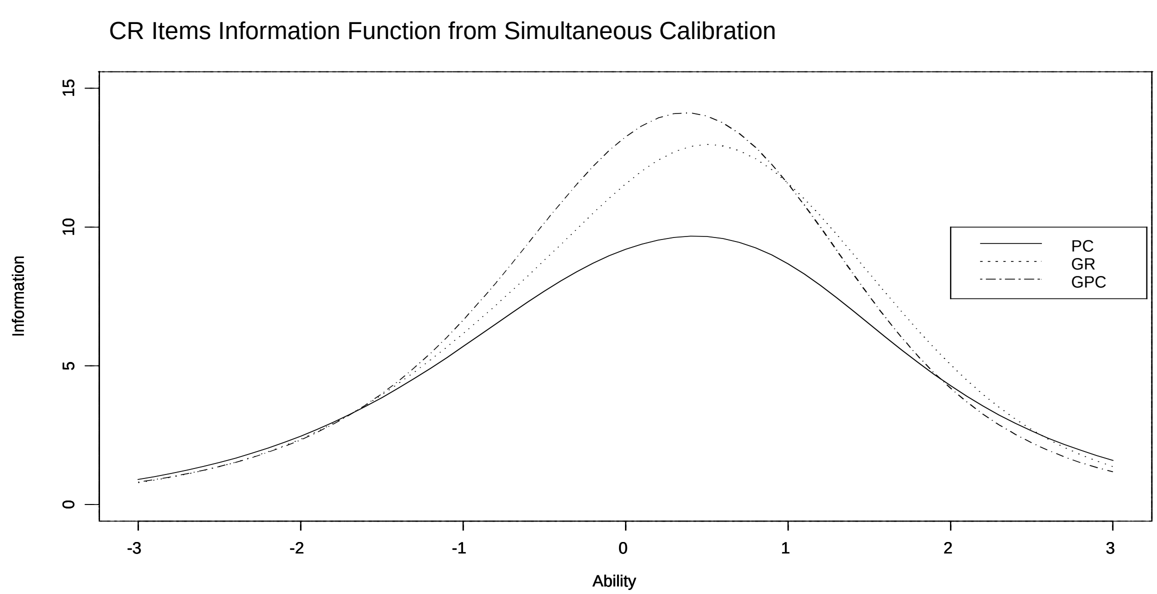
(b)

Figure 7 presents the information functions from the CR items for separate and simultaneous calibrations. For the middle ability level, the GPC yielded relatively larger information while the PC yielded relatively flatter information.

Figure 7. Information functions for CR items.



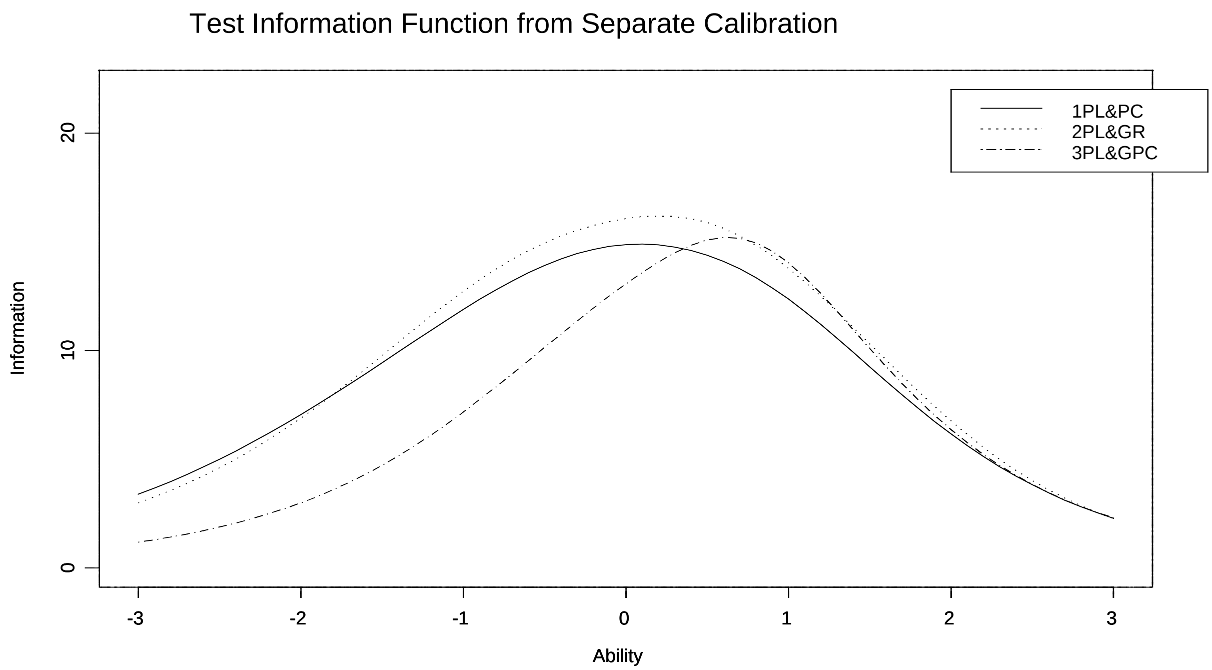
(a)



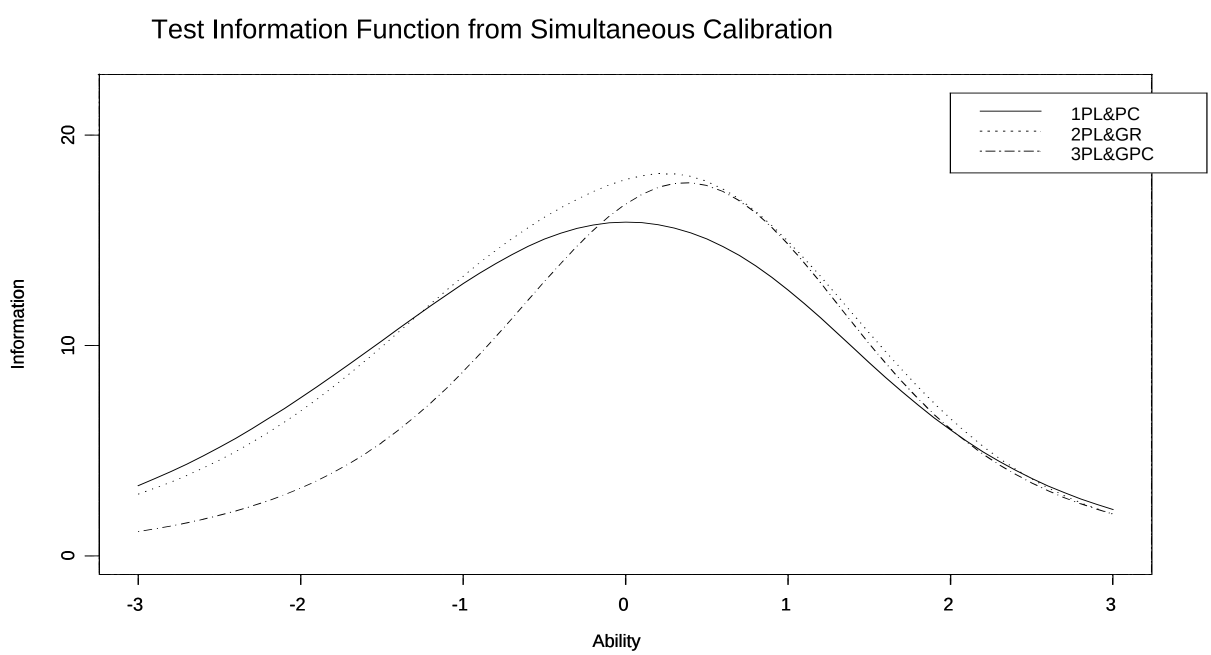
(b)

Figure 8 presents the test information functions from the test as a whole.

Figure 8. Information functions for the test.



(a)

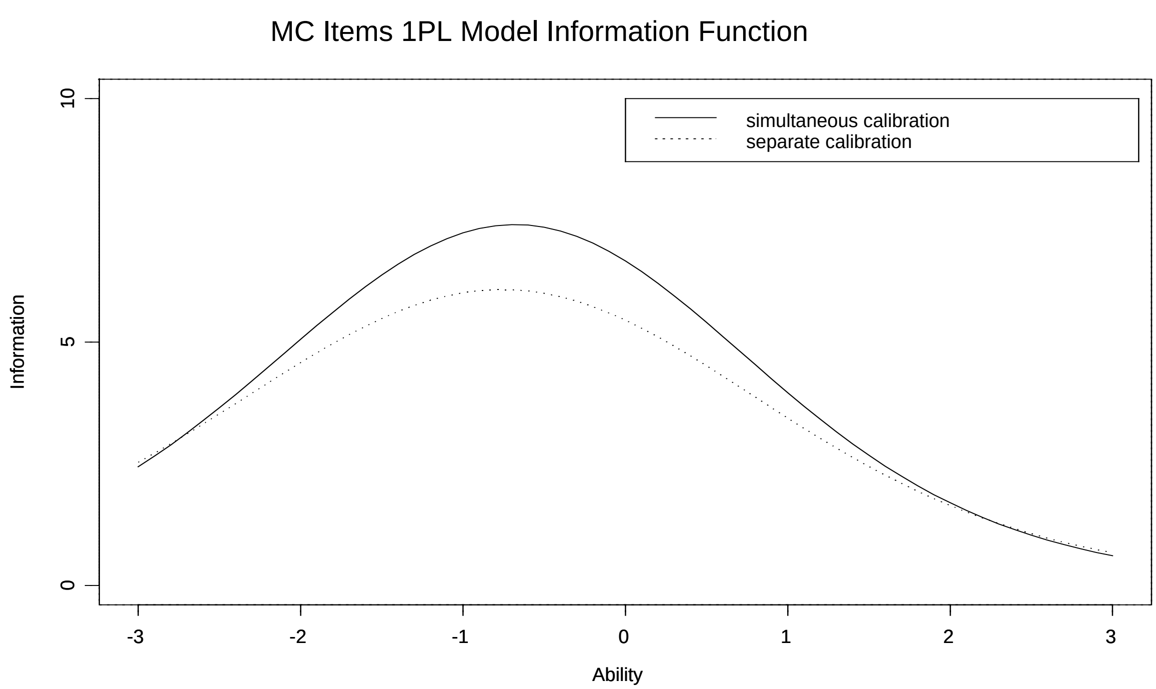


(b)

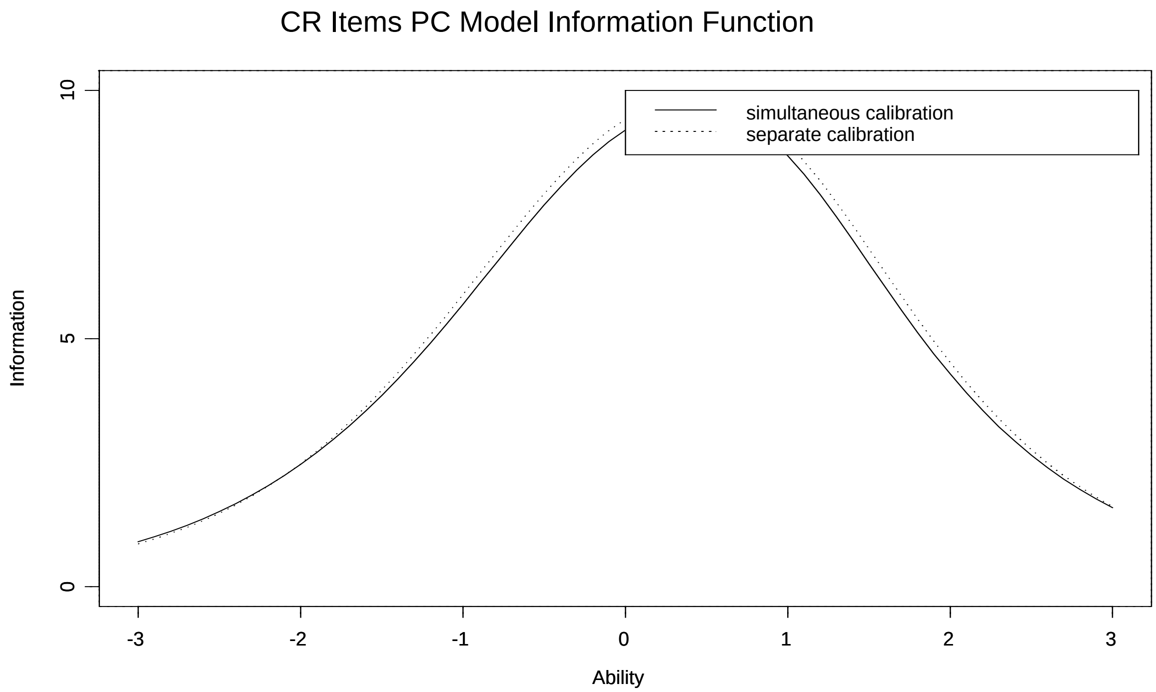
As can be seen from Figure 8, for the most of the ability scale, the 2PL&GR yielded higher information. The 1PL&PC yielded higher information than the 3PL&GPC for the lower ability levels. For the ability levels lower than –1 and higher than 2.5, the 1PL&PC combination produced more information. However, for the ability levels in which students are more likely to be present, the derivatives were smaller for the 2PL & GR.

Figures 9-11 also present the comparison of information functions between separate and simultaneous calibrations for the different models.

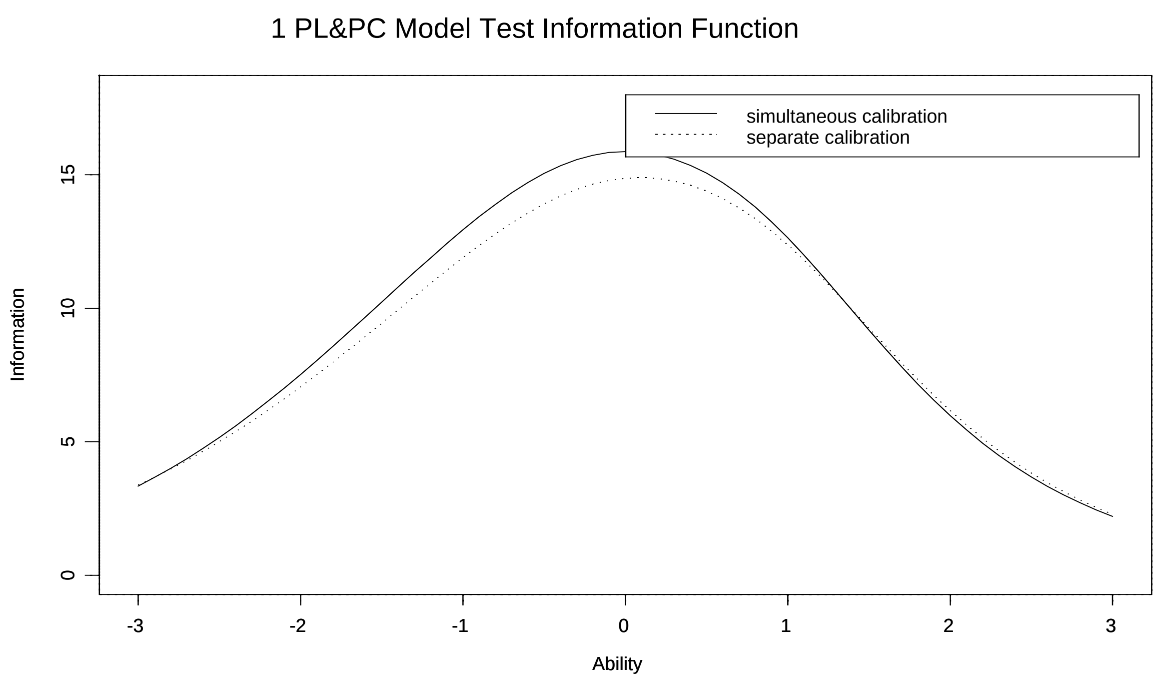
Figure 9. Information functions from 1PL&PC model.



(a)

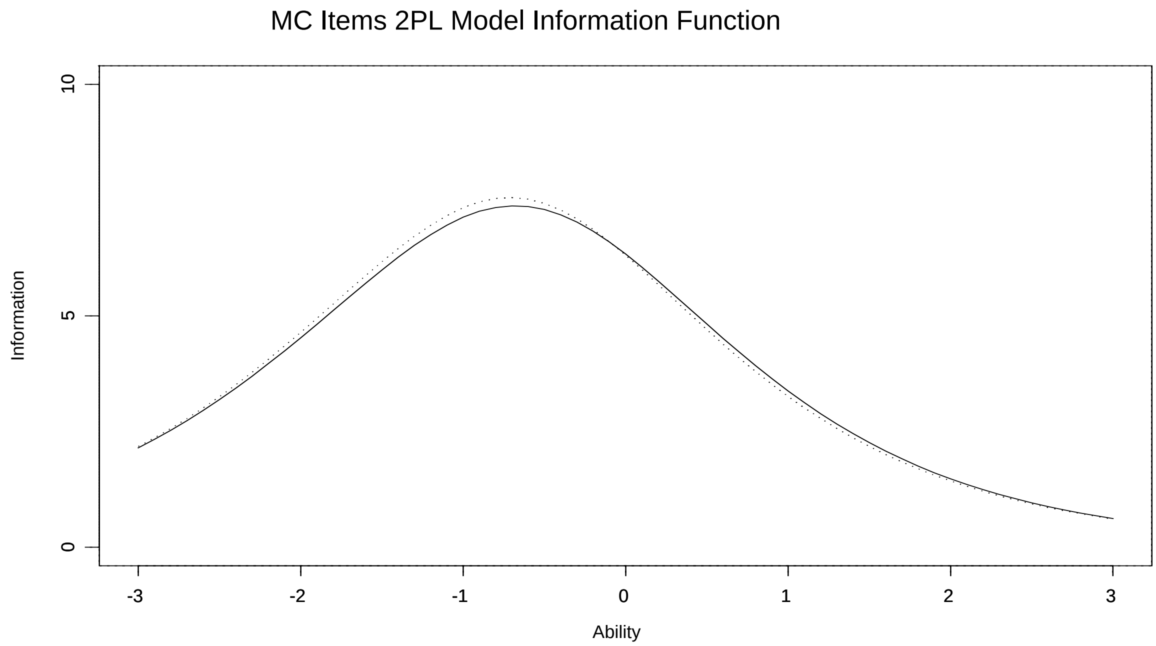

CR Items PC Model Information Function
simultaneous calibration
separate calibration
Information
Ability
0
5
10
-3
-2
-1
0
1
2
3

(b)

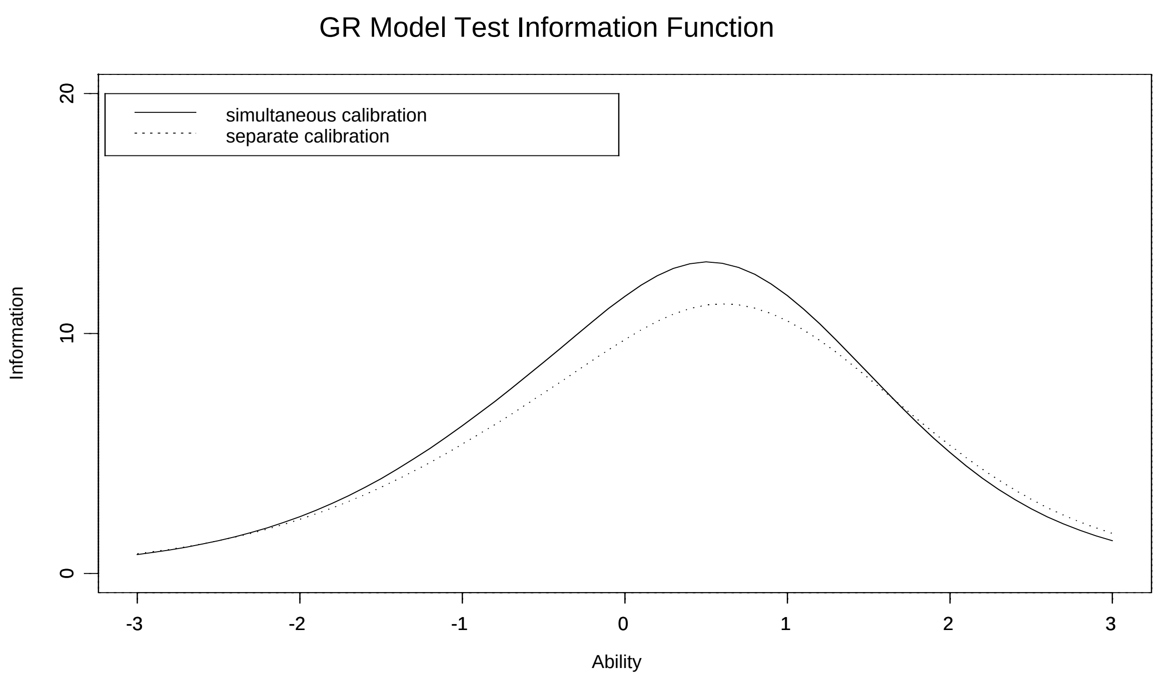

1 PL&PC Model Test Information Function
simultaneous calibration
separate calibration
Information
Ability
0
5
10
15
-3
-2
-1
0
1
2
3

(c)

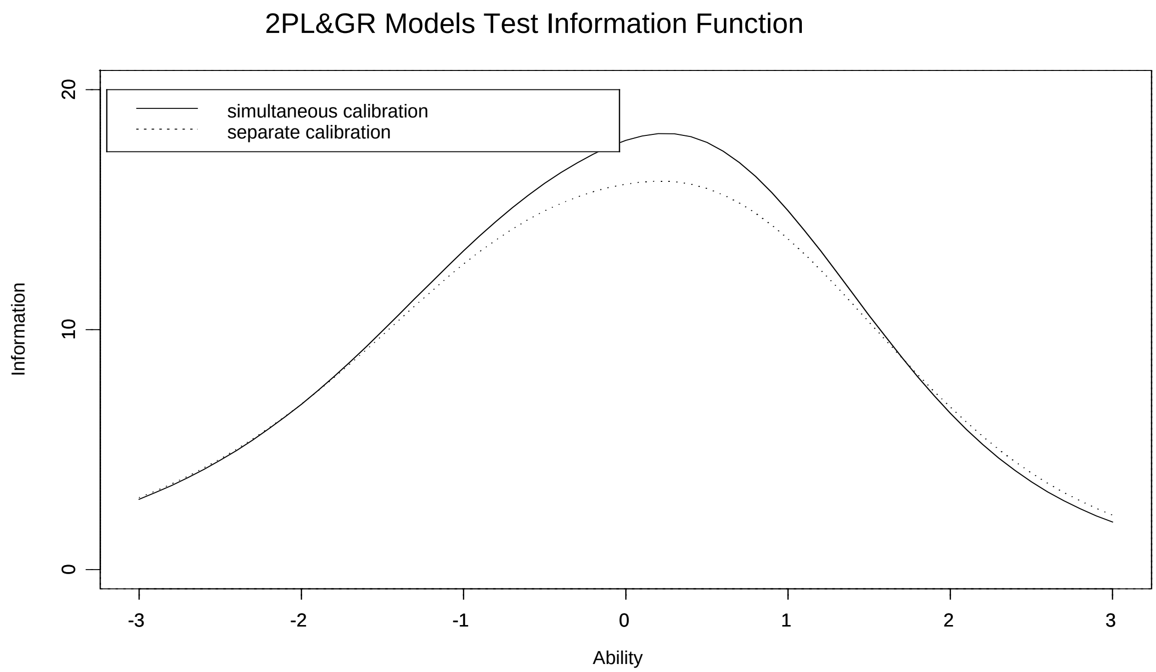
Figure 10. Information functions from 2PL&GR model.



(a)

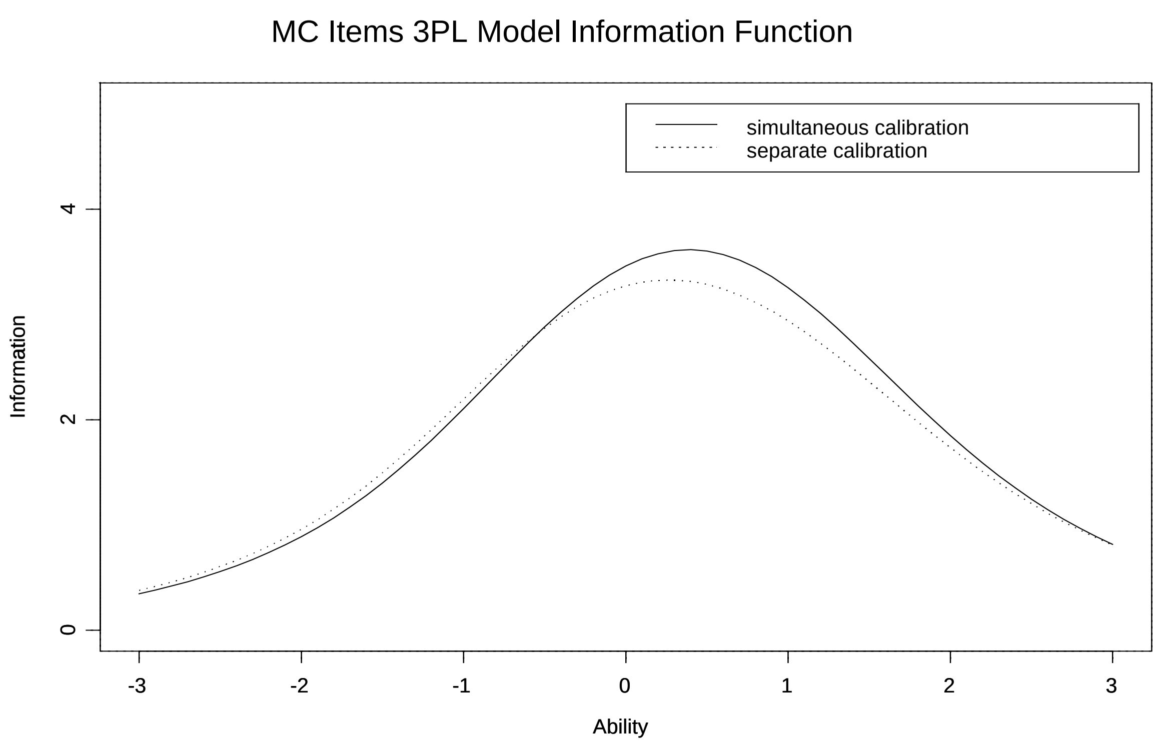


(b)

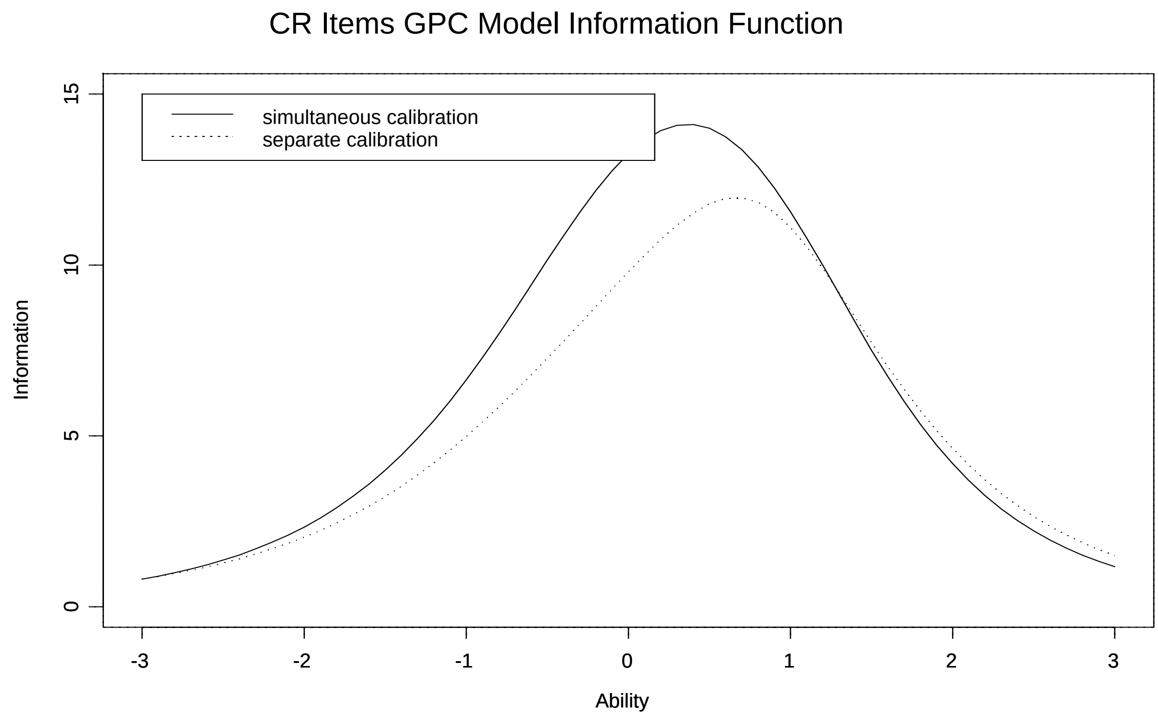


(c)

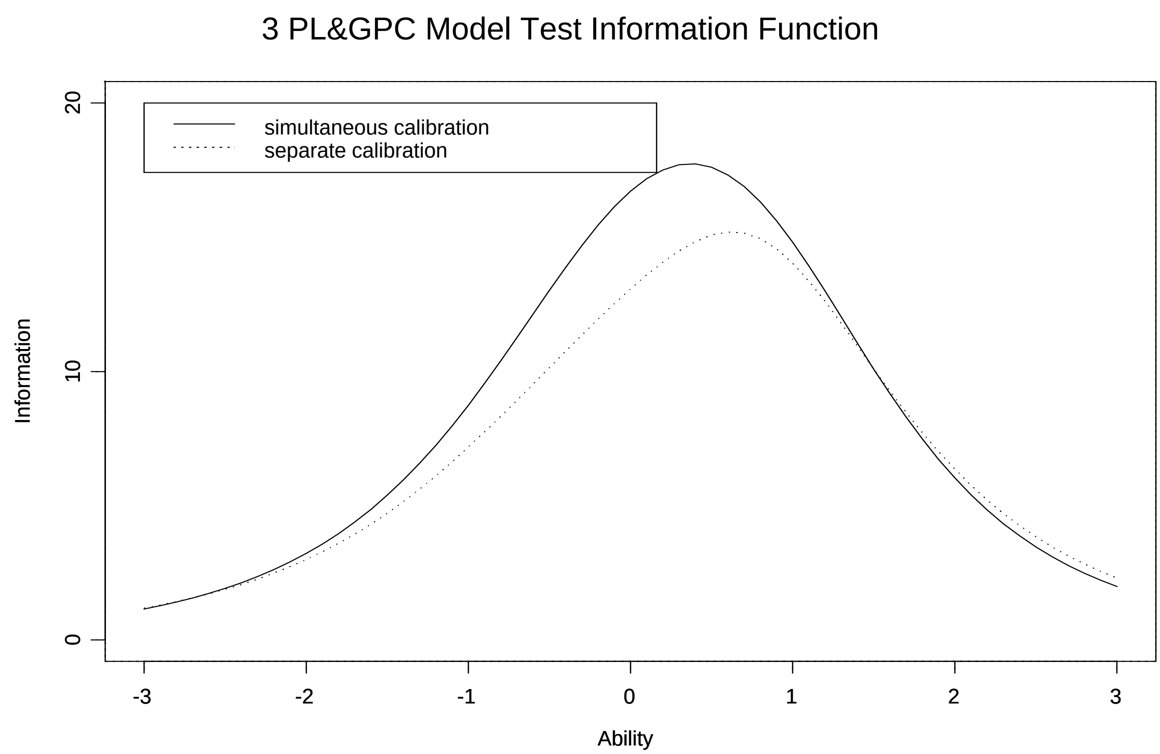
Figure 11. Information functions from 3PL&GPC model.



(a)


CR Items GPC Model Information Function
simultaneous calibration
separate calibration
Information
Ability
0
5
10
15
-3
-2
-1
0
1
2
3

(b)


3 PL&GPC Model Test Information Function
simultaneous calibration
separate calibration
Information
Ability
0
10
20
-3
-2
-1
0
1
2
3

(c)

CHAPTER 5

SUMMARY AND DISCUSSION

Summary

The purpose of the study was to apply item response theory to score multiple-choice and constructed-response items together with various item response functions and investigate the resulting item, ability estimates, and information functions. This study also investigated the best fitting model from among the three dichotomous and three polytomous item response models. Another purpose was to investigate the information loss from simultaneous calibration, which was reported in the previous study (Ercikan et al., 1998).

The factor analytic examination of the tests indicated that multiple choice and constructed response items assessed constructs that were sufficiently similar to construct a common scale and to provide a single set of scores for responses to both item types.

Increasing test length by combining two item types increased overall measurement accuracy and the reliability of the test. The reliabilities from only the MC items were .82, .83, and .84 from 1PL, 2PL, and 3PL, respectively. The reliability from only the CR items was .88. The reliability from the whole test was .93.

The values of the model fit index for only the MC items were 15579, 15298.7, and 15211.7 for the 1PL, 2PL, and 3PL models, respectively. For the CR items, the values of the model fit index were 14380.9, 14112.5, and 14106.8 from the PC, GR, and GPC models. The values of the model fit index were 42549.7, 41877.1, and 41708.7 from the 1PL&PC, 2PL&GR, and 3PL&GPC models. A wider range of difficulty parameter values was obtained from the 3PL model than the 2PL or the 1PL model. Values of the item difficulty estimates for the 1PL model ranged from –1.93 to .62. The values were from –3.08 and .65 for the 2PL model. For the 3PL model, values were –2.13 to .90.

The earlier empirical study (Ercikan et al., 1998) reported that the separate calibration runs yielded higher discrimination estimates than did the simultaneous calibration runs. As can be seen from the Tables 1 and 2, almost half of the items had lower discrimination estimates from the separate calibration runs than the simultaneous calibration runs and the other half had higher ones. However, the patterns were consistent across the different models. If an item had a lower discrimination estimate from the separate calibration run than the simultaneous calibration run for the 2PL model, it had a lower discrimination estimate from the separate calibration run than the simultaneous calibration run for the 3PL model, too.

Correlations of the item parameter estimates from the separate and simultaneous calibration runs were high, ranging from .954 to 1. High correlations were observed ranging from .997 to 1 for the difficulty estimates. Correlations of the difficulty estimates were 1 for the constructed response items with two categories from the separate and simultaneous calibration runs. Correlations among the item difficulty estimates for the multiple-choice items from the simultaneous calibration runs across the three model combinations ranged from .888 to .952.

The comparisons of the ability estimates were made for the nine calibration procedures: three calibration runs with the MC items, three calibration runs with the CR items and three calibration runs with the combination of the MC and CR items. Correlations ranged from .992 to .994 for MC items, from .993 to .998 for the CR items and from .990 to .996 from the combination of the MC and CR items. Relatively smaller correlations between the ability estimates from the MC and CR items were observed. These correlations ranged from .847 to

.852. Note that these last correlations are using different items, which are MC and CR whereas the previous ones are using the same items, but different models.

When calibrated separately, the 2PL model yielded the highest information for the MC items and the 3PL model yielded the lowest information. Simultaneous calibration yielded quite similar information functions for the 1PL and the 2PL models for the MC items. The 3PL information was somewhat different from the 1PL and the 2PL models, having a flatter function. The CR items yielded highest information from the GPC model, and the PC model yielded the flatter function compared to the GR, or the GPC model. The test as a whole yielded highest information functions for the 2PL&GR combination from both separate and simultaneous calibration runs. Information functions resulted from the 3PL&GPC and the 1PL&PC combinations had crossing patterns. The 1PL&PC information function had larger values for lower ability levels than the 3PL&GPC, whereas the 3PL&GPC information function had larger values for higher ability levels. The information functions from the MC items for 1PL and 3PL models yielded higher information for simultaneous calibration than separate calibration. Information functions of the MC items with the 2PL model from both separate and simultaneous calibration runs were quite similar. Separate calibration yielded higher information in most of the scale for the MC items with the 2PL model. Keep in mind that the 2PL model yielded the highest information among the three dichotomous models from separate calibration. Information function of the CR items with the GR and GPC models yielded larger values from simultaneous calibration than separate calibration. The PC model yielded quite similar information function from separate and simultaneous calibration runs; but the magnitude of the differences was not the same for the GR and GPC models. Separate calibration yielded higher information than simultaneous calibration. The magnitude of the test information differences from the separate and simultaneous calibration runs was consistent across the three model combinations. Test information obtained from the 1PL&PC, 2PL&GR, and 3PL&GPC showed higher values for simultaneous calibration than for separate calibration.

Discussion

Studies, which dealt with combining multiple-choice and open-ended items, reported implications for test construction. The use of empirical data makes the generalizability of these results and implications somewhat questionable. It should be kept in mind that this study used empirical data also. However, the results from previous studies were not fully consistent with the results from the current study.

The first issue related to the use of open-ended items in tests was validity improvement. There is a belief that open-ended items improve the validity. To be able to say that adding open-ended items improve the test validity, a researcher is supposed to gather the appropriate evidence. It is maybe a myth that open-ended items measure something different than multiple-choice items. If the general mathematics ability is a construct, which has a component that cannot be measured without open-ended items, adding open-ended items will increase the validity of the test. In my50

study, the rationale behind the idea that two item types measure different levels of cognitive ability could be in conflict by just inspecting the items that the test includes. From the investigation of multiple-choice items, it can be seen that they measured various cognitive levels. The test includes items that try to cover the knowledge, comprehension, application, and analysis and include levels defined by Bloom (1956). The same argument is valid for open-ended items in the test. Furthermore, with the factor analytic investigation, we concluded that the test, which was the combination of open-ended and multiple-choice items, measured one construct. This was also a requirement to apply the item response theory estimation methods to the data. Test scoring and also reporting has been an issue under question with two item types. From the starting point, if we would like to use item response theory models to score tests, test constructors should be careful about not creating items in a way that two types of items measure two different dimensions of cognitive ability. As long as they measure the same or similar cognitive ability dimensions, scoring them under item response theory avoids the weighting of these two parts of the test. The use of open-ended items to measure different cognitive ability dimensions is not viable in this case. The second issue is the reliability of the tests with open-ended items. It is at least two-fold: test reliability and scoring reliability (inter-rater reliability). Test reliability in our study increased when the open-ended items were combined with the multiple-choice items. Scoring reliability was not investigated in this study. The question is whether the increase in the reliability is attributable to the adding open-ended items per se. The increase in reliability might just be due to the increase in the number of items, regardless of the type of the item. In addition, adding more multiple-choice items may increase the reliability further. Looking at this reliability increase may not support the use of constructed response items.

Model selection has not been an issue for the combination of two item types by earlier studies. This study compared three combinations for the test. Prior to looking at the combination, we looked at the two types of item groups separately. For only multiple-choice items, the 3PL model resulted in the lowest model fit value, indicating the best fit. However, it should be noted that the amount of difference in the fit indices from the 2PL and 3PL models was pretty small. It can be concluded that both 2PL and 3PL models fit to the data well. From the investigation of fit indices for constructed response items, it can be concluded that the GR and GPC models fit better than the PC model. For the whole test, the 2PL&GR and 3PL&GPC models fit better than the 1PL&PC model.

It should also be noted that item parameter estimates were comparable from various calibration runs because the computer program MULTILOG employed the marginal maximum likelihood estimation, where it is assumed that the examinees are from a population in which ability is distributed according to the same density function. Investigation of these item parameter estimates indicated that they were not affected in a particular way from the separate and simultaneous calibration runs (e.g., having higher discrimination from one than the other or vice versa). Items had higher and lower discrimination parameter estimates from the separate and simultaneous calibration runs in a consistent manner. Some items, however, had higher discrimination parameter estimates from simultaneous calibration, and others had higher discrimination parameter estimates from separate calibration. Item parameter estimates were highly correlated for the separate and simultaneous calibration runs. This indicates that estimation of multiple-choice and constructed-response items together and separately produced consistent results. The correlation of item parameter estimates from the 3PL&GPC and 2PL&GR combinations was .952, which was higher than the correlation between the 1PL&PC model and the other two combinations. Examinee ability estimates were highly correlated from the multiple-choice items, constructed-response items and the multiple-choice and constructed response item combinations. As expected, the ability estimates from the constructed response and multiple-choice only items resulted the lowest correlation because they used different items, (i.e., they did not have items in common). Correlations between the ability estimates from three models using only multiple-choice items were pretty high, as were the ones using only constructed response items. Correlations of the ability estimates from three combinations were higher because they used all items.

Another investigation was carried out for information functions. Even though total information function for multiple-choice items yielded the largest value from the 2PL model, for several items the amount of information function was higher for the 1PL or 3PL model. The greater information leads to the greater contribution of the items to measure a given ability level. Instead of looking at the amount of information across different models, if we looked at the point in the ability scale where the information is at its highest value, it is apparent that information functions of multiple-choice items from the 1PL and 2PL models have reached their maximum values at very close ability points. On the other hand, the 3PL model had somewhat different shape. First of all, it was flatter across the ability scale. A flatter information function explains the more even contribution of judging ability, whereas sharper plots explain the larger contribution to assess the ability levels. Thus, for the 3PL model, higher ability levels were more precisely measured than the lower ability levels. The information function from the 3PL model reached its highest value at around .5, whereas the 1PL and 2PL information functions had their highest values at around –1. This means that the 1PL and 2PL models assess more accurately for the lower ability levels. For the constructed-response items, information functions were quite similar to each other in terms of the point that they had the maximum information value, which was around .5. The amount of information was also pretty close to each other from the GPC, PC and GR models, where the GPC model had the highest value. The information function includes the inverse of the variance of the conditional distribution of the $\hat{\theta}$ at a given ability level, (i.e. $\sigma^2{}_{\hat{\theta}|\theta}$). Thus the larger this variance, the less precise the estimate of θ and the less information one has about the examinee's unknown ability level. The test information function, which takes all the items into account, was highest from the 2PL&GR combination. Investigation of the information from the simultaneous and separate calibration runs indicated that calibrating multiple-choice and constructed items together did not lead to information loss. For some items, the item information function was higher for separate calibration, and for some other items the information function was higher for simultaneous calibration. In summary, calibrating items together did not lead to any technical problems, but seemed to enhance the precision of estimation of parameters, as well as the reliability of the test.

REFERENCES

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

Baker, F.B. (1992). *Item response theory: Parameter estimation techniques* . New York: Marcel Dekker.

Baker, F.B. (2001). *The basics of item response theory.* ERIC Clearinghouse on Assessment and Evaluation.

Bloom, Benjamin S. (Ed.)(1956) *Taxonomy of educational objectives, handbook 1: The cognitive domain*. New York: McKay.

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

Childs, R.A., & Chen W.H. (1999). Obtaining Comparable Item Parameter Estimates in MULTILOG and PARSCALE for Polytomous IRT models. *Applied Psychological Measurement*, *23*, 371-379

Crocker, L., & Algina, J. (1986). *Introduction to classical & modern test theory*. Belmont, CA: Wadsworth.

Davis, A. (1992). *Using tests to evaluate the impact of curricular reform on higher order thinking.* (ERIC Document Reproduction Service No. ED. 373 114).

Dawadi, B. R. (1999). *Robustness of the Polytomous IRT Model to Violations of the Unidimensionality Assumption*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada. (ERIC Document Reproduction Service No. ED429998).

De Ayala, R. J. (1989). Computerized adaptive testing: A comparison of the nominal response model and the three-parameter model. *Educational and Psychological Measurement, 49,* 789-805.

De Ayala, R. J., Dodd, B. G., & Koch, W. R. (1992). A comparison of partial credit and graded response models in computerized adaptive testing. *Applied Measurement in Education, 5(1),* 17-34.

DeMars, C.E. (2002). *Recovery of graded response and partial credit parameters in MULTILOG and PARSCALE.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED476138).

Dodd, B. G., Koch, W. R. & De Ayala, R.J. (1989). Operational characteristics of adaptive testing procedures using the graded response model. *Applied Psychological Measurement, 13,* 129-144.

Donoghue, J.R. (1993). An empirical examination of the IRT information in polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, 31, 295-311.

Ercikan, K.; Schwarz, R.D.; Julian, M.W.; Burket, G. R.;Weber, M. M.; & Link, V. (1998). Calibration and Scoring of Tests With Multiple-Choice and Constructed-Response Item Types. *Journal of Educational Measurement, 35,* 137-154.

Folk, V. G., & Green, B.F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13(4)*, 373-89.

Grima, A. M., & Weichun, W. M. (2002). *Test Scoring: Multiple-Choice and Constructed-Response Items.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Hambleton, R.K., & Swaminathan, H. (1985). *A look at psychometrics in the Netherlands.* (Tech. Rep. No TM860514) Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED273665).

Haley, D.C. (1952). *Estimation of the dosage mortality relationship when the dosage is subject to error*, (Tech. Rep. No. 15) Stanford, CA: Stanford University, Applied Mathematics and Statistics Laboratory.

Jöreskog, K. & Sörbom, D. (2003). *Lisrel 8.54.* Chicago, IL: Scientific Software International.

Lau, C. A.; Wang, T. (1998). *Comparing and Combining Dichotomous and Polytomous Items with SPRT Procedure in Computerized Classification Testing.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA (ERIC Document Reproduction Service No. ED430033).

Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

Rudner, L. M. (2001). Informed Test Component Weighting. *Educational Measurement: Issues and Practice, 20(1)*, 16-19.

Samejima, F.(1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrica Monograph Supplements, 17.*

Samejima, F. (1996). *Polychotomous responses and the test score*. Paper presented at the annual meeting of the National Council of Measurement in Education, New York, NY. (ERIC Document Reproduction Service No. ED397105).

Sykes, R.C., & Hou, L. (2003) Weighting Constructed-Response Items in IRT-Based Exams. *Applied Measurement in Education, 16(4),* 257-275.

Thissen, D.J., & Steinberg, L. (1984). A Response model for multiple-choice items. *Psychometrica, 49*, 501-519.

Thissen, D.J. (1991) MULTILOG: Multiple, categorical item analysis and test scoring using item response theory (Version 6.0) [Computer program]. Chicago: Scientific Software.

Tuerlinckx, F., & De Boeck, P. (2001) Non-modeled item interactions lead to distorted discrimination parameters: A case study. *Methods of Psychological Research-Online,6,* 159-174.

Lukhele, R., Thissen, D. & Wainer, H. (1994). On the relative value of multiple choice, constructed response and examinee selected items on two achievement tests. *Journal of Educational Achievement, 31,* 234-250.

Wainer, H. &Thissen, D. (1993). Combining Multiple-Choice and Constructed-Response Test Scores: Toward a Marxist Theory of Test Construction. *Applied Measurement in Education,6(2),*103-118.

APPENDICES

A. MULTILOG Command for 1_Parameter Model for 26 Multiple Choice Items

```
1PLS

>PRO RA PA NI=26 NG=1 NP=1000;

>TES AL L1;

>EST NC=100;

>END;

        2

01

11111111111111111111111111

N

(26A1,21X,F2.0)
```

B. **MULTILOG Command for 2_Parameter Model for 26 Multiple Choice Items**

```
2PLS
>PRO RA PA NI=26 NG=1 NP=1000;
>TES AL L2;
>EST NC=100;
>END;
         2
01
11111111111111111111111111
N
(26A1,21X,F2.0)
```

C. MULTILOG Command for 3_Parameter Model for 26 Multiple Choice Items

```
3PLS NORMAL PRIOR
>PRO RA PA NI=26 NG=1 NP=1000;
>TES AL L3;
>EST NC=100;
>PRI AL DK=1 PA=(-1.1,0.5);
>END;
    2
01
11111111111111111111111111
N
(26A1,21X,F2.0)
```

D. MULTILOG Command for Partial Credit Model for 21 Constructed Response Items

```
PCS
>PRO RA PA NI=21 NG=1 NP=1000;
>TES IT=(1(1)15) L1;
>TES IT=(16(1)21) NO NC=(3(0)4,5(0)2) HI=(2(0)4,4(0)2);
>TMA IT=(16(1)21) AK PO;
>TMA IT=(16(1)21) CK PO;
>EQU IT=(16(1)21) AK=1;
>FIX IT=(16(1)19) AK=2 VA=0.0;
>FIX IT=(20(1)21) AK=(2,3,4) VA=0.0;
>EST NC=100;
>END;
    5
01234
111111111111111111111
222222222222222222222
333333333333333333333
444444444444444444444
555555555555555555555
(26X,21A1,F2.0)
```

E. MULTILOG Command for Graded Response Model for 21 Constructed Response Items

```
GRS
>PRO RA PA NI=21 NG=1 NP=1000;
>TES AL GR NC=(2(0)15,3(0)4,5(0)2);
>EST NC=100;
>END;
    5
01234
111111111111111111111
222222222222222222222
333333333333333333333
444444444444444444444
555555555555555555555
(26X,21A1,F2.0)
```

F. MULTILOG Command for Generalized Partial Credit Model for 21 Constructed Response Items

```
GPCS
>PRO RA PA NI=21 NG=1 NP=1000;
>TES IT=(1(1)15) L2;
>TES IT=(16(1)21) NO NC=(3(0)4,5(0)2) HI=(2(0)4,4(0)2);
>TMA IT=(16(1)21) AK PO;
>TMA IT=(16(1)21) CK PO;
>FIX IT=(16(1)19) AK=2 VA=0.0;
>FIX IT=(20(1)21) AK=(2,3,4) VA=0.0;
>EST NC=100;
>END;
    5
01234
111111111111111111111
222222222222222222222
333333333333333333333
444444444444444444444
555555555555555555555
(26X,21A1,F2.0)
```

G. MULTILOG Command for 1-Parameter Logistic Model for 26 Constructed Response and Partial Credit Model for 21 Constructed Response Items

```
1PLPCS
>PRO RA PA NI=47 NG=1 NP=1000;
>TES IT=(1(1)41) L1;
>TES IT=(42(1)47) NO NC=(3(0)4,5(0)2) HI=(2(0)4,4(0)2);
>TMA IT=(42(1)47) AK PO;
>TMA IT=(42(1)47) CK PO;
>EQU IT=(42(1)47) AK=1;
>FIX IT=(42(1)45) AK=2 VA=0.0;
>FIX IT=(46(1)47) AK=(2,3,4) VA=0.0;
>EST NC=100;
>END;
    5
01234
11111111111111111111111111111111111111111111111
22222222222222222222222222222222222222222222222
33333333333333333333333333333333333333333333333
44444444444444444444444444444444444444444444444
55555555555555555555555555555555555555555555555
(47A1,F2.0)
```

H. MULTILOG Command for 2-Parameter Logistic Model for 26 Constructed Response and Graded Response Model for 21 Constructed Response Items

```
2PLGRS
>PRO RA PA NI=47 NG=1 NP=1000;
>TES AL GR NC=(2(0)41,3(0)4,5(0)2);
>EST NC=100;
>END;
    5
01234
11111111111111111111111111111111111111111111111
22222222222222222222222222222222222222222222222
33333333333333333333333333333333333333333333333
44444444444444444444444444444444444444444444444
55555555555555555555555555555555555555555555555
(47A1,F2.0)
```

I. MULTILOG Command for 3-Parameter Logistic Model for 26 Constructed Response and Generalized Partial Credit Model for 21 Constructed Response Items

```
>PRO RA PA NI=47 NG=1 NP=1000;
>TES IT=(1(1)26) L3;
>TES IT=(27(1)41) L2;
>TES IT=(42(1)47) NO NC=(3(0)4,5(0)2) HI=(2(0)4,4(0)2);
>PRI IT=(1(1)26) AJ PA=(1.7,1.0);
>PRI IT=(1(1)26) BJ PA=(0.0,2.0);
>PRI IT=(1(1)26) CJ PA=(-1.1,0.5);
>TMA IT=(42(1)47) AK PO;
>TMA IT=(42(1)47) CK PO;
>FIX IT=(42(1)45) AK=2 VA=0.0;
>FIX IT=(46(1)47) AK=(2,3,4) VA=0.0;
>EST NC=100;
>END;
    5
01234
11111111111111111111111111111111111111111111111
22222222222222222222222222222222222222222222222
33333333333333333333333333333333333333333333333
44444444444444444444444444444444444444444444444
55555555555555555555555555555555555555555555555
(47A1,F2.0)
```