Investigation of IRT Parameter Recovery and Classification Accuracy in Mixed Format

March 20, 2012

Melinda Montgomery
Dr. William Skorupski

University of Kansas

Paper presented at the Annual meeting of the Nation Council of Measurement in Education, April, 2012.

Abstract

The increased interest in innovative item types has led educational measurement to examine the accuracy and reliability of scoring more complex Item Response Theory (IRT) models including mixed format assessments. This study examines parameter recover, classification accuracy, and convergence rates of mixed format assessments using simulated data fit with PARSCALE. The independent variables examined include six item combinations, within four model combinations, across two samples sizes. The interaction between item combination, IRT model, and sample size impacted the root mean square error in the theta recovery, the classification accuracy, and the convergence rate.

Introduction

There are some who argue that the inclusion of some type of constructed response item will provide more information about examinees' understanding and ability than the current reliance on multiple choice items alone (Bennett, 2000; Zenisky & Sireci, 2001). While that argument has merit, it is unlikely that large scale state or national assessments will completely abandon the already existing pool of dichotomous items. It is more likely that polytomous items will be added to the existing assessments, as we see is the addition of an essay portion to the current SAT and ACT, leading to mixed format assessments. Since the goal of this additional piece of information, obtained from the polytomously scored items, is to increase the reliability and accuracy of the examinee's score on the assessment as a representation of the examinees true ability, it is reasonable to ask; which model type will perform the best is scoring mixed format assessments?

Since the goal of large scale state and national assessments is to determine an examinee's true ability, not just to compare performance on one assessment to performance another very similar assessment, ultimately scoring these assessments will employ some form of Item Response Theory model. It may turn out that a testlet response model or a multidimensional model will fit these mixed format assessment best. However, it is reasonable to ask whether a simple 2PL or 3PL IRT model can be used to score the dichotomous items in combination with a polytomous IRT model such as the Generalized Partial Credit Model (GPC) (Muraki, 1992) or the Graded Response Model (GRM) (Samejima, 1969). It is possible that, given a specific combination of polytomous and dichotomous items, one of these IRT model combinations will fit the data with very little error and will accurately classify examinees into pass/fail categories. The purpose of this study is to examine model fit and classification accuracy of mixed format

assessments using simulated data with four model combinations utilizing both a 2PL and 3PL model, six item combinations and two sample size.

Both the GRM and GPC models have been studied for many years and several authors have found differences in performance between the GRM and the GPC. Van der Ark (2005) found that ordering of the expected latent trait was violated more often by the GRM than the GPC. DeMars (2008) confirmed those results but found that this result did not lead to differences in theta values matched on raw scores. Kang, Cohen, & Sung (2009) found that the GPC fit the data generated by the GRM better than the GRM, itself particularly in small sample sizes. This issue of fitting generated data to the other models will not be tested in this study, but may be considered in subsequent studies. It is clear that these polytomous models perform differently when used individually. How these models perform in combination with a 2Pl or 3PL IRT model is the focus of this study.

Simulated data was used in this study so that the recovered parameter values could be compared to the true parameter values. Additionally, since the true theta values are known, the comparison between the numbers of examinees classified into pass/fail categories in the replicated versus the numbers of examinees classified as pass/fail based on the known true scores can be directly evaluated. Since dimensionality is of particular concern when fitting mixed format assessments, simulated data allowed for the assurance of unidimensionality. If the models do not fit well with known unidimensional data what is the likelihood that the models will fit an actual assessments that often have some small measure of dimensionality? In fact, it is possible that differences in format alone between the dichotomously scored and polytomously scored items may create dimensionality.

Traub (1993) found that there can be a format effect resulting from examinees processing items differently. When a formatting effect occurs, the multiple choice and constructed response items may measure different abilities and cause the presence of multidimensionality in the test total score (Kim & Kolen, 2006). A number of authors have discussed the issue of dimensionality in mixed format assessments (Cao, 2008; Kamata & Tate, 2005; Kim & Kolen, 2006; Kim, Walker, & McHale, 2010; Lee, 2010; Yao & Schwarz, 2006). None of these studies provide a comprehensive analysis of the issue of dimensionality in mixed format assessment.

Since this data was generated to meet the assumptions of IRT models, including unidimensionality, it can be expected that the IRT model will fit the data well and recover the parameters with very little error. However, a preliminary study found issues with convergence and parameter recovery. While low convergence rates, error in parameter recovery, and poor classification in this simulated data cannot be attributed to dimensionality as traditionally defined, it is possible that these model combinations, in and of themselves, create a form of dimensionality in the form of noise in data that may cause the models to have low convergence rates and larger than expected errors in parameter recovery.

Methods

All of the combinations consisted of the same total test length. The length of the assessment was designed to approximate existing testing parameters as closely as possible, while allowing for the six item combinations. The ACT and SAT assessments are similar in terms of the number of items on the mathematics portion. The ACT utilizes 60 items while the SAT utilizes 54 mathematics items (ACT, 2011; Kaplan, 2001). Standardized state assessments commonly include four to eight items per indicator making the test length approximately 48

items. The Science and English portions of the ACT and the Language portion of the SAT also contain approximately 48 items. Based on these finding, this study considered a test length of 53 items.

The four model types were created by paring a 2PL model and a 3PL model with each of the models used to fit the polytomously scored items. The polytomously scored items were fit to either the Graded Response Model (Samejima, 1969) or the Generalized Partial Credit model (Murakis, 1992). The GRM essentially turns polytomous responses into dichotomous responses by considering the probability that an examinee provides a correct response at each threshold or higher, while the GPC considers the probability that an examinee selecting a particular response over the previous one. A preliminary study indicated that the 2PL plus GRM model tended to fit the data better but did not recover theta values or classify students into pass/fail categories as well as the 2PL plus GPC.

For the purpose of this study, the number of dichotomously scored to polytomously scored items were allowed to vary across the following six conditions:

- *Combination 1*: 13-40, 25% of the items dichotomously scored, 14% of the points from dichotomously scored items
- *Combination 2*: 18-35, 34% of the items dichotomously scored, 20% of the points from dichotomously scored items
- *Combination 3*: 27-26, 50% of the items dichotomously scored, 34% of the points from dichotomously scored items
- *Combination 4*: 35-18, 66% of the items dichotomously scored, 49% of the points from dichotomously scored items
- *Combination 5*: 40-18, 75% of the items dichotomously scored, 61% of the points from dichotomously scored items
- *Combination 6*: 45-8, 85% of the items dichotomously scored, 74% of the points from dichotomously scored items

The sample sizes used in this study consisted of 5000, and 10,000 examinees with a distribution N (0, 1) which were generated using WinGen (Han, 2007). It is reasonable to

assume that the participants are normally distributed given that this study is designed to provide information about large scale assessments.

For the 2PL plus GRM and the 2PL plus the GPC, the item parameters were set as $a \in U(0.2,2.0)$, $b \in N(0,1)$ and the constant $D = 1.702$. The same parameter values were used for the 3PL plus GRM and 3PL plus GPC with the addition of $c \in U(0,.3)$. One hundred data sets were replicated for each item combination within each model type across two sample sizes. A total of 4800 data sets were generated. The number of quadrature points was set to 101, the convergence criterion was 0.001, the EM cycles were set to 500 and the Newton steps were set to 100, the score estimation method was set to EAP. The EAP option on the SCORE command allows scale scores to be estimated by the Bayes (EAP) method where their posterior standard deviations serve as standard errors (SSI, 2005). All other choices were left at the default PARSCALE (Muraki & Bock, 2003) levels.

To evaluate the theta classification accuracy, average theta value for each participant was calculated across replications and compared to the true theta values generated by WinGin (Han, 2007). Furthermore, for each of the parameters, a, b, c, theta, the Root Mean Square Error (RMSE) and bias was calculated to establish the amount of error present in the recovered parameters. Below is the RMSE calculation for theta. This formula can be extended to the remaining parameters by substituting each of the other parameters for theta.

$$RMSE_\theta = \sqrt{\frac{\sum_{i=1}^{N}\sum_{i=1}^{R}\left(\hat{\theta}_{ir} - \theta_i\right)^2}{NR}}$$

The formal below for bias can likewise be altered to establish the bias for the other parameters.

$$BIAS_{\theta} = \frac{\sum_{i=1}^{N}\sum_{i=1}^{R}\left(\hat{\theta}_{ir} - \theta_i\right)^2}{NR}$$

In order to compare the pass/fail rate between the fitted models with the true score, a true theta score of 0.2088, estimated from the theta distribution served as cut scores for the 55% . The replicated theta scores were averaged by person across replication. Then the number of examines in each category was compared between the true theta values and the estimated theta values.

Finally, since convergence was an issue in the preliminary study, the number of replications, out of 100, that converge was recorded. These values were compared across model types and item combination to determine if there was a pattern to the convergence based on the models.

Results

Convergence

The GCP models (2PL and 3PL) maintained the most consistent convergence rate across samples sizes. Item *combination one* demonstrated a low convergence rate for the 5000 sample 2PL, but the rate was nearly 100% for the 10,000 sample and for both 3PL samples sizes. *Combination two* maintained a very low convergence rate across both samples sizes and both the 2PL and 3PL models. The 2PL GCP demonstrated a high convergence rate for all other combinations across both sample sizes (Figure 1). Overall the 3PL GCP produced more variability across sample sizes and item combinations as compared to the 2PLGCP model. In addition to a poor rate of convergence at *combination two*, the 3PL GCP also performed poorly

at *combinations three and five. Combination six* produced a low convergence rate at the 10,000 sample but nearly perfect convergence at the 5000 (Figure 2). That result was contrary to what was expected as larger sample sizes were expected to improve the rate of convergence.

The GRM model, for sample size 5000, performed the best across the 2PL and 3PL models for all item combinations, except *combination six* on the 2PL GRM model. However, the 10,000 sample size was very inconsistent across all of the item combinations and both IRT models. Overall, the 10,000 sample size performed the best at the first two and last two item combinations but produced a poor convergence rate *at combinations three and four* (Figures; 3, 4).

Issues with convergence using PARSCALE resulting from floating point errors have been reported. DeMars (2005) found that using prior distribution for the item parameters helped in some cases. In this study all of the models were fit using prior distributions. It has also been suggested that changing the constant from 1.7 to 1 can affect the singularity of the matrix in PARSCALE (DeMars, 2005). An adjustment to the constant was not considered for this study. Given that the cycles were set to 500 and the Newton step set at 100, the models had more than adequate opportunity to converge.

Parameter Recovery

In terms of parameter recovery it is expected that those models with low convergence rates will contain a large amount of error in parameter recovery. In general that was true for the GCP (2PL and 3PL) models with a few exceptions. Item *combination five* when fit to the 2PL plus GCP had a high rate of convergence across the two sample sizes but with a large amount of error in a, b, and theta-parameter recovery at the 5000 sample (Tables 1 and 3). As expected, based on preliminary study, the 2PL plus GRM was less consistent (Table 2). The RMSE for

theta increased linearly across the item combinations regardless of the convergence rate in the GRM. The a- and b-parameter RMSE for the 2PL plus GRM produced a pattern more in line with the convergence rate.

The RMSE for the a- and b-parameters in the 3PL plus GRM were more difficult to explain (Table 4). For example, *item combination two* had a fairly high convergence rate for both sample sizes, yet there was a large about of error in the 5000 sample b-parameter recovery. In addition, item *combination five* maintained a high convergence rate across the two samples sizes but produced a higher RMSE for both the a- and b-parameters. Overall the c-parameters produced very little recovery error.

To further examine the parameter recover, an ANOVA of the RMSE and Bias values for all parameters across sample size, model combination, and item combination was conducted. This 2 x 4 x 6 ANOVA (N = 3177) indicated that 99% (partial η^2=.991) of the variance in the Theta RMSE, 54% (partial η^2=.539) of the Theta Bias, and 40% (partial η^2=.403) of the a-parameter bias was attributed to the three way combination (Table 5).

Splitting the file by sample size and examining the interaction between item combination and model type on the dependent variable Theta_RMSE indicated that the item combination accounted for nearly all of the variability for each of the model types except the GRM models at the 5000 sample size. Item combination accounted for 50% of the variability in the 2PL_GRM, 57% of the variability in the 3PL_GRM models and nearly 100% of the variability in the other two models (Table 6). In the 10,000 sample size the item combination accounted for 86% of the variability in the 2PL_GRM model and 96% or more for all other models.

Follow-up analysis indicated most of the variability in a-parameter bias could be explained by the 2PL plus GCP model. In the 5000 sample the largest amount of

THETA_RMSE can be found in the 2PL plus GCP in combination with *item combination five* as well as in the 3PL plus item *combination three* (Figure 5). The Theta_RMSE at the 10,000 sample size is more difficult to explain due in part to the fact that two of the item combinations failed to converge (Figure 6). The effect size in the 2PL plus GRM and 3PL plus GRM increased from sample size 5000 to sample size 10000. This could be due in part to the fact that one of the item combination models in each of the GRM model combinations did not converge for the 10.0000 sample size, increasing the variability across the item combination for the GRM models across sample sizes. It should be noted that there was a great deal of variability across these models in terms of convergence, but in general, as the rate of convergence approached 100%, the RMSE for Theta decreased.

Classification

For both the 5000 and 10,000 sample, item *combination five* of the 2PL plus GCP model produced the most error in classifying examinees. This is understandable since this model and item combination also produced the most Theta_RMSE error. In the 10000 sample 48 examinees did not pass that should have and in the 5000 sample 32 examinees passed that should not have. In addition in the 10,000 sample at item *combinations three and four*, 44 and 30 examinees respectively were classified as passing that should not have passed (Table 7).

The 3PL plus GCP model was less accurate in classifying examinees. The 5000 model 45 examinees in *combination one* and 85 examinees in *combination four* were classified as passing that should not have and 35 examinees in *combination two* and 76 examinees in *combination six* were classified as failing that should not have failed. The 10,000 sample model passed 47 more examinee than should have passed at item *combination one* and failed 718 more examinees than should have failed at *combination two* (Table 8). Since only 19 replications

converged in the *combination two* model, this result is most likely a product of the low convergence rate.

As a general trend, the 5000 sample 2Pl plus GCP model produced more examinees classified as passing when most of the items were polytomous or most were dichotomously scored (Figure 8). Examinees were classified more often a failing when they should not have been when number of polotomously scored and dichotomously scored items were approximately equal. In the 10000 sample more examinees were classified as passing when the items were equally distributed. Furthermore, more examinees were classified as passing when there were more polytomously scored items and more were classified as failing when there were more dichotomously scored items. There did not seem to be a pattern to the classification for the GRM across sample sizes (Figure 9).

Discussion and Limitations

It is clear there is an interaction between the type of IRT model combination, item combination, and sample size in terms of model fit particularly the theta recovery. However, it is unclear whether this interaction is a function of those independent variables or of the rate of convergence. It may be possible to make some of the adjustments suggested by DeMars (2005) and increase the convergence rate sufficiently so that a clearer understanding of the interaction could be obtained. However, given that the data was generated to be unidimensional and that the cycles and Newton step were sufficiently long enough for the model to converge, the low rate of convergence may in fact be a function of noise in the data resulting from the model and item combinations. There is evident of this trend when considering both GPC models. There seems to be a pattern across both sample sizes in convergence based on the item combination.

Specifically the 3PL plus GCP model had a high convergence rate when the items are either mostly dichotomous, mostly polytomous, or evenly split. The 2Pl plus GCP has a low convergence rate when the assessment consists of mostly polytomously scored items. The GRM models are difficult to classify, however, it is clear the when the item were mostly polytomously scored or dichotomously scored, both GRM models had a high convergence rate. When the item combinations were more evenly split the GRM produced a low rate of convergence.

The ability to classify examinees into pass fail categories was inconsistent across sample sizes and model combination. In terms of percentage of examinees misclassified the numbers are mostly small. However, with frequently 35 to 45 examinees misclassified and as many as 70 examinees misclassified in several different model combinations, those numbers are troubling. Since the goal of these assessments is to more accurately determine examinee ability, this level of misclassification does raise some concern.

Certainly there are limitations to this study. While the assessment studied contained a reasonable number of items, item combinations, and model types, it is based on simulated data. It would be beneficial to fit real examinee data to the GCP models to see if the same pattern of model fit exists in the real data. More study is needed on the GRM as well. It is possible that fitting real data would clarify some of the remaining questions and add clarity to which IRT model combination will perform the best in scoring mixed format assessments.

References

ACT. (2011, November 26, 2011). The ACT Test, from http://www.act.org/newsroom/factsheets/view.php?p=160

Bennett, R. E., Morley, M., & Quardt, D. (2000). Three Response Types for Broadening the Conception of Mathematical Problem Solving in Computerized Tests. *Applied Psychological Measurement, 24*, 294-309.

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of Free-Response and Multiple-Choice Items. *Journal of Educational Measurement, 28*(1), 77-92.

Cao, Y. (2008). Mixed-format test equating: Effects of test dimensionality and common-item sets.

DeMars, C. E. (2005). Type I error rates for PARSCALE's fit index. *Educational and psychological measurement, 65*(1), 42-50.

DeMars, C. E. (2008). Polytomous differential item functioning and violations of ordering of the expected latent trait by the raw score. *Educational and psychological measurement, 68*(3), 379.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses.
Applied Psychological Measurement, 31(5), 457-459.

Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts, School of Education.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*(7), 499.

Kamata, A., & Tate, R. (2005). The Performance of a Method for the Long-term Equating of Mixed-Format Assessment. *Journal of Educational Measurement, 42*(2), 193-213.

Kaplan. (2001). Kaplan's Guide to Taking the New SAT, from http://www.kaptest.com/newsat

Kim, S., & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education, 19*(4), 357-381.

Kim, S., Walker, M. E., & McHale, F. (2010). Comparisons among Designs for Equating Mixed-Format Tests in Large-Scale Assessments. *Journal of Educational Measurement, 47*(1), 36-53.

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement, 47*(1), 1-17.

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

Muraki, E., & Bock, R. (2002). PARSCALE (Version 4.1). *Computer Software]. Lincolnwood, IL: Scientific Software International*.

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

SSI (2005) Scientific Software International. Parscale Manual. http://www.ssicentral.com/irt/example4-2.html

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. *Construction versus choice in cognitive measurement*, 29-44.

van der Ark, L. A. (2005). Stochastic ordering of the latent trait by the sum score under various polytomous IRT models. *Psychometrika, 70*(2), 283-304.

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format tests. *Applied Psychological Measurement, 30*(6), 469.

Zenisky, A. L., & Sireci, S. G. (2001). Feasibility review of selected performance assessment item types of the Computerized Uniform CPA Exam. *Laboratory of Psychometric and Evaluative Research Report*(406).

Figure 1

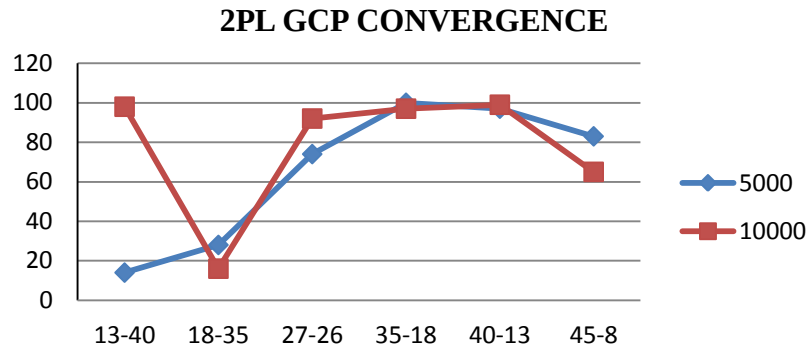


Figure 2

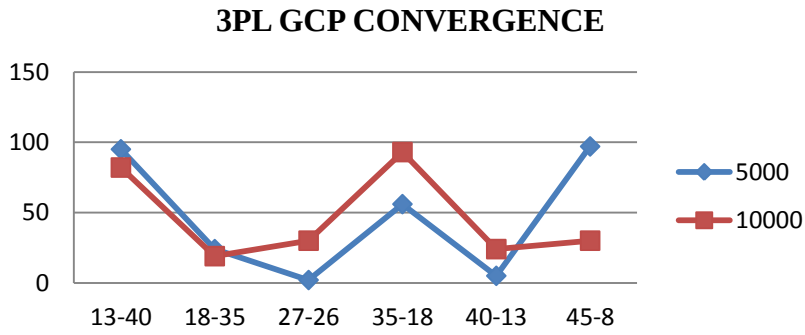


Figure 3

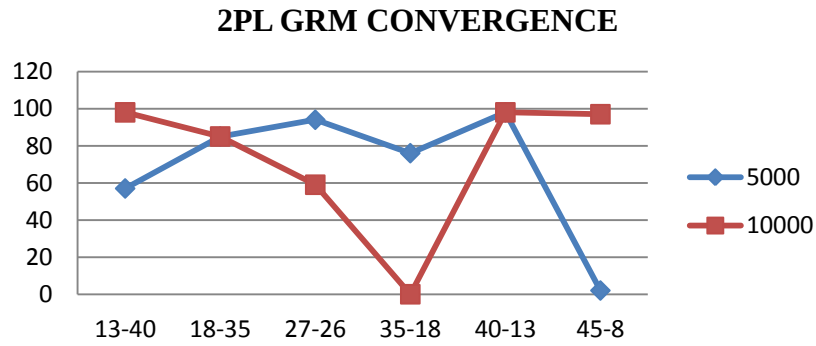
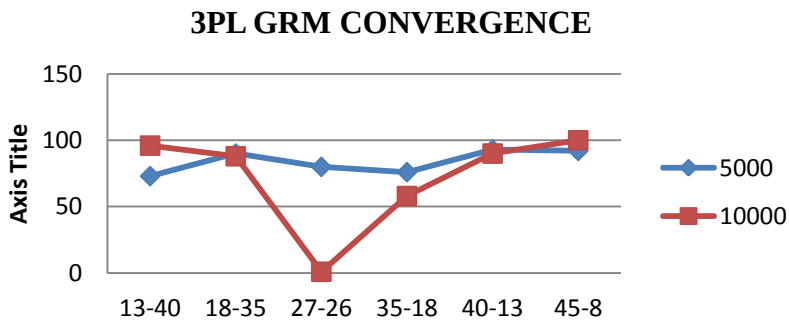


Figure 4



.

Table 1

			5000			10000		
Model	**Item Distribution**	**parm**	**Convergence**	**BIAS**	**RMSE**	**Convergence**	**BIAS**	**RMSE**
2PL + GPCM	13-40 (Di,Poly)		14/100			98/100		
		a		0.0223	0.0857		0.0116	0.0280
		b		0.0236	0.2350		0.0045	0.0197
		θ		0.0170	0.1728		0.0038	0.1723
	18-35		28/100			16/100		
		a		0.0426	0.6194		0.0527	0.6105
		b		0.0029	0.7495		-0.0044	0.8009
		θ		-0.0040	0.1648		0.0018	0.1712
	27-26		74/100			92/100		
		a		0.1268	0.0446		0.0063	0.0272
		b		0.0089	0.0293		0.0079	0.0202
		θ		0.0087	0.1725		0.0080	0.1771
	35-18		100/100			97/100		
		a		0.0100	0.0427		0.0182	0.0340
		b		-0.0104	0.0420		0.0003	0.0254
		θ		-0.0114	0.1815		-0.0018	0.1832
	40-13		97/100			99/100		
		a		0.0494	0.7881		0.0012	0.0291
		b		0.1912	0.9969		-0.0176	0.0309
		θ		0.0035	1.4063		0.0167	0.2030
	45-8		83/100			65/100		
		a		-0.0099	0.0475		-0.0058	0.0405
		b		-0.0053	0.0467		0.0003	0.0365
		θ		-0.0059	0.2171		0.0010	0.2065

Table 2

			5000			10000		
Model	**Item Distribution**	**parm**	**Convergence**	**BIAS**	**RMSE**	**Convergence**	**BIAS**	**RMSE**
2PL + GRM	13-40		57/100			98/100		
		a		0.0094	0.0662		-0.0065	0.0274
		b		0.1012	0.3356		-0.0190	0.0238
		θ		0.0214	0.1907		-0.0186	0.1703
	18-35		85/100			85/100		
		a		0.0176	0.1389		0.0232	0.1579
		b		0.2522	0.5198		0.0033	0.0565
		θ		0.0118	0.1921		0.0070	0.1932
	27-26		94/100			59/100		
		a		-0.0013	0.0445		0.0464	0.2131
		b		-0.0128	0.0108		-0.0308	0.2465
		θ		-0.0128	0.1818		-0.0033	0.1884
	35-18		76/100			0/100		
		a		0.0018	0.0439		N/A	N/A
		b		-0.0071	0.0364		N/A	N/A
		θ		-0.0073	0.2083		N/A	N/A
	40-13		98/100			98/100		
		a		0.0089	0.0432		0.0078	0.0356
		b		0.0040	0.0408		-0.0031	0.0351
		θ		0.0035	0.2155		-0.0015	0.2275
	45-8		2/100			97/100		
		a		0.5030	0.9223		0.0071	0.0278
		b		-0.4710	0.9720		0.0029	0.0285
		θ		0.0321	0.2471		0.0033	0.2154

Table 3

			5000			10000		
Model	**Item Distribution**	**parm**	**Convergence**	**BIAS**	**RMSE**	**Convergence**	**BIAS**	**RMSE**
3PL + GPCM	13-40		95/100			82/100		
		a		-0.0155	0.1087		0.0058	0.0405
		b		-0.0022	0.7554		0.0010	0.4303
		c		0.0030	0.0518		-0.0061	0.0587
		θ		0.0068	0.1683		0.0092	0.1743
	18-35		24/100			19/100		
		a		-0.0200	0.0674		-0.0068	0.0523
		b		-0.0545	1.1356		-0.0465	0.9364
		c		-0.0037	0.0674		-0.0064	0.0639
		θ		-0.0160	0.1788		0.0059	0.1850
	27-26		2/100			30/100		
		a		0.2075	0.5788		0.0015	0.0524
		b		0.3571	1.9780		-0.0072	0.6160
		c		0.1394	0.2201		-0.0016	0.0569
		θ		0.0028	0.8542		0.0070	0.1886
	35-18		56/100			93/100		
		a		-0.0220	0.1067		0.0033	0.1743
		b		-0.0891	1.0273		-0.0135	0.8046
		c		0.0078	0.0808		0.0036	0.0630
		θ		0.0337	0.2288		0.0001	0.2107
	40-13		5/100			24/100		
		a		0.0175	0.1083		0.0004	0.1805
		b		0.0396	0.5066		-0.0241	0.9706
		c		0.0174	0.0786		0.0024	0.0628
		θ		-0.0027	0.2107		-0.0060	0.2119
	45-8		97/100			30/100		
		a		-0.0002	0.1688		0.0036	0.0654
		b		-0.0752	0.9162		-0.0065	0.6473
		c		0.0133	0.0849		0.0076	0.0647
		θ		-0.0140	0.2345		-0.0086	0.2223

Table 4

			5000			10000		
Model	**Item Distribution**	**parm**	**Convergence**	**BIAS**	**RMSE**	**Convergence**	**BIAS**	**RMSE**
3PL + GRM	13-40		73/100			96/100		
		a		-0.0078	0.1370		-0.0110	0.0393
		b		0.0038	1.2828		-0.0390	0.9276
		c		0.0020	0.0590		0.0003	0.0552
		θ		0.0012	0.2024		-0.0045	0.1953
	18-35		90/100			88/100		
		a		-0.0045	0.0766		-0.0091	0.1028
		b		-0.0550	1.2077		-0.0637	0.4159
		c		-0.0028	0.0924		0.0013	0.0543
		θ		0.0073	0.1913		-0.0010	0.2068
	27-26		80/100			1/100		
		a		-0.0193	0.2133		N/A	N/A
		b		-0.0977	0.9570		N/A	N/A
		c		0.0106	0.0838		N/A	N/A
		θ		-0.0018	0.2215		N/A	N/A
	35-18		76/100			58/100		
		a		0.0168	0.1237		-0.0059	0.1142
		b		-0.1358	0.9020		-0.0218	0.9218
		c		0.0168	0.1237		0.0016	0.0751
		θ		-0.0107	0.2371		-0.0039	0.2200
	40-13		93/100			90/100		
		a		-0.0484	0.21053		-0.0121	0.1640
		b		-0.0957	1.18009		-0.0517	1.1291
		c		0.0174	0.09726		0.0014	0.0861
		θ		0.0222	0.21577		-0.0130	0.2530
	45-8		92/100			100/10		
		a		-0.0159	0.1443		-0.0068	0.0786
		b		-0.0679	0.9599		-0.0320	0.6006
		c		0.0059	0.0863		-0.0021	0.0552
		θ		-0.0130	0.2457		0.0047	0.2520

Table 5

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	A_Bias	1.736[a]	45	.039	17.184	.000	.198
	A_RMSE	71.527[b]	45	1.589	196.701	.000	.739
	B_Bias	15.314[c]	45	.340	20.235	.000	.225
	B_RMSE	326.223[d]	45	7.249	33.435	.000	.324
	THETA_Bias	.843[e]	45	.019	121.435	.000	.636
	THETA_RMSE	264.735[f]	45	5.883	35240.808	.000	.998
Intercept	A_Bias	.375	1	.375	166.866	.000	.051
	A_RMSE	18.761	1	18.761	2321.702	.000	.426
	B_Bias	.111	1	.111	6.605	.010	.002
	B_RMSE	194.521	1	194.521	897.143	.000	.223
	THETA_Bias	.174	1	.174	1130.039	.000	.265
	THETA_RMSE	77.416	1	77.416	463746.188	.000	.993
Sample	A_Bias	.250	1	.250	111.256	.000	.034
	A_RMSE	3.315	1	3.315	410.233	.000	.116
	B_Bias	.783	1	.783	46.574	.000	.015
	B_RMSE	16.723	1	16.723	77.128	.000	.024
	THETA_Bias	.288	1	.288	1870.719	.000	.374
	THETA_RMSE	2.221	1	2.221	13304.830	.000	.809
Model	A_Bias	.756	3	.252	112.296	.000	.097
	A_RMSE	3.795	3	1.265	156.557	.000	.130
	B_Bias	2.868	3	.956	56.854	.000	.052
	B_RMSE	76.028	3	25.343	116.882	.000	.101
	THETA_Bias	.299	3	.100	646.383	.000	.382
	THETA_RMSE	2.197	3	.732	4387.842	.000	.808
Combination	A_Bias	.353	5	.071	31.448	.000	.048
	A_RMSE	4.437	5	.887	109.825	.000	.149
	B_Bias	1.182	5	.236	14.051	.000	.022
	B_RMSE	15.926	5	3.185	14.690	.000	.023
	THETA_Bias	.335	5	.067	434.525	.000	.410
	THETA_RMSE	7.377	5	1.475	8838.453	.000	.934
Sample * Model	A_Bias	.249	3	.083	36.920	.000	.034
	A_RMSE	.170	3	.057	7.016	.000	.007
	B_Bias	1.003	3	.334	19.888	.000	.019
	B_RMSE	1.276	3	.425	1.962	.118	.002
	THETA_Bias	.198	3	.066	427.523	.000	.291
	THETA_RMSE	5.769	3	1.923	11518.465	.000	.917

Sample * Combination	A_Bias	.364	5	.073	32.416	.000	.049
	A_RMSE	1.878	5	.376	46.471	.000	.069
	B_Bias	.812	5	.162	9.651	.000	.015
	B_RMSE	2.949	5	.590	2.720	.019	.004
	THETA_Bias	.323	5	.065	419.158	.000	.401
	THETA_RMSE	6.079	5	1.216	7283.082	.000	.921
Model * Combination	A_Bias	.902	15	.060	26.792	.000	.114
	A_RMSE	24.930	15	1.662	205.669	.000	.496
	B_Bias	4.244	15	.283	16.823	.000	.075
	B_RMSE	51.143	15	3.410	15.725	.000	.070
	THETA_Bias	.548	15	.037	236.830	.000	.531
	THETA_RMSE	58.832	15	3.922	23494.555	.000	.991
Sample * Model * Combination	A_Bias	.887	13	.068	30.406	.000	.112
	A_RMSE	17.066	13	1.313	162.455	.000	.403
	B_Bias	2.836	13	.218	12.972	.000	.051
	B_RMSE	45.586	13	3.507	16.173	.000	.063
	THETA_Bias	.566	13	.044	282.093	.000	.539
	THETA_RMSE	57.225	13	4.402	26368.903	.000	.991
Error	A_Bias	7.032	3132	.002			
	A_RMSE	25.309	3132	.008			
	B_Bias	52.672	3132	.017			
	B_RMSE	679.090	3132	.217			
	THETA_Bias	.483	3132	.000			
	THETA_RMSE	.523	3132	.000			
Total	A_Bias	8.797	3178				
	A_RMSE	125.924	3178				
	B_Bias	68.129	3178				
	B_RMSE	1365.906	3178				
	THETA_Bias	1.327	3178				
	THETA_RMSE	506.361	3178				
Corrected Total	A_Bias	8.768	3177				
	A_RMSE	96.837	3177				
	B_Bias	67.986	3177				
	B_RMSE	1005.313	3177				
	THETA_Bias	1.326	3177				
	THETA_RMSE	265.258	3177				

a. R Squared = .198 (Adjusted R Squared = .186)

b. R Squared = .739 (Adjusted R Squared = .735)

c. R Squared = .225 (Adjusted R Squared = .214)

d. R Squared = .324 (Adjusted R Squared = .315)

e. R Squared = .636 (Adjusted R Squared = .630)

f. R Squared = .998 (Adjusted R Squared = .998)

Table 6

Tests of Between-Subjects Effects

Dependent Variable:THETA_RMSE

Sample	Model	Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
5000	2PL_GCP	Corrected Model	108.818[a]	5	21.764	5095605.747	.000	1.000
		Intercept	34.946	1	34.946	8182140.190	.000	1.000
		Combination	108.818	5	21.764	5095605.747	.000	1.000
		Error	.002	388	4.271E-6			
		Total	202.390	394				
		Corrected Total	108.820	393				
	3PL_GCP	Corrected Model	92.720[b]	5	18.544	19113.665	.000	.997
		Intercept	8.352	1	8.352	8608.511	.000	.969
		Combination	92.720	5	18.544	19113.665	.000	.997
		Error	.265	273	.001			
		Total	197.497	279				
		Corrected Total	92.984	278				
	2PL_GRM	Corrected Model	.076[c]	5	.015	81.552	.000	.502
		Intercept	2.271	1	2.271	12183.491	.000	.968
		Combination	.076	5	.015	81.552	.000	.502
		Error	.075	405	.000			
		Total	16.201	411				
		Corrected Total	.151	410				
	3PL_GRM	Corrected Model	.181[d]	5	.036	130.403	.000	.568
		Intercept	23.644	1	23.644	84959.625	.000	.994
		Combination	.181	5	.036	130.403	.000	.568
		Error	.138	495	.000			
		Total	24.227	501				
		Corrected Total	.319	500				
10000	2PL_GCP	Corrected Model	.081[e]	5	.016	3429.137	.000	.974
		Intercept	10.331	1	10.331	2190661.044	.000	1.000
		Combination	.081	5	.016	3429.137	.000	.974
		Error	.002	460	4.716E-6			
		Total	16.286	466				
		Corrected Total	.083	465				
	3PL_GCP	Corrected Model	.089[f]	5	.018	7621.206	.000	.993
		Intercept	7.736	1	7.736	3304583.550	.000	1.000
		Combination	.089	5	.018	7621.206	.000	.993
		Error	.001	272	2.341E-6			
		Total	10.898	278				
		Corrected Total	.090	277				
	2PL_GRM	Corrected Model	.197[g]	4	.049	682.616	.000	.864
		Intercept	16.531	1	16.531	229200.450	.000	.998
		Combination	.197	4	.049	682.616	.000	.864
		Error	.031	431	7.213E-5			
		Total	17.653	436				
		Corrected Total	.228	435				
	3PL_GRM	Corrected Model	.246[h]	4	.062	2739.648	.000	.964
		Intercept	20.249	1	20.249	901977.612	.000	1.000
		Combination	.246	4	.062	2739.648	.000	.964
		Error	.009	410	2.245E-5			
		Total	21.263	415				
		Corrected Total	.255	414				

a. R Squared = 1.000 (Adjusted R Squared = 1.000)

b. R Squared = .997 (Adjusted R Squared = .997)

c. R Squared = .502 (Adjusted R Squared = .496)

d. R Squared = .568 (Adjusted R Squared = .564)

e. R Squared = .974 (Adjusted R Squared = .974)

f. R Squared = .993 (Adjusted R Squared = .993)

g. R Squared = .864 (Adjusted R Squared = .862)

h. R Squared = .964 (Adjusted R Squared = .964)

Figure 5

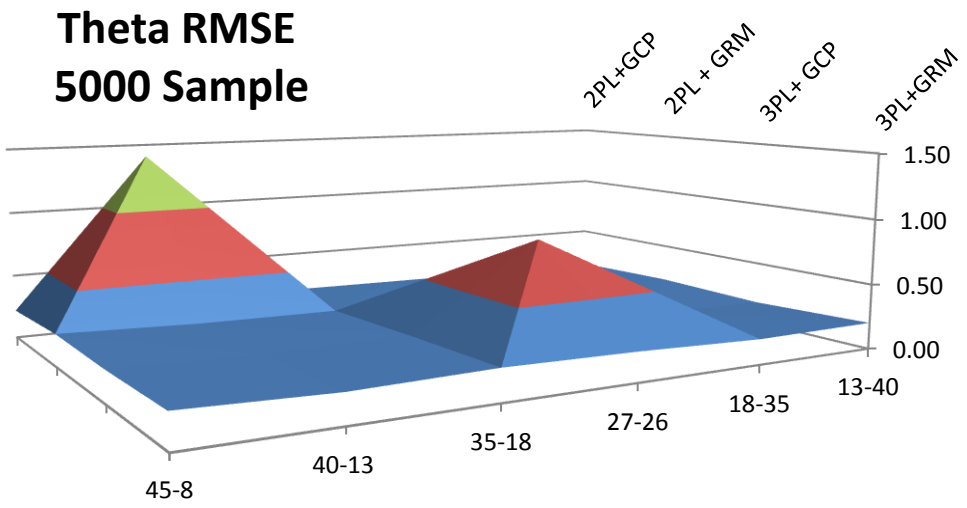


Figure 6

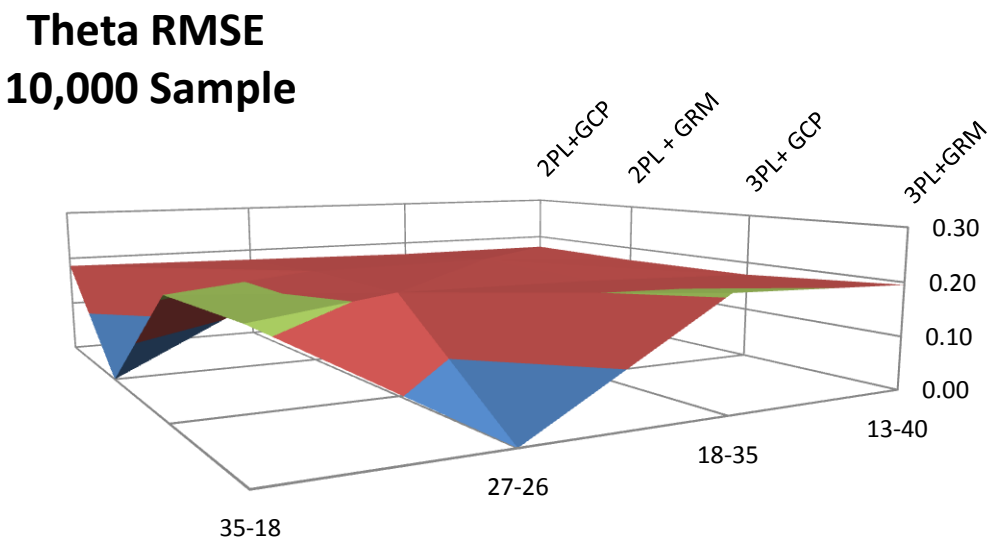


Table 7

Examinee Pass/Fail Classification

		5000		10000	
Model		**Passing ($\theta \geq 0.208$)**		**Passing ($\theta \geq 0.208$)**	
2PL + GPCM		True Theta	Theta Recovered	True Theta	Theta Recovered
	13-40	2058	2090	4141	4150
	18-35	2099	2091	4211	4208
	27-26	2091	2098	4138	4182
	35-18	2083	2061	4161	4191
	40-13	2049	2081	4249	4201
	45-8	2099	2117	4185	4186
2PL + GRM		True Theta	Theta Recovered	True Theta	Theta Recovered
	13-40	2062	2096	4255	4183
	18-35	2084	2112	4119	4145
	27-26	2079	2058	4207	4192
	35-18	2139	2115	N/A	N/A
	40-13	2049	2077	4152	4177
	45-8	2026	2125	4207	4225
3PL + GPCM		True Theta	Theta Recovered	True Theta	Theta Recovered
	13-40	2021	2063	4154	4201
	18-35	2142	2107	4142	3424
	27-26	2122	2131	4183	4190
	35-18	1979	2064	4127	4102
	40-13	2061	2084	4218	4219
	45-8	2122	2046	4233	4187
3PL + GRM		True Theta	Theta Recovered	True Theta	Theta Recovered
	13-40	2082	2084	4225	4191
	18-35	2071	2102	4130	4134
	27-26	2071	2102	N/A	N/A
	35-18	2121	2075	4203	4160
	40-13	2010	2056	4230	4195
	45-8	2082	2093	4167	4252

Figure 7

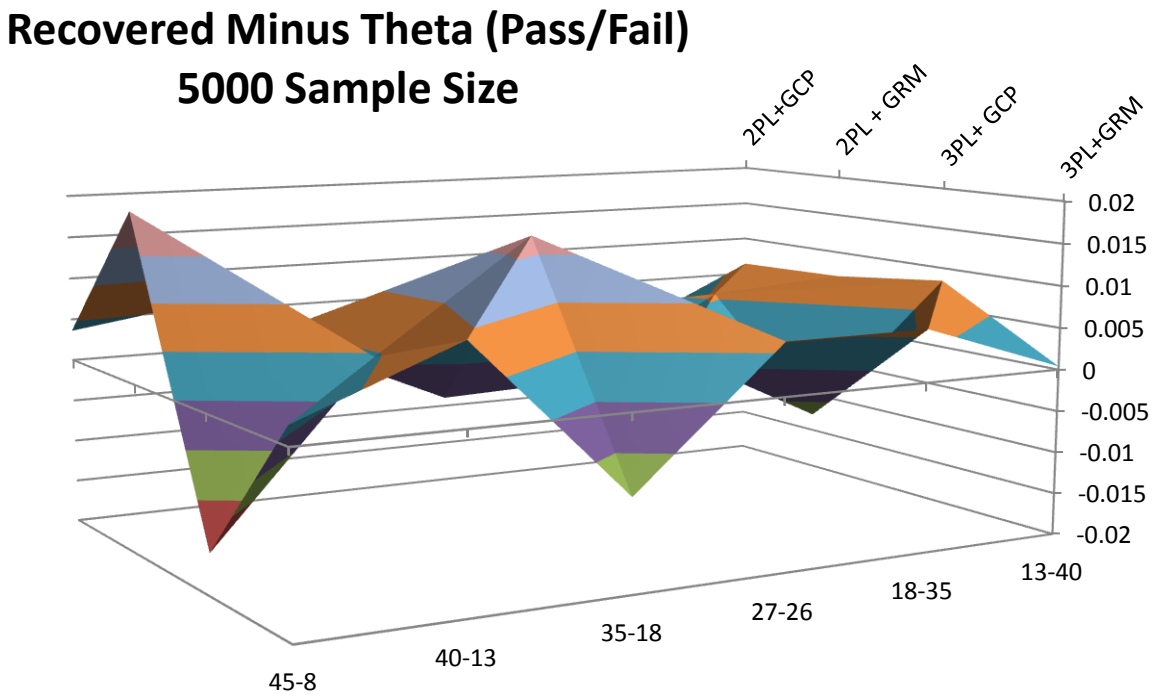


Figure 8