
**Effects of Random Rater Error on Parameter Recovery of the
Generalized Partial Credit Model and Graded Response Model**

Keith A. Boughton, Don A. Klinger, and Mark J. Gierl

Center for Research in Applied Measurement and Evaluation
University of Alberta

Paper presented at the annual meeting of the National
Council on Measurement in Education, Seattle, WA

April 10 – 14, 2001

Effects of Random Rater Error on Parameter Recovery of the Generalized Partial Credit Model and Graded Response Model

The format of large-scale achievement testing has shifted over the past 15 years with an increased use of extended response items and performance-based assessments that are scored using polytomous scales (i.e., three or more response categories). For example, extended response items are used extensively in Canada for the Communication Skills Assessment in British Columbia (1993) and the School Achievement Indicators Project (Council of Ministers of Education, 1996) and in the United States for the Goals Assessment program (Psychological Corporation, 1994), the Stanford Achievement Test, the Golden State Exams, the Illinois Goal Assessment Program (IGAP), the Portfolio Assessment in Vermont, and the Core Assessment CRT Program (Council of Chief State School Officers, U.S.A., 1998, Council of Ministers of Education, Canada, 1999). In order to score such assessments, different rating scales have been developed and used. While Linn and Gronlund (1995) recommend the use of rating scales between 3 and 7 scale points (most rating scales tend to be within this range), other rating schemes can also be found (e.g., Blok, 1985). One of the continuing issues in performance-based assessments is the unreliability of rating scales due to raters (e.g., Wainer & Thissen, 1993). Due to practical constraints, item scoring is often completed by a pool of raters with each item being scored by one or more raters from the pool. Several techniques have been developed to address systematic rater error such as rater severity and leniency (Donoghue and Hombo, 2000; Hombo, Thayer, and Donoghue, 2000; Houston, Raymond, and Svec, 1991; Patz, Junker, & Johnson, in-press; Raymond and Viswesvaran, 1993). However, coefficients of interrater agreement are often in the range of 0.70 and 0.90, with much lower values being reported in the literature, signifying a moderate amount of random error in the rating process (Dunbar, Koretz, & Hoover, 1991; Fitzpatrick, Ercikan, & Yen, 1998; Klein, Stetcher, Shavelson, McCaffrey, Ormseth, BellComfort, & Othman, 1998).

Historically, achievement tests have used total test score to measure student achievement. Whether in a criterion or a norm-referenced framework, the theoretical foundation for such tests has largely been in the realm of classical test theory (CTT). However with the increased use of computer technology, there has also been a steady rise in the use of item response theory (IRT) to develop and score large-scale assessments. Recent research has focused on the use of

polytomous item response theory (PIRT) for achievement tests containing polytomously-scored items. As with their dichotomous counterparts, these models enable psychometricians to use examinee response vectors to obtain the examinees' ability and the item parameter estimates.

The generalized partial credit model (GPCM) (Muraki, 1992) and the graded response model (GRM) (Samejima, 1969) are both generalized versions of the two-parameter IRT model for use with polytomously-scored items since they both allow the discrimination parameter to vary across items. However, the two models differ in the function used to develop the overall probability function for the polytomously-scored items. While some research has been conducted on parameter recovery for these polytomous models (e.g., Fitzpatrick, Link, Yen, Burket, Ito, & Sykes, 1996; Maydeu-Olivares, Drasgow, & Mead, 1994; Reise & Yu, 1990), researchers acknowledge the need for further study in this area. Reise and Yu (1990) conducted research regarding parameter recovery in the GRM. Using simulated data, they were able to obtain stable estimates of the ability, slope, and item-category threshold parameters. However, the study was limited by the use of a single test length (25 items), the use of only five response categories, and item parameters that may not have been realistic for the GRM.

Given the prevalence and nature of assessments using polytomously-scored test items, the purpose of this study is threefold. First, it extends the research of Reise and Yu by examining the effects of the number of response categories (i.e., 4, 6, and 8) and test length (i.e., 4, 8, and 16 items) on ability and item parameter recovery. These conditions reflect the diversity of testing conditions found in many testing programs. Second, this study investigates the parameter recovery for both the GPCM and the GRM because, until now, there has been little research examining or comparing the two models. Given that there is a difference in the structure between the GPCM and GRM, it is important to compare the accuracy of the two models. Finally, this study investigates the effect of random rater error on the estimation of the ability and item parameters in both the GPCM and the GRM. The results of this study will provide practitioners and researchers with some guidelines regarding the accuracy of each model as well as the effects of differing response categories, test length, and rater accuracy on the estimation process.

The Generalized Partial Credit Model

Muraki's generalized partial credit model (1992, 1997) is an extension of the one-parameter polytomous model developed by Masters (1982). For a test item with k possible response categories, the GPCM determines the probability of selecting the k th category compared to the $k-1$ category. The probability function (P_{jk}) for the GPCM can be expressed as

$$P_{jk}(\mathbf{q}) = \frac{\exp\left(\sum_{v=0}^{k-1} Da_j(\mathbf{q} - b_{jv})\right)}{\sum_{c=0}^{m_j-1} \exp\left(\sum_{v=0}^c Da_j(\mathbf{q} - b_{jv})\right)},$$

where j is the item being analyzed, k is 1, 2, 3, ..., m_j , and m_j is the number of possible response categories, v is the response category being analyzed, $P_{jk}(\mathbf{q})$ represents the probability that an examinee with a given \mathbf{q} value will achieve category k on item j , c is the response categories 1, 2, ..., m_j , $D = 1.7$ and is a scaling constant which places the theta scale close to the metric of the normal ogive model, a_j is the slope parameter for each of the response categories, and b_{jv} is the item category parameter, at which a category score of k or $k-1$ is equally likely.

The GPCM is considered to be a divide by total model because the denominator in the probability function represents the total amount of information provided by a specific item (Thissen & Steinberg, 1986). Figure 1a contains a series of item category response functions for a six-category item. The a -parameter mediates the steepness and height of each of the curves, while the threshold parameters indicate the location where two adjacent response categories intersect. (i.e., the point where the probability of either score is equal).

The Graded Response Model

In contrast to the GPCM, Samejima's graded response model (1969, 1972, 1997) uses a series of two-parameter dichotomous probability functions to determine the overall probability function. This overall probability function is created for each category boundary. For example, with an item having four possible scores, the first dichotomous function is zero versus one, two, or three; the second function is zero or one versus two or three; the third function is zero, one, or two versus three. For an item with k possible categories, $k-1$ probability functions will be calculated using

$$P_{jk}(\mathbf{q}) = \frac{\exp[Da_j(\mathbf{q} - b_{jk})]}{1 + \exp[Da_j(\mathbf{q} - b_{jk})]}$$

where D is the scaling constant of 1.7, a_j is the slope parameter for item j , b_{jk} is the item-category threshold parameter for category k and item j . The probability of responding above the lowest category bound or higher ($P_o(\Theta)$) is set to 1 and the probability of responding beyond the highest category ($P_{k+1}(\Theta)$) is zero. The graded response model is an example of a difference model since the overall probability is found by subtraction (Thissen & Steinberg, 1986). In other words,

$$P(u_j = k|\mathbf{q}) = P(u_j \geq k|\mathbf{q}) - P(u_j \geq k + 1|\mathbf{q}).$$

Figure 1b contains a set of item category response functions for a six-category item using the GRM. Within the dichotomous IRT models, the slope is the same as the discriminating power. However, within the GRM, it is the combination of the slope parameter and the distance between adjacent category thresholds that mediates discrimination. The item-category threshold indicates the location where the probability of achieving a given response category is .5 for the dichotomous response function, and it represents the steepest point on the response function for that category. The same item parameters were used for each model in Figures 1a and 1b, demonstrating the differences in the response functions between the two models.

Method

Simulated data were generated using the computer program RESGEN 3 (Muraki, 1999) by specifying normal unidimensional latent trait distributions and polytomous item responses for both the GPCM and the GRM. Based on the findings of Reise and Yu (1990), a sample size of 2000 was used in all conditions in order to ensure stable parameter estimates. A model (GPCM vs. GRM) X response category (4, 6, and 8 categories) X number of items (4, 8, and 16 items) design was used, first, in the absence of rater error and then in the presence of random rater error. In order to obtain stable estimates, 30 replications of each condition were completed. These conditions were chosen to represent the types of variables (i.e., 4, 6, and 8 categories for 4, 8, and 16 items) often found in practice (cf. Council of Chief State School Officers, 1998).

In the case of the GPCM, a -parameters were generated such that they were uniformly distributed between 0.75 and 1.50 representing items with adequate but varied discriminations. In order to simulate items with increasing threshold values, random threshold values were generated between -2.25 and 2.25 with the first score category having the lowest value and the subsequent categories having a value greater than the previous value. Further, the range for each score category was limited to ensure the threshold values would be relatively equal. This step was necessary to obtain score distributions in both the GPCM and the GRM that would be similar. For items having more response categories, the distance between threshold values was decreased in order to keep the total range between -2.25 and 2.25. For a given number of response categories, parameters were generated for the 4-, 8-, and 16-item tests. Using these item parameters, examinee ability estimates and the associated item response vectors were generated using the computer program RESGEN. Then, the computer program PARSCALE (Muraki & Bock, 1997) was used to estimate the ability and item parameters using the examinee response vectors generated with RESGEN. The default settings were used in PARSCALE with the exception of the REPEAT command. The REPEAT command was necessary in order to obtain separate category parameters for each item.

In order to simulate rater error, a random selection of examinee item scores from the non-error conditions simulated using the RESGEN program were increased or decreased. Uniformly distributed random numbers between zero and one were generated and subsequently used to determine if a given score changed, the direction of the change (decrease or increase), and the extent of the change (either 1 or 2 response categories). The random numbers were selected in order to maintain a correlation of approximately 0.85 between each of the non-error and error conditions. Further, an error of one score point was more common than an error of two score points and the total number of errors increased as the number of score categories increased. PARSCALE was then used to estimate the ability and item parameters from the modified simulated response vectors. The estimates derived from PARSCALE for each model were rescaled onto their true parameters using a linear transformation method:

$$l_x(y) = x = s(X) \left\{ \frac{y - \bar{x}(Y)}{s(Y)} \right\} + \bar{x}(X),$$

where $l_x(y)$ is the transformed parameter for the estimates placed onto the scale of the true parameters; $\bar{x}(Y)$, $\bar{x}(X)$ and $s(Y)$, $s(X)$ are the means and standard deviations of the estimated and true parameter, respectively (Kolen & Brennan, 1996).

To evaluate the accuracy of the parameter estimates, the root mean square error (RMSE) values between the estimated and true values were calculated and compared across conditions and models. A small RMSE indicates close agreement between the estimated and true parameters.

Results

Table 1 contains the RMSE value for the ability and item parameter estimates across the test length, model, and error conditions when 4-, 6-, and 8-score points were evaluated. The results in each horizontal panel of Table 1 are discussed, starting with 4-score points. Figure 2 contains a visual summary of the RMSE values in this panel. For the no error conditions, the RMSE values were generally low for the ability parameter estimates for both models, albeit slightly higher for the GRM. Moreover, the RMSE values decreased as the number of items increased. The RMSE values were also low for the a -parameter estimates across models. These parameter estimates also improved as the number of items increased. Recovery of the b -parameters was also very good. Lower RMSE values were obtained as the number of items increased. The RMSE values were higher for extreme categories because few students received these extreme scores, and the smaller number of examinees at the extreme score points likely decreased the accuracy of the estimation process. However, as the number of items increased, the RMSE for the extreme values decreased. The categories in the middle of the scale did not have markedly different RMSE values as the number of items increased because the parameters were accurately estimated with four items.

For the error conditions, the RMSE values were noticeably higher for the ability parameter estimates across both models. The RMSE values decreased as the number of items increased. The RMSE values were also higher in the error conditions compared to the no error conditions for

the a -parameter estimates across models. Further, there was a large increase in the RMSE under the GPCM. These parameter estimates improved as the number of items increased. Recovery of the b -parameters was poorer in the error conditions compared to the no error conditions but improved as the number of items increased. The RMSE values were higher for extreme categories. These values also tended to be higher for the GRM compared to the GPCM.

The second horizontal panel contains the results when 6-score categories were evaluated. The results are also displayed in Figure 3. For the no error conditions, the RMSE values were generally low for the ability parameter estimates across models with slightly higher values for the GRM. Moreover, the RMSE values decreased as the number of items increased. The RMSE values were also low for the a -parameter estimates across models. These estimates also improved as the number of items increased. Recovery of the b -parameters was also very good. As with the 4-category conditions, the RMSE values tended to be higher for extreme categories. As the number of items increased, the RMSE values decreased for the extreme categories. Conversely, the mid-score category RMSE values were largely unaffected by the number of items.

For the error conditions, similar results occurred to the 4-category conditions with the RMSE values being noticeably higher for the ability parameter estimates under both models. The RMSE values decreased as the number of items increased for the Θ -estimates. The RMSE values were also higher in the error conditions compared to the no error conditions for the a -parameter estimates across models. As before, these values decreased as the number of items increased. Larger increases in the RMSE values were found for the GPCM. As with the 4-category conditions, recovery of the b -parameters was poorer in the error conditions especially for the extreme categories but improved as the number of items increased. Again, the RMSE values for the extreme categories were higher under the GRM compared to the GPCM.

The third horizontal panel contains the results when 8-score categories were evaluated. These results are displayed in Figure 4. For the no error conditions, the RMSE values were generally low for the ability parameter estimates for both models, and decreased as the number of items increased. Higher values were observed for the GRM. The RMSE values were also low for the a -parameter estimates across models. However, a decline in RMSE did not occur as the number of

items increased from 8 to 16. This result is unexpected given that research with dichotomous models has indicated the number of examinees required for accurate item parameter estimation is largely a function of the model (e.g., Hulin, Lissak, & Drasgow, 1982). The results of the current study suggest that, in the case of polytomous data, the number of examinees required for accurate parameter estimation is based on the model as well as the number of items, number of score categories, and the number of examinees in each score category (i.e., distribution of the examinees' data). Recovery of the b -parameters was very good; higher RMSE values were found for the extreme categories. In contrast to the 4- and 6-category conditions, the RMSE values for these extreme categories did not decrease as the number of items increased from 8 to 16. The leveling of the RMSE values suggests that, given 2000 examinees, there is potentially an upper limit to the number of items and score categories that can be accurately estimated. In other words, to decrease the RMSE values in this condition, more examinees are needed because a large number of parameters are estimated.

Compared to the no error conditions, the RMSE values for ability parameter estimates with random error were higher under both models. The RMSE values decreased as the number of items increased for the ability estimates. As in the previous conditions, higher RMSE values were observed for the GRM. The RMSE values were also higher in the error conditions compared to the no error conditions for the a -parameter estimates across models, with the largest values occurring under the GPCM. In contrast to the other score categories, these values did not decrease as the number of items increased from 8 to 16. Although the RMSE was lower for the 16 item conditions compared to the 4 item conditions, the lowest RMSE values were obtained under the 8 item conditions for both models. Recovery of the b -parameters was poorer in the error conditions compared to the no error conditions, especially for the extreme categories, with the largest values occurring under the GRM. As with the a -parameter estimates, there was an increase in the RMSE as the number of items increased from 8 to 16 items. The lowest RMSE values for both the GPCM and GRM occurred when 8 items were used.

Discussion

Over the last two decades, IRT research has largely focused on the dichotomous response models even though extended response items and performance-based assessments having

polytomously-scored items are increasingly being used in large-scale testing. Given the increased use of polytomously-scored items in testing, it is necessary to examine the PIRT models under different testing conditions. In the absence of random rater error, both the GRM and GPCM demonstrated good item and ability parameter recovery with item recovery being slightly superior. Increasing the number of items is the best way to improve ability estimates, while parameter estimates remain relatively stable with increased items or categories. The results suggest that the accuracy of the 2-parameter PIRT models used in this study is very similar in the absence of random rater error under the different testing conditions that were used in this study.

However in the presence of random error, an increase in the absolute error between the estimated and true parameters was found. In particular, the a -parameters were largely affected by rater error. Our analysis of the estimated and true a -parameters also revealed that the estimated values were systematically lower than the actual values. Not unexpectedly, random rater error lowers the discrimination of a polytomously-scored item. This finding has direct implications for practitioners. Lower item discrimination values translate into longer tests in order to achieve a given amount of test information. In the presence of random rater error, the a -parameter estimates are most likely an underestimate of the potential discrimination of an item. Unfortunately, low a -parameter estimates could signify a poor item, random rater error, or a combination of the two (i.e., we will never know in practice if an item has low discrimination because it is a poor item or because raters were not very accurate in rating the examinees). Thus, careful attention must be given to the precision of the scoring process in order to minimize random rater error and to obtain accurate parameter estimates. This would also include clearly defined scoring rubrics, increased training for raters, and continual monitoring of the scoring process.

This research provides practitioners with some guidelines regarding the accuracy of the GPCM and GRM under different testing designs and in the absence and presence of rater error. Given these results, further research into the effectiveness of multiple raters and the effects of systematic rater error in the recovery of item and ability parameters is needed. This work is currently being conducted by the authors.

References

- Blok, H. (1985). Estimating the reliability validity, and invalidity of essay ratings. Journal of Educational Measurement, *22*, 41-52.
- British Columbia Ministry of Education. (1993). Communication Skills Reading Assessment. Victoria, BC: Queens Press.
- Council of Chief State School Officers. (1998). Key State Education policies on K-12 Education: Standards, Graduation, Assessment, Teacher Licensure, Time and Attendance. Washington, DC: Council of Chief State School Officers.
- Council of Ministers of Education, Canada. (1999). Schedule of Current and Recent Assessments. [On-line]. <http://www.cmec.ca>.
- Donoghue, J. R., & Hombo, C. M. (2000, April). A comparison of different model assumptions about rater effects. Annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, *4*, 289-303.
- Fitzpatrick, A. R., Ercikan, K., and Yen, W. M. (1998). The Consistency between raters scoring in different test years. Applied Measurement in Education, *11*, 195-208.
- Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. Journal of Educational Measurement, *33*, 291-314.
- Hombo, C. M., Thayer, D. T., & Donoghue, J. R. (2000, April). A simulation study of the effects of crossed and nested rater designs on ability estimation. Annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Houston, W. M., Raymond, M. R., & Svec, J. C. (1991). Adjustments for rater effects in performance assessment. Applied Psychological Measurement, *15*, 409-421.
- Hulin, C. L., Lissak, R. I., Drasgow, R. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. Applied Psychological Measurement, *6*, 249-260.

Klein, S. P., Stetcher, B. M., Shavelson, R. J., McCaffrey, D., Ormseth, T., Bell, R. M., Comfort, K., Othman, A. R. (1998). Analytic versus holistic scoring of science performance tasks. Applied Measurement in Education, 11, 121-138.

Kolen, M. J., & Brennan, R. L. (1996). Test equating: Methods and practices. New York: Springer.

Linn, R. L., & Gronlund, N. E., (1995). Measurement and assessment in teaching (7th Ed.). Upper Saddle River, NJ: Prentice Hall.

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

Maydeu-Olivares, A., Drasgow, F., & Mead, A. (1994). Distinguishing between parametric item response models for polychotomous ordered data. Applied Psychological Assessment, 18, 245-256.

Muraki, E. (1990). Fitting a polytomous item response model to likert-type data. Applied Psychological Measurement, 14, 59-71.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement, 16, 159-176.

Muraki, E. (1993). Information functions of the generalized partial credit model. Applied Psychological Measurement, 17, 351-363.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook Of Modern Item Response Theory (pp. 153-164). New York, NY: Springer.

Muraki, E. (1999). RESGEN: Item response generator [Computer program]. Princeton NJ: Educational Testing Service.

Muraki, E., & Bock, R. D. (1997). PARSCALE 3: IRT item analysis and test scoring for rating-scale data [Computer program]. Chicago, IL: Scientific software, Inc.

Patz, R. J., Junker, B. W., & Johnson, M. S. (in press). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. Journal of Educational and Behavioral Statistics.

The Psychological Corporation. (1994). GOALS: A performance-Based Measure of Achievement. San Antonio, TX: Harcourt Brace and Company.

Raymond, M. R., & Viswesvaran, C. (1993). Least squares models to correct for rater effects in performance assessment. Journal of Applied Measurement, 30, 253-268.

Reise, S. P., & Yu, J. (1990). Parameter recovery of the graded response model using MULTILOG. Journal of Educational Measurement, 27, 133-144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, 17.

Samejima, F. (1972). A general model for free-response data. Psychometrika Monograph Supplement, 18.

Samejima, F. (1997). In W. J. van der Linden & R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory (pp. 85-100). New York, NY: Springer.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.

Wainer, H., & Thissen, D. (1993). Combining multiple choice and constructed response test scores: Towards a Marxist theory of test construction. Applied Measurement in Education, 6, 103-118.

Table 1

The Root Mean Square Error Values for 4-, 6- and 8-Point Scales as a Function of Number of Item, Model, and Error Conditions

| | <u>No Error</u> | | | | | | <u>Error</u> | | | | | |
|-----------------|-----------------|----|----|------------|----|----|--------------|----|----|------------|----|----|
| | <u>GPCM</u> | | | <u>GRM</u> | | | <u>GPCM</u> | | | <u>GRM</u> | | |
| | 4 | 8 | 16 | 4 | 8 | 16 | 4 | 8 | 16 | 4 | 8 | 16 |
| 4 Points | | | | | | | | | | | | |
| ⊖ | 46 | 34 | 25 | 52 | 39 | 29 | 62 | 47 | 35 | 67 | 51 | 38 |
| A | 11 | 8 | 5 | 11 | 8 | 5 | 52 | 48 | 45 | 44 | 39 | 35 |
| B ₁ | 22 | 13 | 10 | 22 | 13 | 10 | 60 | 40 | 27 | 72 | 54 | 38 |
| B ₂ | 5 | 5 | 4 | 3 | 4 | 3 | 15 | 12 | 10 | 5 | 8 | 5 |
| B ₃ | 22 | 13 | 9 | 22 | 13 | 8 | 64 | 40 | 27 | 79 | 52 | 39 |
| 6 Points | | | | | | | | | | | | |
| ⊖ | 37 | 27 | 20 | 49 | 36 | 26 | 53 | 39 | 28 | 61 | 46 | 35 |
| A | 9 | 6 | 4 | 9 | 4 | 4 | 61 | 57 | 53 | 41 | 36 | 32 |
| B ₁ | 14 | 10 | 8 | 15 | 10 | 7 | 46 | 31 | 23 | 69 | 53 | 43 |
| B ₂ | 5 | 6 | 5 | 8 | 5 | 5 | 10 | 9 | 9 | 20 | 15 | 13 |
| B ₃ | 5 | 4 | 5 | 3 | 3 | 4 | 10 | 8 | 9 | 6 | 7 | 8 |
| B ₄ | 9 | 6 | 5 | 8 | 5 | 4 | 12 | 9 | 10 | 19 | 12 | 11 |
| B ₅ | 19 | 12 | 9 | 18 | 11 | 8 | 50 | 33 | 23 | 71 | 55 | 44 |
| 8 Points | | | | | | | | | | | | |
| ⊖ | 32 | 23 | 17 | 49 | 36 | 26 | 46 | 34 | 25 | 60 | 45 | 34 |
| A | 6 | 3 | 4 | 6 | 3 | 4 | 57 | 37 | 52 | 34 | 19 | 27 |
| B ₁ | 13 | 7 | 10 | 12 | 7 | 9 | 40 | 21 | 27 | 74 | 39 | 52 |
| B ₂ | 10 | 4 | 6 | 8 | 4 | 6 | 13 | 9 | 15 | 25 | 9 | 13 |
| B ₃ | 8 | 4 | 5 | 5 | 3 | 4 | 16 | 8 | 11 | 13 | 7 | 10 |
| B ₄ | 4 | 3 | 4 | 3 | 2 | 3 | 10 | 7 | 10 | 9 | 5 | 8 |
| B ₅ | 7 | 3 | 5 | 6 | 3 | 4 | 14 | 8 | 10 | 10 | 7 | 9 |
| B ₆ | 11 | 6 | 6 | 9 | 5 | 6 | 14 | 9 | 15 | 26 | 13 | 14 |
| B ₇ | 17 | 9 | 10 | 14 | 8 | 8 | 46 | 22 | 30 | 78 | 41 | 55 |

Figure Caption

Figure 1a. The item category response functions for the Generalized Partial Credit Model.

Figure 1b. The item category response functions for the Graded Response Model.

Figure 2. The RMSE values for the ability and item parameters across model, error conditions, and number of items (i.e., 4, 8, and 16 items from top to bottom) with 4-score points.

Figure 3. The RMSE values for the ability and item parameters across model, error conditions, and number of items (i.e., 4, 8, and 16 items from top to bottom) with 6-score points.

Figure 4. The RMSE values for the ability and item parameters across model, error conditions, and number of items (i.e., 4, 8, and 16 items from top to bottom) with 8-score points.

Figure 1a.

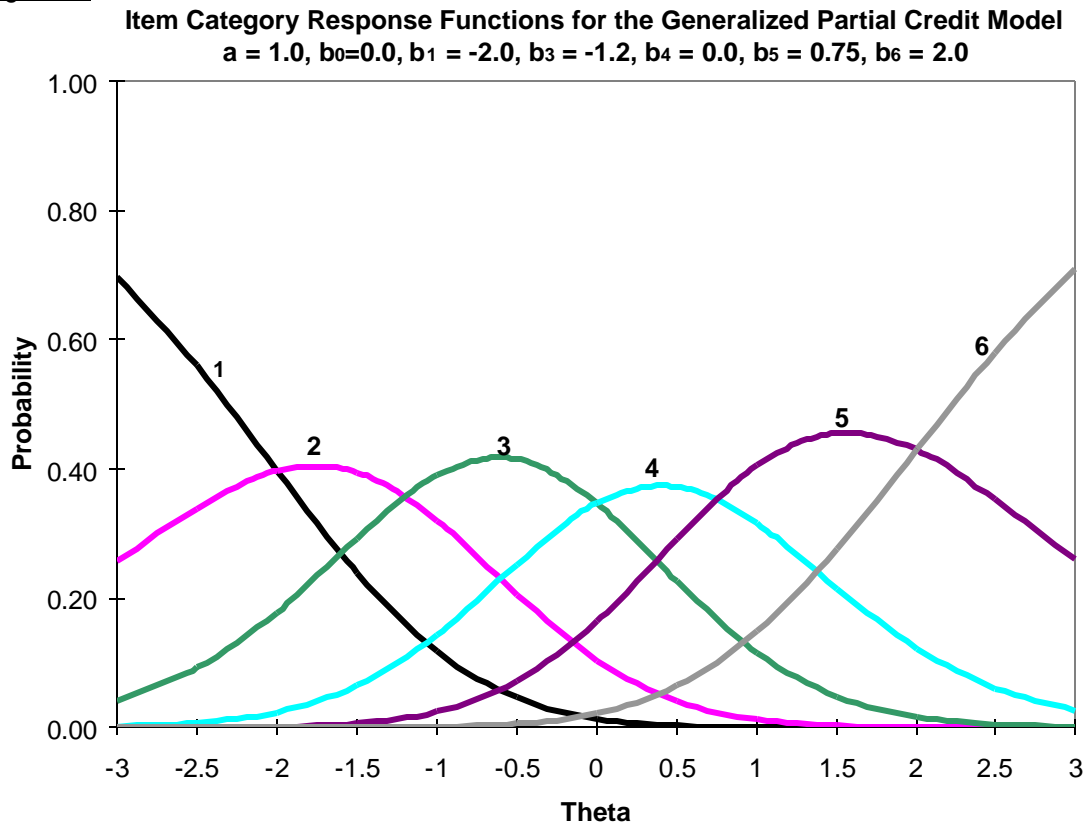


Figure 1b.

