Relative Precision of Ability Estimation in Polytomous CAT: A Comparison under the Generalized Partial Credit Model and Graded Response Model

Shudong Wang

CAT*ASI

Tianyou Wang

University of West Michigan

Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April 2002. All correspondence should be sent to Shudong Wang, CAT*ASI, Three Bala Plaza West, Bala Cynwyd, PA 19004 -3481. E-mail: *Shudong_wang@asisvcs.net*.

Relative Precision of Ability Estimations in Polytomous CAT: A Comparison under the Generalized Partial Credit Model and Graded Response Model

Introduction

Computerized adaptive testing (CAT) using dichotomous scored item response models, such as Rasch or 1-PL, 2-PL, and 3-PL logistic models, are now found in many high- stakes educational and professional assessment programs. However, in practice, there are a few CAT applications that have based on items with more "nature" format of using polytomous models, such as Samejima's (1969) graded response model (GRM), Muraki's (1992) generalized partial credit model (GPCM), Master's (1982) partial credit model, Bock's (1972) normal model, Andrich's (1978) rating scale model, and etc.. In some situations, given the richer and more realistic form of assessment of polytomous scored items comparing to dichotomous scored items, the CAT with polytomously scored items could be a more valid and reasonable choice. in general, advantages of a polytomous model are: (a) the amount of item information provided by a polytomous item is greater than that of a dichotomous response item (Baler, 1992; Bock, 1972; Sympson, 1983; Thissen & Steinberg, 1984, Samejima, 1969). (b) the rate of detecting mismeasured examinees using a polytomous item is greater than that of a dichotomous response item. However, polytomous CATs are not widely used in the educational and professional testing settings because machine scoring of polytomous items is still difficult to achieve. Bennett, Steffen, Singley, Morley¸ and Jacquemin's (1997) research in computer scoring of open-end format items has shown new hope for polytmous item-based CAT.

In CAT, an examinee's ability is estimated after each item response is given. The ability estimates not only affect the final outcome of testing, but also affect which item is to be selected at each CAT stage. Four IRT-based ability estimates have been popular in CAT research and applications in the past: (a) Warm's weighted likelihood estimate (WLE), (b) maximum likelihood estimate (MLE), expected a posterior estimate (EAP), and (d) maximum a posterior estimate (MAP). Previous studies (Bock & Mislevy, 1982; Wang & Vispoel, 1998; Weiss & McBride, 1984; Wang, 1995; Wang, Hanson & Lau, 1999; Wang, 1999; Wang & Wang, 2001) have shown that the Bayesian methods, such as EAP and MAP, are severely biased toward the mean of the prior distribution and are thus unacceptable to many standardized testing programs. MLE was found to have smaller bias with a direction opposite to that of the Bayesian methods, (i.e., low ability examinees are negatively biased and high ability examinees are positively biased), but have a notably larger standard error (SE) than the Bayesian methods. Warm (1989) found that for 2- and 3- parameter IRT models, WLE were less biased than either MLE or Bayesian methods. Wang and Wang (2001) showed that for the Muraki's (1992) generalized partial credit mode (GPCM), WLE have better precision than MLE when the GPCM for fixed test length CAT was used in the CAT environment. It was also found that WLE and MLE have smaller bias but larger SE than both EAP and MAP, which is consistent with the previous finding. Samejima (1998) adopted Warm's approach and expanded it to the polytomous models and formulated it with the graded response model (GRM). Wang, Hanson and Lau (1999) and Wang & Wang (2001) demonstrated that Warm and Samejima's approach is a special case of a general approach proposed by Firth (1993) which has a more rigorous theoretical basis.

The GPCM and GRM models are the two most commonly used IRT models for polytomously scored item. Both models have item discrimination parameters, but GRM is a "difference model" and the GPCM is a "divide-by-total model" (Thissen & Steinberg, 1986). The two models differ in that with GPCM value of the item category parameters are not necessarily in successive order like those of the graded response model.

A few studies have examined the relative precision of those four ability estimation methods using different polytomous IRT models (Gorin, Dodd, Fitzpatrick, & Shieh, 2000; Wang, 1999; Wang & Wang, 2001). In particular, Wang and Wang (2001) systematically compared all four estimation methods under the GPCM model. However, no study has systematically compared the four ability estimation methods under the GRM and no study has made the comparison between the GRM and GPCM models under a similar set of conditions. The present study not only extends the Wang and Wang (2001) study to GRM model, but also makes some comparisons between the two models. It should be noted that the error indices under the two models can not be compared in a strict sense because the trait scales are slightly different for these two. Thus, the two models can only be compared in some general sense. For example, it can be examined if the relative precision of these ability estimation methods are consistent across the two models. The comparison may also provide some guidelines to practitioners about which model they should use in implementing CAT.

Objectives

The purposes of this paper are: (a) To evaluate relative precision (bias, SE, RMSE and others) of four ability estimation methods: Warm's weighted likelihood estimate (WLE), maximum likelihood estimate (MLE), expected a posterior estimate (EAP), and maximum a posterior estimate under two polytomous models in CAT. (b) To compare the ability estimations of the two polytomous models: the generalized partial credit model (GPCM) and the graded response model (GRM) under various computerized adaptive testing (CAT) conditions.

Method and Data

A Monte Carlo simulation method was used to evaluate the ability estimation methods between two polytomous models in this study. A real item bank consisted of 263 polytomously scored 1996 NEAP's science items (Allen, Carlson, & Zelenak, 1999) and a simulated item bank were used for this study. The item bank was originally calibrated using the GPCM model. To construct the item bank using the GRM model, item responses to the entire item bank were generated for a large sample of simulees from a normally distributed population. The response data were then calibrated using the GRM model using PARSCALE. Three items were deleted from the calibration process due to poor fit, thus reducing the bank size to 260 for the GRM model. These item parameters are treated as true item parameters in the simulation study. The items in the two smaller banks are randomly drawn from the larger bank with 260 items. Table 1 and 2 show the descriptive statistics for the estimates of item parameters of three item banks under the generalized partial credit model and graded response model.

The simulations were conditioned at 21 true ability values ranging from -4.0 to 4.0 by increments of 0.4 for both the GPCM and GRM. A CAT was simulated for 500 simulees at each of the 21 ability parameter points. A maximum-information item selection procedure was used. The effects of independent variables, size of item banks (260, 66, and 33), test termination rules (fixed test length and fixed test reliability), estimation methods (WLE, MLE, EAP, and MAP), and polytomously IRT models (GRM and GPCM) were examined by using both descriptive and inferential procedures. The dependent variables are the bias, standard error (SE), root mean square error (RMSE), fidelity (correlation of the estimated and true ability parameters), and administrative efficiency (the mean numbers of items needed to reach a criterion SE level).

Conditional indexes:

$$\mathrm{Bias}(\hat{\theta}) = \sum_{r=1}^{N} (\hat{\theta}_r - \theta) ,$$

$$SE(\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}\left(\hat{\theta}_r - \frac{\sum_{t=1}^{N}\hat{\theta}_t}{N}\right)^2} \quad ,$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\theta}_r - \theta)^2} \ ,$$

where θ is the true ability of simulees, which was used to generate responses in the simulation, $\hat{\theta}_r$ is the estimated ability for the *r*th replication, and N is the number of replications. The number of replications in this MC study is the analogue of sample size. Because the primary goal is to assess the relative accuracy of ability estimation methods, the significance of a statistic is tested, and the empirical sampling distributions for the statistics are generated. In order to minimize the sample variance and increase the power to detect the effects of interest, a large number of replications are desired. In this study, relative accuracy is assessed by comparing the differences between the ability parameter estimates and the true ability across replications. In such a study, 500 replications are considered sufficient (Stone, 1993).
The RMSE can be separated into two components, Bias and SE ($RMSE^2 = Bias^2 + SE^2$).

Overall indexes:

$$AVERAGE_{Bias} = \sum_{i=1}^{21} |\,Bias(\hat{\theta})\,|\,|\theta_i * weight(\theta_i),$$

$$AVERAGE_{SE} = \sqrt{\sum_{i=1}^{21} SE^2(\theta)|\theta_i * weight(\theta_i)},$$

$$AVERAGE_{RMSE} = \sum_{i=1}^{21} RMSE(\hat{\theta})|\theta_i * weight(\theta_i),$$

where the weight(θ_i) are quadrature weights based on the standard normal distribution, and the θ_i are the 21 equally spaced true ability levels that range from –4 to 4 in increments of 0.4.

Four experimental designs were used in the analyses of the overall indices. For the fixed-length tests, a 4 θ estimation methods x 3 bank size x 4 test length and a 4 θ estimation methods x 4 test length x 2 model completely crossed analysis of variance (ANOVA) designs were used. For the fixed reliability tests, a 4 θ estimation methods x 3 bank size x reliability level and a 4 θ

estimation methods x 3 reliability x 2 model completely crossed analysis of variance (ANOVA) designs were used.

Results

Conditional Indices

Figures 1 to 3 show the bias, SE, and RMSE of four ability estimates of fixed test length of 10 items under both models. It can been seen that the WLE has the smallest absolute biases and less SE over the entire or almost entire ability range among all the methods for both GRM and GPCM. Both WLE and MLE have considerably less bias than the two Bayesian methods for both models. Both models have approximately the same precision characteristics along almost all ability levels for both fixed test length CATs although they are not strictly comparable.

Figures 4 to 6 show the bias, SE, and RMSE of four ability estimates of fixed test reliability of 0.9 under both models. First, for both models, the WLE and MLE have remarkably smaller bias than EAP and MAP, especially at both extreme ability levels. Second, for both models, all methods show the same amount of SEs, and last, for both models, WLE and MLE have smaller RMSE than EAP and MAP. In general, there is no large differences of bias, SE, and RMSE between GPCM and GRM.

In general, the results of graded response model agree with those for the generalized partial credit model (Wang & Wang, 2001).

Overall Indices

Table 3 summarizes the results of the three-way ANOVA of absolute bias, SE, and RMSE (averaged across θ levels) for the fixed test length termination and fixed reliability condition under the graded response model. In general, the results for the overall indices further support the results of conditional indices for both models. For the GRM, θ estimation methods and fixed test length termination rule that each accounted for 27.5% and 29.3% of the total variance of absolute bias had large influence on absolute bias while for the GPCM, the θ estimation methods had the most influence on absolute bias (Wang & Wang, 2001). θ estimation methods for the fixed reliability termination conditions under the GRM has the most influence on absolute bias - they accounted for 80.5% of the total variance of absolute bias and this results match the result of GPCM. Like GPCM, the each of fixed test length termination condition and fixed test reliability termination condition under GRM had large influences on RMSE (log of RMSE) and accounted for 51.1% and 90.9% of total variance of RMSE.

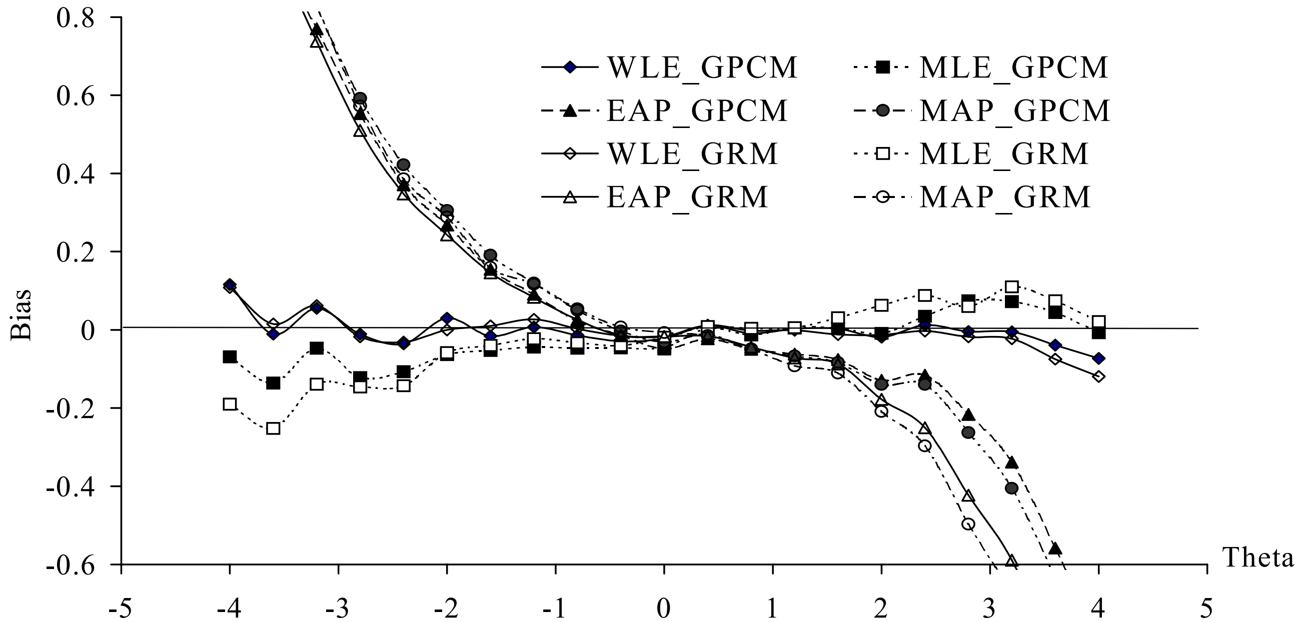


Figure 1. Bias curves of the ability estimation methods of two models, test length = 10, bank sizes = 263(260)

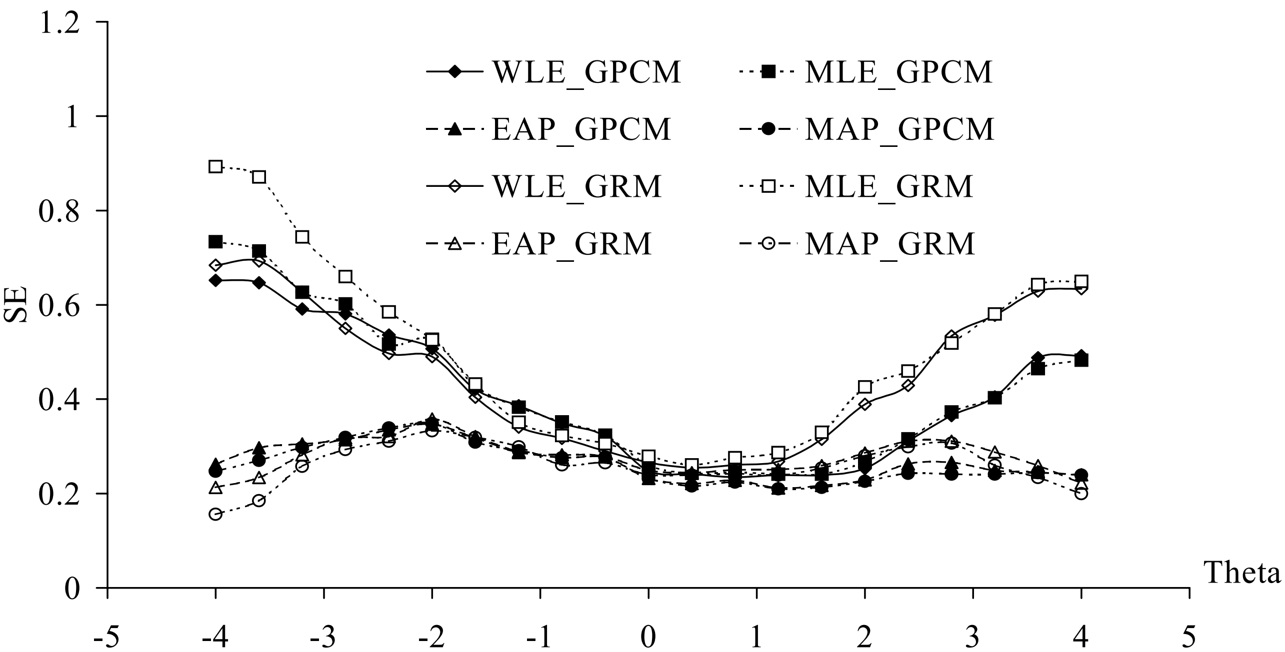
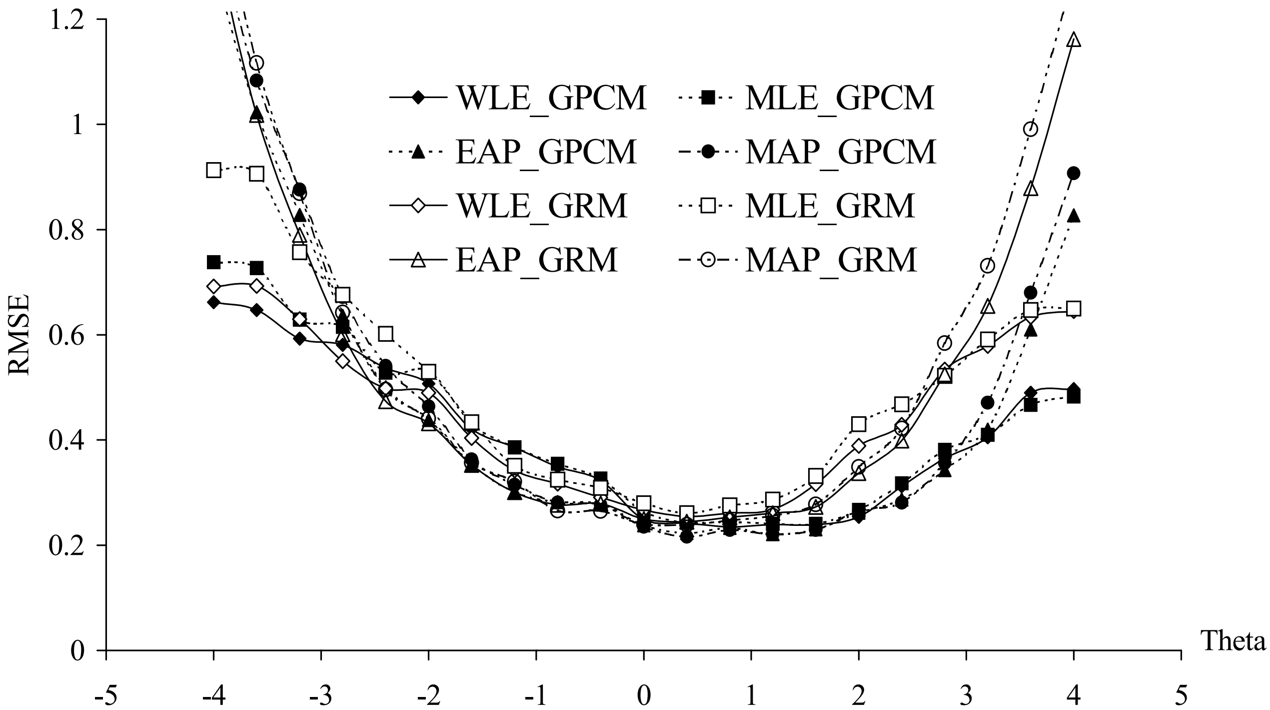
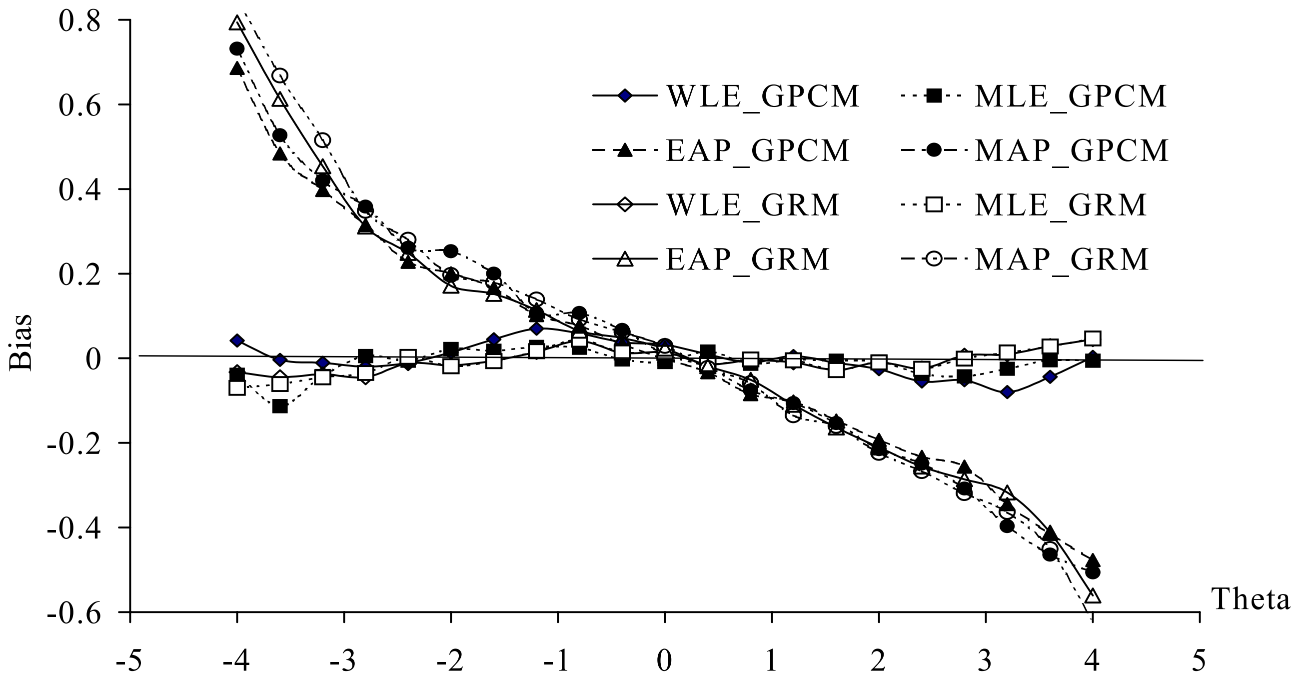


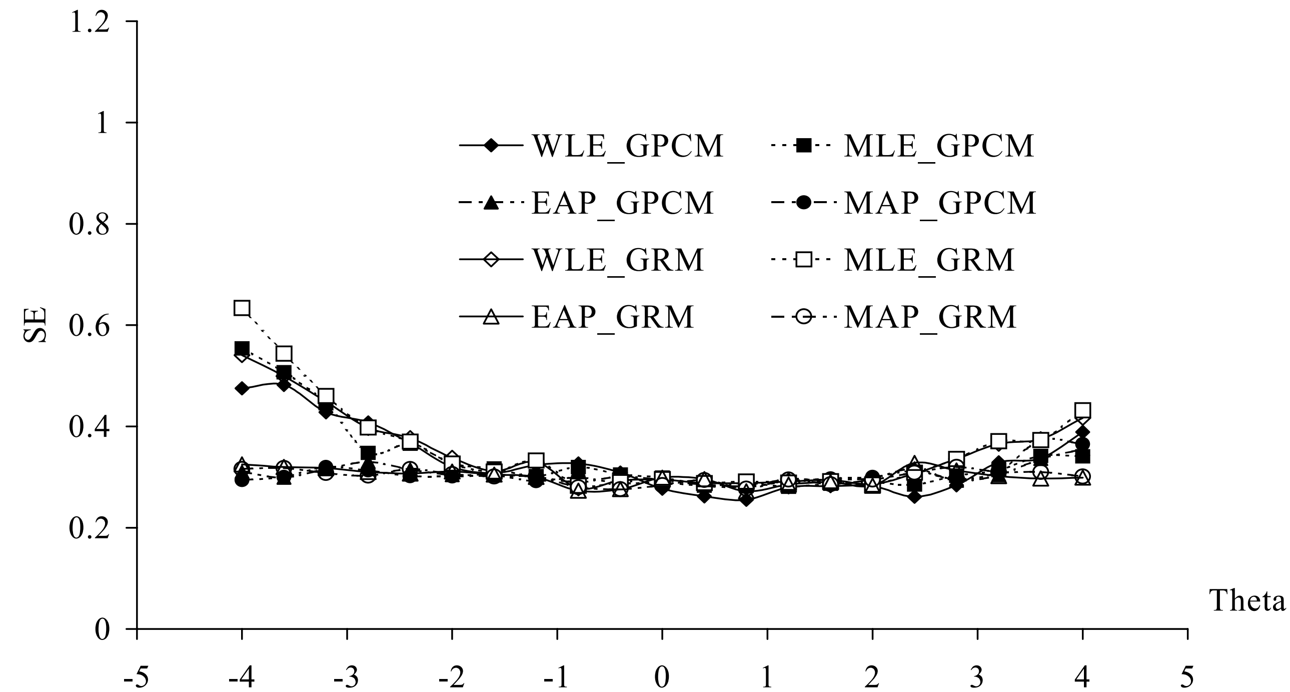
Figure 2. SE curves of the ability estimation methods of two models, test length = 10, bank size = 263(260)



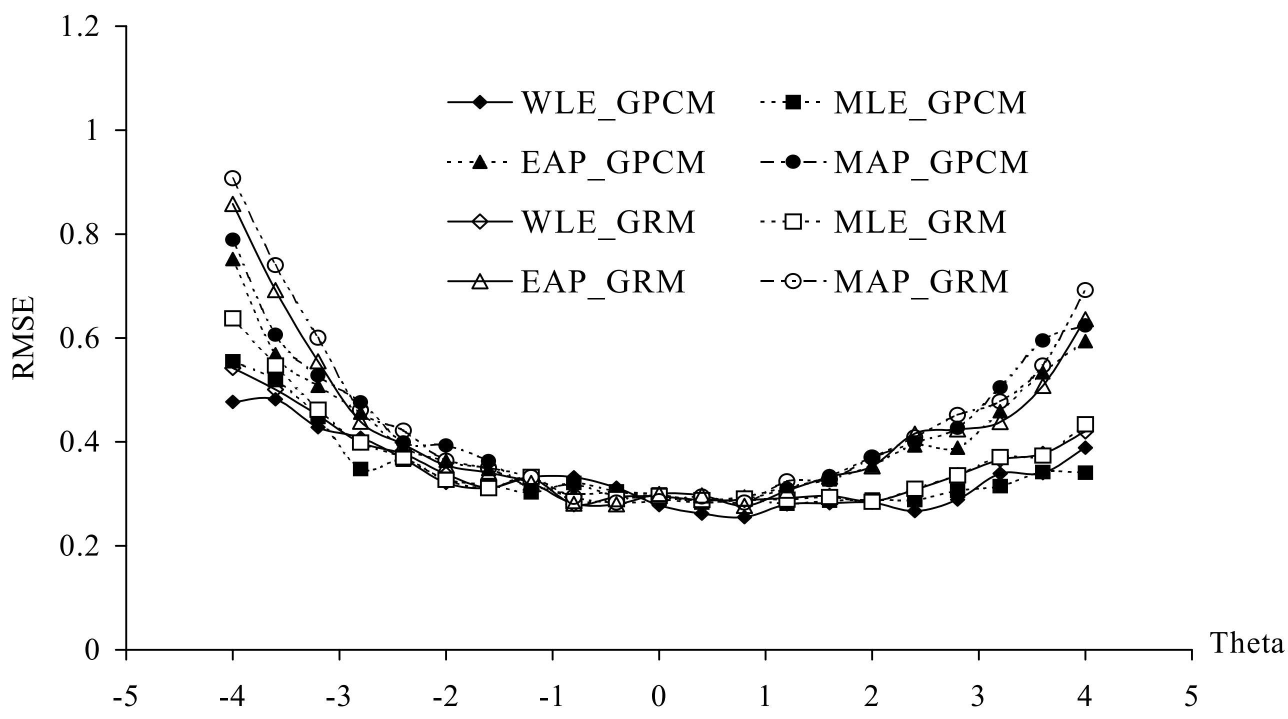
<u>Figure 3.</u> RMSE curves of the ability estimation methods of two models, test length = 10, bank size = 263(260)



<u>Figure 4.</u> Bias curves of the ability estimation methods of two models, Reliability = 0.9, bank size = 263(260)



<u>Figure 5.</u> SE curves of the ability estimation methods,
reliability = 0.9, bank size = 263(260)



<u>Figure 6.</u> RMSE curves of the ability estimation methods of two models,
reliability = 0.9, bank size = 263(260).

Table 4 provides the results of the three-way ANOVA of absolute bias, SE, and RMSE (averaged across θ levels) for the fixed test length termination and fixed reliability condition under the both models. Instead of testing the effect of bank size, the effect of model as one among three factors of method, test length, and model was tested.

For fixed test length termination condition, all main effects of method, test length, and model on absolute bias, SE, and RMSE were statistically significant. Although, model only accounted for 4% and 6.9% of the total variances of bias and SE, it accounted for 31.2% of the total variance of RMSE. All interaction effects for bias, SE, and RMSE are not statistically significant at 0.01 level except for interaction between method and test length for SE and RMSE. θ estimation methods had the most influence on absolute bias - it accounted for 31.8% of the total variance of absolute bias; test length had the most influence on SE - it accounted for 51.5% of the total variance of SE.

For fixed test reliability termination condition, all main effects of method, test reiability, and model on absolute bias, SE, and RMSE were statistically significant at 0.01 level except for the effect of model on SE. For bias, the three-factor interaction was not significant and all three two-factor interactions were significant. For SE and RMSE, all two-factor and three-factor interactions were not statistically significant. Again, θ estimation methods had the most influence on absolute bias - it accounted for 76.7% of the total variance of absolute bias; test reliability had the most influence on SE and RMSE - it accounted for 54.7% of the total variance of SE and 88.5% of the total variance of RMSE.

Summary and Discussion

This study examined the relative precision of four ability estimation methods (WLE, MLE, EAP, and MAP) under two polytomous models (GPCM and GRM) in CAT environment and comparison of relative precision between GCPM and GRM were provided. In general, for all four θ estimation methods, conditional and overall bias, SE, and RMSE decreased as the test length, test reliability, and item bank size increased. The magnitudes of the differences among the dependent variables decreased as the values of independent variables increased. For both models, WLE out performed MLE in terms of all the dependent variables studied, and WLE performed better than the Bayesian methods in terms of bias. MLE had less bias than both Bayesian methods. Both EAP and MAP showed more favorable results with SE and fidelity than either the WLE or MLE did; EAP did a better job than MAP for almost all conditions. Different test termination rules had significant impact on those dependent variables for given ability estimation methods, especially for WLE and MLE methods. Although the quality of item banks has vast effects on the conditional distribution of bias, SE, RMSE, and test efficiency (Wang & Vispoel, 1998), the size of item bank had less impact on the differences among the dependent variables than test termination rules. This study confirms Warm's conclusions that (a) WLE is unbiased to first order, while MLE, EAP, and MAP are biased, and (b) the WLE method has small variance over entire range of θ for fixed test length CAT testing.

In general, for the fixed test length, both GPCM and GRM model, estimation method and test length had same impact on bias, SE, and RMSE. But, the factor of model had the most impact on RMSE and it accounted for 31.2% of the total variance of RMSE under GRM. For the fixed test reliability, practically, the factor of model had almost no influence on bias, SE, and RMSE under GRM.

As CAT with polytomous models can be applied to a variety of polytomously scored items, and can be implemented in more and more testing programs, the search for a sound ability estimation method with a particular polytomously IRT model becomes even more important. MLE has been widely used in many CAT programs due to its small bias. The present study shows that under both GRM and GPCM for fixed test length rule, WLE not only reduced the bias of MLE to almost zero, but also reduced its SE as well. As computer scoring for polytomously scored items becomes more of a reality, the results of this study will have more practical significance.

References

Allen, N. L., Carlson, J. E., & Zelenak, C. A. (1999). *The NAEP 1996 Technical Report, NCES 1999-452*. Washington, DC: National Center for Educational Statistics.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 69-81.

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

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

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Firth, D. (1993). Bias reduction of maximum likelihood estimates [Correction: 95V82, p667]. *Biometrika, 80*, 27-38. *Bayesian estimation of θ.* August 26, 1983 (Internal Memorandum). Princeton, NJ: Educational Testing Service.

Gorin, Dodd, Fitzpatrick, & Shieh (2000). *Computerized adaptive testing with the Partial Credit Model: Estimation procedures, population distributions, and item pool characteristics*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

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

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

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychomtrika monograph, No. 17*.

Samejima, F. (1998). *Expansion of Warm's weighted likelihood estimator of ability for three-parameter logistic model to general discrete responses*. Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, April).

Sympson, J. B. (1983, June). *A new IRT model for calibrating multiple choice items.* Paper presented at the annual meeting of the Psychometric Society, Los Angeles CA.

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

Thissen, D. J., & Steinberg, L. (1986). Taxonomy of item response models. *Psychometrika, 51*, 567-577.

Wang, S. (1999). *The Precision of ability estimation methods for computerized adaptive testing using the generalized partial credit model*. Unpublished doctoral dissertation, University of Pittsburgh.

Wang, S. & Wang, T. (2001). Precision of Warm's weighted likelihood estimation of ability for a polytomous mode in CAT. *Applied Psychological Measurement, 25,* 317-331.

Wang, T., Hanson, B. A., & Lau, C. M. (1999). Reducing bias in CAT trait estimation: A comparison of approaches. *Applied Psychological Measurement, 23* 263-278.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Education Measurement*, *35*, 109-135.

Warm, T. A. (1989). Weighted likelihood estimation of ability in the item response theory. *Psychometrika, 54*, 427-450.

Weiss, D. J. & McBride, J. R. (1984). Bias and information of Bayesian adaptive testing. *Applied Psychological Measurement, 8*, 273-285.

Table 1
Descriptive Statistics for the Estimates of Item Parameters of the Three Item Banks, 1GPCM, 2GPCM, and 3GPCM, under the Generalized Partial Credit Model

Bank/ Parameter	No. Items	Mean	Median	S.D.	Minimum	Maximum
1GPCM	263					
a		0.549	0.522	0.229	0.105	1.871
b_1		0.713	0.720	2.011	-6.972	11.746
b_2		1.270	1.264	2.640	-17.381	13.926
b_3		1.034	1.004	2.371	-6.369	7.187
b_4		0.822	0.822	2.546	-3.159	4.924
2GPCM	66					
a		0.539	0.527	0.171	0.171	1.200
b_1		1.066	1.000	1.728	-3.204	7.399
b_2		1.679	1.491	2.519	-2.665	13.926
b_3		1.832	1.412	1.656	-0.856	5.506
b_4		4.270	4.270	0.535	0.535	4.925
3GPCM	33					
a		0.560	0.523	0.190	1.90	1.055
b_1		0.752	0.631	1.384	-2.738	3.437
b_2		1.695	1.684	2.495	-3.638	7.293
b_3		1.467	1.680	3.480	-6.369	7.187
b_4		2.000	2.000	0.000	2.000	2.000

Table 2
Descriptive Statistics for the Estimates of Item Parameters of the Three Item Banks, 1GRM, 2GRM, and 3GRM, under the Graded Response Model

Bank/ Parameter	No. Items	Mean	Median	S.D.	Minimum	Maximum
1GPCM	260					
a		0.658	0.668	0.347	0.180	2.206
b_1		-0.889	-0.568	2.066	-20.105	3.066
b_2		1.496	1.163	2.245	-9.962	10.627
b_3		1.837	1.941	3.475	-17.578	12.767
b_4		2.033	2.096	1.500	-0.158	4.312
2GPCM	66					
a		0.620	0.647	0.273	0.074	1.098
b_1		-0.834	-0.620	1.385	-5.565	3.066
b_2		1.590	1.108	2.291	-3.079	8.627
b_3		2.184	2.140	1.229	0.600	4.312
b_4		3.072	3.072	0.000	3.072	3.072
3GPCM	33					
a		0.678	0.693	0.333	0.065	1.301
b_1		-0.980	0.803	1.594	-6.853	2.688
b_2		1.374	1.125	1.683	-1.390	5.446
b_3		1.703	1.164	1.639	0.304	5.231
b_4		2.096	2.096	0.000	2.096	2.096

Table 3
Results of ANOVA with Fixed Test Length and Fixed Test Reliability Termination Rules for the GRM

Rule/ Effect	DF	Absolute Bias			Average SE			Average ln(RMSE)		
		F	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
Fixed Test Length										
Main Effects										
Method (M)	3	28.621	0.000	0.275	1299.514	0.000	0.189	25.607	0.000	0.101
Bank Size (S)	2	11.733	0.000	0.075	1667.641	0.000	0.162	79.169	0.000	0.209
Test Length (L)	3	30.517	0.000	0.293	3521.125	0.000	0.513	129.252	0.000	0.511
Interactions										
M x S	6	2.408	0.041	0.046	72.800	0.000	0.021	1.734	0.134	0.014
M x L	9	1.281	0.272	0.037	232.102	0.000	0.101	5.619	0.000	0.067
S x L	6	2.368	0.044	0.046	15.849	0.000	0.005	1.255	0.296	0.010
M x S x L	18	1.275	0.246	0.074	7.444	0.000	0.007	1.068	0.410	0.025
Error	48									
Fixed Test Reliability										
Main Effects										
Method (M)	3	4241.688	0.000	0.805	277.100	0.000	0.092	36.666	0.000	0.020
Bank Size (S)	2	33.022	0.000	0.004	769.988	0.000	0.170	104.525	0.000	0.037
Reliability (R)	2	761.547	0.000	0.096	2340.038	0.000	0.517	2554.179	0.000	0.909
Interactions										
M x S	6	9.392	0.000	0.004	28.488	0.000	0.019	3.142	0.014	0.003
M x R	6	215.322	0.000	0.082	56.738	0.000	0.038	6.259	0.000	0.007
S x R	4	16.170	0.000	0.004	246.731	0.000	0.109	13.417	0.000	0.010
M x S x R	12	4.359	0.000	0.003	38.731	0.000	0.051	3.752	0.001	0.008
Error	36									

Table 4

Results of ANOVA with Fixed Test Length and Fixed Test Reliability Termination Rules for the GRM and GPCM

Rule/		Absolute Bias			Average SE			Average ln(RMSE)		
Effect	DF	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
Fixed Test Length										
Main Effects										
Method (M)	3	58.130	0.000	0.318	54.575	0.000	0.157	16.924	0.000	0.052
Test Length (L)	3	52.078	0.000	0.284	178.535	0.000	0.515	132.094	0.000	0.015
Model (MO)	1	22.944	0.000	0.042	71.545	0.000	0.069	56.549	0.000	0.312
Interactions										
M x L	9	1.785	0.075	0.029	9.189	.000	0.080	2.811	0.004	0.054
M x MO	3	2.179	0.093	0.012	3.451	.018	0.010	1.596	0.193	0.072
L x MO	3	1.918	0.129	0.010	2.021	.113	0.006	0.236	0.871	0.082
M x L x MO	9	0.809	0.609	0.013	1.084	.378	0.009	0.820	0.598	0.161
Error	160									
Fixed Test Reliability										
Main Effects										
Method (M)	3	1635.523	0.000	0.767	5.956	0.001	0.052	24.952	0.000	0.037
Reliability (R)	2	396.923	0.000	0.124	93.240	0.000	0.547	903.584	0.000	0.885
Model (MO)	1	7.640	0.007	0.001	3.293	0.072	0.010	13.694	0.000	0.007
Interactions										
M x R	6	89.758	0.000	0.084	1.751	0.115	0.031	2.368	0.034	0.007
M x MO	3	5.296	0.002	0.002	0.839	0.475	0.007	2.197	0.092	0.003
R x MO	2	5.940	0.003	0.002	0.057	0.945	0.000	0.432	0.650	0.000
M x R x MO	6	0.229	0.967	0.000	0.065	0.999	0.001	0.625	0.710	0.002
Error	120									