Multidimensional Adaptive Testing Using the Weighted Likelihood Estimation: A Comparison of Estimation Methods

Fen-Lan Tseng & Tse-Chi Hsu

University of Pittsburgh

March 2001

Paper presented at the annual meeting of NCME, Seattle, 2001

Introduction

Selecting an appropriate ability (θ) estimation method is very crucial to a computer adaptive testing (CAT) system. Two general types of θ estimates have been proposed and researched in the CAT literature: maximum likelihood estimates (MLE) and Bayesian estimates. Among the Bayesian estimates are the maximum a posteriori (MAP), expected a posteriori (EAP), and Owen's normal approximation method (1975). Previous research (Wang & Vispoel, 1998; Warm, 1989; Weiss & McBride, 1984; Bock & Mislevy, 1982) found that MLE was generally less biased than the Bayesian methods but had larger standard error (SE) and root mean square error (RMSE). If the item bank lacked sufficient items at the extreme ability levels, the MLE could also produce bias in a direction toward the extremes of the θ scale. The bias of the Bayesian methods was in the direction towards zero on the θ scale if a standard normal prior was used. All of these estimation methods produce estimates that are biased to some degree. Bias in ability estimation is problematic in most standardized testing settings. For example, test score equatings will be erroneous and item banks will contain items with noncomparable statistics.

To reduce bias in the ability estimation, Warm (1989) derived a new method, weighted likelihood estimation (WLE) for the three parameter IRT model , which was proved to be less biased than MLE with the same asymptotic variance and normal distribution. The basic idea of WLE is to remove the first order bias term from MLE. Wang, Hanson, and Lau (1999) compared WLE with other estimation methods under a variety of CAT conditions for dichotomous models and found that the WLE method had even less bias than the MLE method when a fixed termination rule was used.

Most CAT procedures are based on unidimensional IRT models. However, some researchers (e.g. Reckase, 1985) argued that most tests required more than one ability to respond correctly. Ackerman (1991) pointed out that if a CAT item pool contained items from several content areas measuring different ability composites, examinees with different unidimensional abilities might receive disparate proportions of items from the diverse content

areas. Therefore, the estimated abilities in CAT were not comparable across the estimated unidimensional θ range. To circumvent this potential obstacle of the CAT, it is important to have a multidimensional computer adaptive test. However, the problems of CAT item selection and parameter estimation become more complex in a multidimensional context. Segall (1996) demonstrated that multidimensional computerized adaptive testing (MCAT) might be worth the added complications. Segall compared a unidimensional CAT for nine power achievement subtests in an Armed Services Vocational Aptitude Battery (ASVAB) to a multivariate CAT, fixing the covariance structure in the latter case so that the items in each subtest loaded on individual trait composites. He also implemented a Bayes modal estimation procedure that allowed the population covariances among the nine traits to enter into the solutions. By maximizing the determinant of the posterior variance-covariance matrix as the statistical objective function for the MCAT item selections, Segall demonstrated some detectable gains in the reliabilities of the outcome subscores when compared to simulated unidimensional CATs.

Statement of the Problem

The purpose of this study was to extend Warm's (1989) weighted likelihood estimation to a multidimensional adaptive test setting. The multidimensional 3PL compensatory model (M3PL) was used in the MCAT simulation, specifically in the case of three-dimensions. WLE was compared with the other three scoring methods: MLE, MAP and EAP. The goals of this investigation were (1)to conduct a Monte Carlo study on the weighted likelihood estimation of ability for a three-dimensional item response model in the CAT environment; (2) to evaluate the accuracy of ability estimates for multidimensional computer adaptive testing under WLE, MLE, MAP and EAP ability estimation methods; (3) to examine the effects of the degree of the intercorrelation between underlying abilities on ability estimation; and (4) to provide guidelines for the selection of a particular ability estimation method for the multidimensional IRT models in the CAT environment. MLE and MAP were selected for comparison because, like WLE, they are estimating the mode of an ability distribution

and they are commonly used in CAT (Ho & Hsu, 1989; Wang & Vispoel, 1998, Warm, 1989). EAP was selected because EAP should have the smallest mean square error over the population specified by the prior.

Multidimensional Weighted Likelihood Estimation (MWLE)

Warm's WLE was extended into a multidimensional WLE (MWLE). The derivation is given in Appendix A. An estimate satisfying (26) was called a multidimensional weighted likelihood estimate (MWLE).

$$\begin{bmatrix} \frac{\partial}{\partial\theta_1}\ln L(\vec{u}|\vec{\theta}) + \frac{J_1}{I_1} \\ \frac{\partial}{\partial\theta_2}\ln L(\vec{u}|\vec{\theta}) + \frac{J_2}{I_2} \\ \vdots \\ \frac{\partial}{\partial\theta_p}\ln L(\vec{u}|\vec{\theta}) + \frac{J_p}{I_p} \end{bmatrix} = \vec{0} \tag{1}$$

where

$$I_k = \sum_{i\in V} \frac{(P_i'(\vec{\theta}))^2}{P_i(\vec{\theta})Q_i(\vec{\theta})}, \quad k = 1, \cdots, p. \tag{2}$$

$$J_k = -3.4\sum_{i\in V} a_{ki}\frac{(P_i'(\vec{\theta}))^2}{P_i(\vec{\theta})Q_i(\vec{\theta})}[\frac{P_i(\vec{\theta}) - c_i}{1 - c_i} - \frac{1}{2}], \quad k = 1, \cdots, p. \tag{3}$$

Since the weighted likelihood equations (26) have no closed form solutions, an exhausting search procedure was used to compute the global minimum or a near global minimum. We define an objective function in the following.

$$\text{Min}\{|\frac{\partial}{\partial\theta_1}\ln L(\vec{u}|\vec{\theta}) + \frac{J_1}{I_1}| + |\frac{\partial}{\partial\theta_2}\ln L(\vec{u}|\vec{\theta}) + \frac{J_2}{I_2}| + \cdots + |\frac{\partial}{\partial\theta_p}\ln L(\vec{u}|\vec{\theta}) + \frac{J_p}{I_p}|\} \tag{4}$$

It is clear to see that when the objective function attains its minimum, (26) also attains its minimum. In our simulation, the number of dimensions was three (i.e., $p = 3$) and the values of θs ranged from -3 to 3. The range of each dimension was equally partitioned into 61 grids with a distance of 0.1 between two adjacent grids. We evaluated the objective function at each grid point and then selected the minimum from a total of $61 \times 61 \times 61$ values. The reason for selecting an exhausting search was to avoid local minima, which often occurred

through the iterative procedure (Luenberger, 1984). In the exhausting search, in order to obtain global minimum, the distance between two adjacent grids should be infinite small. In our simulation, 61 quadrature points at each dimension were equally partitioned between -3 and 3. Estimation precision would not drastically change for quadrature points of 60 or more.

Methods

Model

In this study, a multidimensional three-parameter logistic model (Segall, 1996) was used as a response generation model. This model is a generalization of the familiar unidimensional three-parameter logistic model to the multidimensional case. It could be written as

$$P_i(\vec{\theta}) = P(U_i = 1|\vec{\theta}) = c_i + \frac{1 - c_i}{1 + exp[-1.7\vec{a}_i^T(\vec{\theta} - b_i\mathbf{1})]} \tag{5}$$

where

U_i is the binary random variable, containing the response to item i. $U_i = 1$ if item i is answered correctly; and $U_i = 0$, otherwise,

c_i is the probability that a person will guess item i correctly,

b_i is the difficulty parameter for item i,

$\mathbf{1}$ is a $p \times 1$ vector of 1's,

Construction of Item Banks

The quality of the item pool is paramount to CAT performance. Two factors that determine the item pool quality are the locations of the items and their discrimination parameters. Items should be evenly and equally distributed throughout the θ continuum of interest (Urry, 1977; Weiss, 1982). Therefore, uniform distributions are assumed for all item parameters with the discrimination parameter at the interval of 0.4 and 2.0, the difficulty parameter at the interval of -3.0 and 3.0, and the guessing parameter at the interval of 0.0 and 0.3. These are the ranges of parameter values typically found in simulation studies

(e.g. Zhang & Stout, 1999; De Ayala, 1989; Urry, 1977). A three dimensional vector of the discrimination parameters was generated for each item, whereas only one scalar of the difficulty and guessing parameter was generated for each item.

Two item banks were generated and differed only in the distribution of discrimination parameters. The first item bank, which is named Bank I for the remainder of this paper, consists of 300 randomly generated items. The dimensional structure of Bank I is assumed to consist of three equally dominant factors. The discrimination parameters were designed to be uniformly distributed and to range between 0.4 and 2.0. In addition, the difficulty parameters were also generated based on uniform distributions and ranged from -3.0 to 3.0. The guessing parameters were also obtained based on the uniform distribution and ranged between 0.0 and 0.3. The summary statistics for the item parameters of Bank I are listed in Table 1.

Table 1: Summary Statistics of the Item Parameters in Bank I

	a_1	a_2	a_3	b	c
mean	1.192	1.2068	1.2041	-0.0033	0.1503
	(1.2)	(1.2)	(1.2)	(0)	(0.15)
std.	0.4471	0.4548	0.4529	1.72	0.0846
	(0.4619)	(0.4619)	(0.4619)	(1.7321)	(0.0866)
max.	1.9897	1.9941	1.9984	2.9878	0.2997
min.	0.4202	0.4015	0.4151	-2.9957	0.0006

*The numbers in the parenthesis are the population mean and standard deviation.

The other multidimensional test structure is called approximate simple structure. Zhang & Stout (1999) stated that a test is said to have approximate simple structure if the test consists of dimensionality-based and sufficiently separated clusters with each cluster consisting of dimensionally homogeneous items. To generate a three-dimensional test with approximate simple structure based on the same multidimensional 3PL model, each test item generated has a relatively large discrimination parameter on one dimension and relatively small discrimination parameters on the other two dimensions. Zhang & Stout (1999) defined the ability axis with a relatively large discrimination parameter for an item as the primary dimension

of the item, and the other ability axes as the secondary dimensions.

Among the 300 generated items in the second item bank, which is named Bank II for the remainder of this paper, the primary dimension of the first 100 items is θ_1. From items 101 to 200, θ_2 is the primary dimension. θ_3 is the primary dimension for the rest of the items. The discrimination parameters for the primary dimension of each item were designed to be uniformly distributed and to range between 0.4 and 2.0. On the other hand, the discrimination parameters for the secondary dimensions of each item were designed to be uniformly distributed and to range between 0.1 and 0.4. Note that the same samples of difficulty and guessing parameters were used in the two banks. The summary statistics for the item parameters of Bank II are listed in Table 2.

Independent Variables

The primary independent variable examined in this study was the ability estimation methods: WLE, MLE, EAP, and MAP. These methods were compared under two different item bank structures: One primary or multiple equally dominant factors

This study was specifically designed based on three-dimensional 3PL IRT model. To test the effects of correlated θ dimensions, three levels of correlation were chosen. They were no correlation (0.0), low correlation (0.3), and high correlation (0.6). In the previous researches on unidimensional CAT, simulation points were usually selected along the true ability or θ scale. Since there were three dimensions in our study and intercorrelations between each dimension was also considered, the previous sampling method was not appropriate for our situation. Therefore, 21 ability combinations were randomly generated for each correlation level. To reach the proposed criteria (i.e. correlation between sampled dimensions to be 0.0, 0.3, and 0.6, respectively), the combinations were obtained through two do-loops. First, 20000 samples of 21 ability points along a standard normal distribution were generated. Among these 20000 replications, qualified samples were selected if their sample means were close enough to 0.0 ($-0.005 < mean < 0.005$) and sample variances were

Table 2: Summary Statistics of the Item Parameters in Bank II

The first item to the 100th item					
	a_1	a_2	a_3	b	c
mean	1.1946	0.2494	0.2501	-0.2045	0.1541
	(1.2)	(0.25)	(0.25)	(0)	(0.15)
std.	0.4575	0.0839	0.0861	1.7434	0.0828
	(0.4619)	(0.0866)	(0.0866)	(1.7321)	(0.0866)
max.	1.9542	0.3962	0.3959	2.9678	0.2997
min.	0.4041	0.1057	0.1050	-2.9819	0.0006
The 101th item to the 200th item					
	a_1	a_2	a_3	b	c
mean	0.2489	1.2043	0.2492	0.0172	0.1441
	(0.25)	(1.2)	(0.25)	(0)	(0.15)
std.	0.0861	0.4533	0.0834	1.7265	0.0856
	(0.0866)	(0.4619)	(0.0866)	(1.7321)	(0.0866)
max.	0.3998	1.98	0.3923	2.9581	0.2980
min.	0.1048	0.4215	0.1031	-2.9957	0.0019
The 201th item to the 300th item					
	a_1	a_2	a_3	b	c
mean	0.2498	0.2510	1.2059	0.1773	0.1527
	(0.25)	(0.25)	(1.2)	(0)	(0.15)
std.	0.0799	0.0863	0.4418	1.6854	0.0857
	(0.0866)	(0.0866)	(0.4619)	(1.7321)	(0.0866)
max.	0.3998	0.3966	1.9718	2.9878	0.2966
min.	0.1027	0.1047	0.4130	-2.9015	0.0052

*The numbers in the parenthesis are the population mean and standard deviation.

around 1.0($0.99 < std < 1.01$). At the second (outer)loop, the first three qualified samples were selected to form the ability combinations $\vec{\theta} = (\theta_1, \theta_2, \theta_3)^T$. In the case of zero correlation, the computed correlation matrix was checked to see if the pairwise correlations were around 0.0 (i.e. $-0.05 < r < 0.05$). If the ability combination failed to meet the criterion, another 20000 samples of 21 ability points would be generated again until a set of qualified ability combinations were found.

In the cases of correlations 0.3 and 0.6, after the selection of the first three qualified samples to form the ability combinations, a Cholesky factorization was conducted to reach the desired covariance matrix. This followed by the same pairwise correlation computation.

Table 3: The 21 ability combinations at each correlation level

Correlation zero			Correlation 0.3			Correlation 0.6		
θ_1	θ_2	θ_3	θ_1	θ_2	θ_3	θ_1	θ_2	θ_3
-1.0608	-0.6022	-0.0613	-0.3625	0.6203	0.6061	0.7456	0.5141	0.2030
-0.1605	-0.2593	-2.2625	0.1090	0.8991	0.1424	-0.2206	0.5439	0.5845
-0.6665	1.5407	0.3809	-0.5980	0.3748	-1.2350	-0.1898	0.4468	-1.1742
-1.2094	-0.6288	-1.2545	-0.4047	-0.9719	0.1587	0.9927	0.4856	0.6723
0.8678	1.2602	-0.6045	-0.2161	0.7989	0.3751	-0.0339	-0.9653	-1.2353
0.3685	-1.3398	2.0521	1.5361	0.7659	-0.6542	-0.5380	-1.0996	-0.6165
0.2050	0.2415	1.3271	0.7892	-0.1639	-1.8583	-1.5095	-1.4189	-2.0371
1.7038	-0.6830	-1.1732	0.3090	0.5018	0.6207	0.5468	0.2746	-0.2876
-1.7767	1.7482	-0.3762	-1.1166	-0.6423	-2.2183	0.6452	1.1306	0.1933
-0.7355	-1.3467	-0.2333	2.0217	0.4583	1.7097	1.5993	0.9850	1.2278
0.5683	-0.7332	-0.2076	-0.5561	0.6430	0.3503	1.2558	-0.4750	0.7035
-1.6270	-0.8455	0.6147	-1.5828	-0.1699	-0.6448	-1.2861	-0.5675	1.1683
-0.3926	-1.0253	-0.5998	-1.9975	-1.9115	0.3195	-0.8293	-0.8132	-0.4131
0.8313	0.8387	0.5944	0.1941	-0.7866	-0.7366	0.5755	-0.6714	-0.0504
0.7481	0.3899	-1.0694	1.4661	-0.1238	0.6985	0.0120	-0.0264	-0.6960
-1.2381	0.8684	1.0415	0.0944	-0.3065	1.4019	0.7440	1.0081	1.1386
-0.2178	-0.8186	0.7637	0.2711	-1.6905	0.1297	0.1961	1.4592	1.1999
1.0263	-0.1641	-0.4467	-0.3313	-0.9797	-0.1959	-1.0055	0.8955	-0.9847
0.8255	0.8322	0.4964	1.0842	-0.0607	0.2823	-2.4622	-2.3222	-1.0200
0.6907	-0.7810	0.9007	-0.7425	0.1764	-0.3848	0.4408	1.2551	1.4407
1.1956	1.4806	0.1036	0.0436	2.6403	1.0670	0.2422	-0.7353	-0.1502

Like the zero correlation case, if the ability combinations failed to meet the preset criterion ($0.25 < r < 0.35$ for the 0.3 level, and $0.55 < r < 0.65$ for the 0.6 level), another 20000 samples of 21 ability points would be generated again until a set of qualified ability combinations were found. At each ability combination, 50 multidimensional adaptive testing replications were performed. Table 3 shows the generated 21 ability combinations at each intercorrelation level.

Dependent Variables

Evaluation of ability estimation methods were typically based on variables such as root mean squared error (RMSE), bias, and standard error (SE). RMSE is a measure of total error of estimation that has two components: systematic error (bias), and random or sampling

error (SE). These indexes are related as follows

$$\text{RMSE}^2 = \text{SE}^2 + \text{Bias}^2 \tag{6}$$

Indexes of bias, SE, and RMSE were computed at each θs combination based on the following formulas:

$$\text{Bias}(\widehat{\theta}_i) = \frac{1}{N}\sum_{j=1}^{N}(\widehat{\theta}_{i,j} - \theta_i) \tag{7}$$

$$\text{SE}(\widehat{\theta}_i) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\widehat{\theta}_{i,j} - \frac{1}{N}\sum_{k=1}^{N}\widehat{\theta}_{i,k})^2} \tag{8}$$

$$\text{RMSE}(\widehat{\theta}_i) = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(\widehat{\theta}_{i,j} - \theta_i)^2} \tag{9}$$

where $i = 1, 2$ or 3, and N is the number of replications. The number of replications in this study is the analogue of sample size. To minimize the sample variance and increase the power to detect the effects, a large number of replications are desired. Stone (1993) showed that 500 replications were considered sufficient. In this study, at each θ point, 50 replications were conducted due to the computation complexity.

The last criterion variable is the test information. Similar to Segall's (1996) formulation, the item information for the four estimation methods are computed based on Equations 11 and 13.

$$I_{rr}(\vec{\theta}) = -2.89\sum_{i\in V}\frac{a_{ri}^2 Q_i(\vec{\theta})(P_i(\vec{\theta}) - c_i)(c_i P_i(\vec{\theta}) - P_i(\vec{\theta})^2)}{P_i(\vec{\theta})^2(1-c_i)^2} \tag{10}$$

$$= \sum_{i\in V}\frac{[\frac{\partial P_i(\vec{\theta})}{\partial\theta_r}]^2}{P_i(\vec{\theta})Q_i(\vec{\theta})} \tag{11}$$

$$I_{rs}(\vec{\theta}) = -2.89 \sum_{i \in V} \frac{a_{ri} a_{si} Q_i(\vec{\theta})(P_i(\vec{\theta}) - c_i)(c_i P_i(\vec{\theta}) - P_i(\vec{\theta})^2)}{P_i(\vec{\theta})^2 (1 - c_i)^2} \quad (12)$$

$$= \sum_{i \in V} \frac{\frac{\partial P_i(\vec{\theta})}{\partial \theta_r} \frac{\partial P_i(\vec{\theta})}{\partial \theta_s}}{P_i(\vec{\theta}) Q_i(\vec{\theta})} \quad (13)$$

Experimental Design

A $4 \times 3 \times 21$ factorial design was used for each intercorrelation level to investigate the effects of different θ estimation methods and the intercorrelation of the underlying θs on the accuracy of ability estimation. For each intercorrelation level, a $4 \times 3 \times 21$ analysis of variance was performed on the bias. The four nominal levels of the ability estimation methods manipulated were WLE, MLE, EAP, and MAP.

The second factor in the design was the theta for each dimension included in each ability combination. At each intercorrelation level, 21 ability combinations were randomly generated and used as subjects in the simulation. For each ability combination, 50 replications were performed. Even though the differences among the three theta estimates were included in the analysis, we do not expect any differences among the three theta estimates since each of them was generated independently from the same specifications.

Computer Simulation Procedure

In this study, four ability estimation procedures were implemented: WLE, MLE, EAP, and MAP using Matlab 5.0 in a computerized adaptive testing system. As mentioned in the previous section, 21 ability combinations were generated at each correlation levels. At each ability combination point, 50 replications were made. This defined a sample of 1050 simulees for each intercorrelation level (i.e., 21 ability combinations $\times$ 50 replications for each combination).

Subroutines and Simulation Steps

The subroutines implemented for WLE, MLE, EAP, and MAP procedures included

Routine a. Computation of the probability of equation (5) and its first order derivatives. Input: item parameters $\vec{a_i}$, $\vec{b_i}$ and c_i of ith item and ability vector $(\theta_1, \theta_2, \theta_3)^T$. Output: $P_i(\vec{\theta})$ and $\frac{\partial}{\partial \vec{\theta}} P_i(\vec{\theta})$.

Routine b. Computation of the determinants of item information matrix from the first to the last item. Input: an ability vector $(\theta_1, \theta_2, \theta_3)^T$. Output: an array storing the determinants of item information matrix from the first to the last item.

Routine c. Searching the maximum determinant using exhausting search from the output array given by **Routine b** and returning the item index. Input: an ability vector $(\theta_1, \theta_2, \theta_3)^T$. Output: maximum determinant and item index.

Routine d. Computation of simulated item response. Input: item parameters $\vec{a_i}$, $\vec{b_i}$ and c_i of ith item and ability vector $(\theta_1, \theta_2, \theta_3)^T$. Output: 1 (correct answer) or 0 (incorrect answer).

The main procedure for WLE, MLE, MAP, and EAP included the following steps

Step 1. Read all the item parameters $\vec{a_i}$, $\vec{b_i}$ and c_i, $i = 1, \cdots, 300$.

Step 2. For each of the 21 ability combinations at each correlation level, **Step 3** to **Step 4** was done for 50 times.

Step 3. Randomly choose the initial item with its difficulty parameter ranging between -0.1 and 0.1. Then use this item index as the input for **Routine d** to determine correct or incorrect responses.

Step 4. After a response was generated, the provisional ability level was estimated using one of the four ability methods, WLE, MLE, MAP, and EAP. Based on this provisional estimate, the next item was selected using the maximum information procedure, **Routine c**. were administered.

Step 5. After 50 items were administered, the final ability estimates and error variance estimates were recorded.

Validation of the Generated Data

To address the goodness-of-fit of the model and the correlation between dimensions, a response matrix of 2000 simulees on 300 items was generated for each item bank.

Goodness-of-fit of the Models

Table 4 provides a summary of the results of the confirmatory factory analysis (CFA). Since LISREL83 only imports data file with 100 variables at the most, response matrix of 90 randomly selected items from the pool of 300 were used to perform the CFAs. For Bank II, 30 items were randomly selected from each of the three primary dimensions. For Bank I, 90 items were randomly selected from the 300 items because each item is loaded with three dominant dimensions. Each analysis was done with sample size of 2000. From the table, it can be seen that both one-factor and three-factor models fit the data except Bank I with correlation .30. It is evident by both nonsignificant chi-square statistics and adjusted goodness of indexes that are close to 1. For Bank II, the drops in χ^2 from one factor to three-factor are all greater than 11.34, which is significant at .01 for a χ^2 with $df = 3$. Obviously the data fit the three dimensional model better.

Correlations between Simulated Dimensions

From the LISREL output, polychoric correlation of independent variables was available, which is listed on Table 5, on the upper half of the matrix under 90 items which were randomly selected to run the CFAs. For example, in the case of correlation 0.3, the pairwise correlations are 0.7, 0.68, and 0.63. The second half of the matrix is based on the Pearson product moment correlations computed from the subscores representing each dimension. In the case of correlation 0.3, the pairwise correlations are 0.53, 0.48, and 0.49. These numbers are higher than the correlations that were intended to simulate if compared with those listed in Table 6. In Table 6, the correlation matrix of the 21 ability combinations at each

Table 4: Model-Fit Statistics for Bank I and Bank II

Model-Fit Statistics With Zero Correlation Among the Dimensions

	Bank I		Bank II	
	One-Factor	Three-Factor	One-Factor	Three-Factor
χ^2	1637.17	1636.90	3802.02	3352.29
df	3915	3912	3915	3912
p	1.00	1.00	0.90	1.00
AGFI	0.98	0.98	0.96	0.97
N	2000	2000	2000	2000

Model-Fit Statistics With Correlation 0.3 Among the Dimensions

	Bank I		Bank II	
	One-Factor	Three-Factor	One-Factor	Three-Factor
χ^2	6417.05	6414.42	2058.55	1897.56
df	3915	3912	3915	3912
p	0.00	0.00	1.00	1.00
AGFI	0.95	0.95	0.98	0.98
N	2000	2000	2000	2000

Model-Fit Statistics With Correlation 0.6 Among the Dimensions

	Bank I		Bank II	
	One-Factor	Three-Factor	One-Factor	Three-Factor
χ^2	2157.25	2155.61	1087.75	1039.95
df	3915	3912	3915	3912
p	1.00	1.00	1.00	1.00
AGFI	0.97	0.97	0.99	0.99
N	2000	2000	2000	2000

*AGFI = adjusted goodness of fit index.

intercorrelation level is presented. In addition, the sampled mean and standard deviation of the 21 ability combinations for each dimension at each intercorrelation level are listed. To get a clearer picture of the correlation, the Pearson product moment correlations were computed based on the subscores of each dimension from 300 items. Again the numbers are higher than the intended ones. However, the computed correlations based on the simulated response matrix consistently became higher as the intended correlations were getting higher. One possible reason contributed to this discrepancy is the random error occurred in the procedure of determination of correct or incorrect response. A random number between 0

and 1 is generated to be compared with the calculated probability. These random errors during the course of simulation are inevitable. These inflated correlation should be taken into account in interpreting the results of the study.

Table 5: Correlations Between Simulated Dimensions

Correlations Between Dimensions Based on Simulated Item Response with r=0.0

	90 Items			300 Items	
	Dimension1	Dimension2	Dimension3	Dimension1	Dimension2
D1		0.31	0.47		
D2	0.2785		0.28	0.4123	
D3	0.2722	0.2563		0.3930	0.3984

Correlations Between Dimensions Based on Simulated Item Response with r=0.3

	90 Items			300 Items	
	Dimension1	Dimension2	Dimension3	Dimension1	Dimension2
D1		0.7	0.68		
D2	0.5264		0.63	0.6543	
D3	0.4795	0.4873		0.6452	0.6531

Correlations Between Dimensions Based on Simulated Item Response with r=0.6

	90 Items			300 Items	
	Dimension1	Dimension2	Dimension3	Dimension1	Dimension2
D1		0.87	0.84		
D2	0.6786		0.83	0.8023	
D3	0.6265	0.6616		0.7855	0.8095

* The data are based on the analysis of Bank II. The upper half of the left matrix is obtained through LISREL and the lower half is based on the Pearson product moment correlations computed from the subscores representing each dimension.

Table 6: Correlation Matrix of the 21 ability combinations at each correlation level

Correlation	Thetas	θ_1	θ_2	θ_3	Mean	Std.
Cor 0.0	θ_1	1.000	0.049	-0.0454	-0.002	1.005
	θ_2		1.000	-0.0094	-0.001	1.002
	θ_3			1.000	-0.001	0.999
Cor 0.3	θ_1	1.000	0.2927	0.3046	0.001	1.001
	θ_2		1.000	0.2546	0.003	1.001
	θ_3			1.000	-0.003	0.981
Cor 0.6	θ_1	1.000	0.6232	0.5739	-0.004	0.991
	θ_2		1.000	0.6041	-0.005	1.015
	θ_3			1.000	-0.006	0.981

Results and Discussions

In this section, simulated data are analyzed and presented. In the first part, bias results are presented in terms of plots and ANOVAs. In the second part, tables of the average standard errors are listed based on each 21 ability combination at each correlation level. In the third part, plots of total RMSE for each ability combination are presented at each correlation level. Finally, the relationship between the actual SEs and the test information-based SEs are discussed for the two non-Bayesian methods.

Bias

Bias Plots for the Estimated and True Abilities

Figures 1 to 18 in Appendix B show the bias plots of the four estimation methods for the two item banks at each correlation level at a test length of 50 items. These plots show that MLE is the most consistent estimator among the four simulated methods. If we draw a line for both +0.1 and -0.1 on the bias scale for each condition under Bank II and a line for both +0.2 and -0.2 on the bias scale for each condition under Bank I, almost all the bias points of MLE fall into this region. These plots also indicate that the bias for MLE is either smaller than or comparable with that for the two Bayesian methods along the theta continuum, especially at extreme ability levels. In most of the conditions, the Bayesian methods show a bias toward the prior mean (0.0), especially evident by Figures 1, 7, and 15. The two Bayesian methods yield comparable estimates for all the conditions under Bank I, which is evident by the two almost identical solid lines in the bias plots.

By comparing the bias plots of the estimations of each theta scale at different correlation levels under both item banks, it is obvious that the magnitude of bias under Bank II is only half of the size of that under Bank I. This indicates that item banks with items which consists of only one primary factor and some secondary factors are more desirable in reducing estimation bias for a multidimensional CAT system that we are investigating. The estimation of WLE shows the greatest discrepancies between the two banks. This is especially evident by Figures 4, 10, and 16. The greatest differences all occur at the high end of the scales,

where the WLE bias is as large as -0.6. Notice that the two Bayesian estimates also yield great bias at this end of ability scale. When Bank II is used, WLE is as consistent as MLE at all correlation levels.

A comparison of bias plot for the four methods under different correlation levels indicates similar pattern of bias resulted by each method across the three correlation levels. In addition, the magnitude of bias of each of the four estimation methods is comparable across the three correlation levels. For the two Bayesian, the direction of bias is inward (i.e., toward the middle of the scale) at the low and high extremes of the ability scale for all three correlation levels. There are exceptions in Figures 6 and 14 at correlation 0.0 and 0.6, where the ability is overestimated between 0.0 and 1.0 on the true theta scale. Compared with the two Bayesian, WLE and MLE are considered to be more consistent estimator along the entire ability scale across all three correlation levels. There are a few outliers for MLE estimates under bank II, seen in Figure 13, and for WLE under Bank I, as seen in Figures 4, 10, and 16. These large biases may reflect the effect of the factorial structure of the items in the two banks. The results from these bias plots fail to show any effect of different correlation levels on the ability estimation for the four methods.

Taken collectively, the results for bias show that Bayesian ability estimates at high and low levels are pulled toward the prior mean and that this bias is further exaggerated when Bank I is used, i.e. under an item bank with items of equally dominant on all dimensions. The two Bayesian display comparable bias across the three correlation levels, especially when Bank I is used for the simulation. WLE and MLE ability estimates are more consistent and less biased compared with the Bayesian along the true theta continuum. However, when Bank I is used, WLE shows a few large biased estimates at the high end of the ability scale. This result may reflect the combined effects of different item structure and the weighting function of WLE estimation. Finally, the bias plots does not indicate any effect of intercorrelation between dimensions on the ability estimation for any of the four methods.

ANOVA Results for the Bias

Biases for the four estimation methods are analyzed at each correlation level for each item bank. A $4 \times 21 \times 3$ ANOVA was conducted for each intercorrelation level separately since the 21 ability combinations at each correlation level are unable to be matched properly due to the complexity of three dimensions. At each intercorrelation level, we have four different estimation methods, 21 ability combinations, and three dimensions of thetas that are estimated. Even though the differences among the three theta estimates were analyzed, we do not expect any differences among the three theta estimates since each of them was generated independently from the same specifications.

Table 7 shows the results of the three-way ANOVA of bias for Bank I. Using $\alpha = 0.001$ for each hypothesis tested: two main effects, methods(M) and ability combination (AC), and three interaction effects, $M \times AC$, $AC \times T$, and $M \times AC \times T$, were statistically significant for all correlation levels.

Because the total number of replications is large, effect sizes were used to provide additional information on significant effects. The magnitude of significant effects was estimated using the eta-squared η^2. Cohen(1988) provided some advice on classification of the effect size in terms of η^2: (a) no effect ($\eta^2 < 0.0099 \approx 0.01$), (b) small effect ($0.01 < \eta^2 < 0.0588 \approx 0.06$), (c) medium effect ($0.06 < \eta^2 < 0.1379 \approx 0.14$), and (d) large effect ($\eta^2 > 0.14$). In terms of η^2, the two-way $AC \times T$ interaction effect and the three-way $M \times AC \times T$ interaction effect accounted for most of the variance for all correlation levels. More specifically, 4.84% and 4.40% of the total sum of squares of the bias for the ability estimate was due to the interaction effects of $AC \times T$ and $M \times AC \times T$ at correlation 0.0. Although these interaction effect are statistically significant, their effect sizes belong to the small range ($0.01 < \eta^2 < 0.0588 \approx 0.06$). Among the main effects, even though methods(M) and ability combination (AC) effects are statistically significant, their effect sizes are classified into no effect level according to Cohen (1988), i.e. ($\eta^2 < 0.0099 \approx 0.01$).

For the two-way interaction, the significance means that mean differences among the

levels of any factor are not constant across all levels of the remaining factor. To further explore the nature of interactions, two sets of cell means are given in Tables 8 and 9. The significance of differences among the means of the main effect of M (methods) are also given in Table 8. The only significant differences at .001 level are WLE and other three methods. We did not compute the mean differences of the ability combinations because there are 21 levels and their differences can be examined through the bias plots presented before. The main effect of T (thetas) was not included due to nonsignificance.

In terms of the ability estimation methods, if we take the absolute value of the cell means, WLE has the largest mean of bias among all ability estimation methods, and the mean of WLE is significantly different all three other methods at all three correlation levels. Overall, WLE seems to underestimate the abilities, especially at the high end of the ability scale, which is evident by the bias plots presented in the previous section. MLE yields the second largest mean bias, but the mean of MLE is not significantly different from the Bayesian methods at all three correlation levels. The two Bayesian methods again yield very comparable results in terms of the bias at all three correlation levels.

Since the complexity of the 21 ability combinations, it is not clear how the methods and different ability combinations interact with one another. To get a clearer picture of the interaction between estimation methods and levels of ability, the 21 ability combinations were collapsed into 5 different levels along the ability scale: (1) $\theta < -1.5$, (2) $-1.5 < \theta < -.5$, (3) $-.5 < \theta < .5$, (4) $0.5 < \theta < 1.5$, and (5) $\theta > 1.5$. Table 9 shows the cell means for the 4×5 factorial design. The data indicate very large underestimates occurred at the high end of the ability for all four methods; and very large overestimates were yielded at the low end of the ability scale for all four methods. The data clearly show that all four methods yield large bias under Bank I, especially at the extremes of the ability scale.

The results from ANOVA on bias for Bank I simulation output indicate that WLE reduced too much of the bias of MLE. However, the following results from Bank II show a very different outcome. Table 10 shows the results of the three-way ANOVA of bias for Bank II.

Using $\alpha = 0.001$ for each hypothesis tested: two interaction effects, $AC \times T$ and $M \times AC \times T$, were statistically significant for correlation equal zero; one main effect, ability combination (AC), and two interaction effects, $AC \times T$ and $M \times AC \times T$, were statistically significant for correlation equal 0.3; one main effect, ability combination (AC), and two interaction effects, $AC \times T$ and $M \times AC \times T$, were statistically significant for correlation equal 0.6.

For Bank II, the two-way $AC \times T$ interaction effect and the three-way $M \times AC \times T$ interaction effect also accounted for most of the variance for all correlation level. More specifically, 2.43% and 3.26% of the total sum of squares of the bias for the ability estimate was due to the interaction effects of $AC \times T$ and $M \times AC \times T$, respectively, at correlation 0.0. Even though this interaction effect is statistically significant, its effect size belong to small ranges ($0.01 < \eta^2 < 0.0588 \approx 0.06$). The main effect of ability combination (AC) is statistically significant. Its effect size is classified into no effect level according to Cohen (1988), i.e. ($\eta^2 < 0.0099 \approx 0.01$).

For Bank II, there is only one significant main effect, i.e. the ability combination. The two-way interaction, $AC \times T$, and the three-way interaction, $M \times AC \times T$, are significant, which means the mean differences among the levels of any factor are not constant across all levels of the remaining factor. The main effect of ability combination can be examined through the bias plots presented in the previous section. To further explore the nature of interactions, two sets of cell means are provided in Tables 11 and 12.

Although the main effect of methods is not statistically significant, the cell means and the mean differences of multiple comparison are still presented to compare the four estimation methods. Table 11 shows the cell means and the mean differences of bias at each correlation level. For Bank II, if we take the absolute value of the cell means, WLE has the smallest mean of bias among all ability estimation methods, followed by EAP, MAP, and WLE, at correlation 0.3 level. For correlation equal 0.0 and 0.6, WLE yields the second smallest bias and the ability estimates of WLE is slightly underestimated. It is clear that WLE reduces the MLE bias when Bank II is used. Table 12 shows the cell means for the 4×5 factorial

design. It is clear that under Bank II, both WLE and MLE yield smaller bias at the extreme ends of the ability scale. The two Bayesian methods again resulted in underestimation at the high end of the ability distribution and in overestimation at the low end of ability scale.

The above results indicate that the two non-Bayesian do not yield relatively small bias as expected in comparison with the two Bayesian methods under Bank I and that under Bank II WLE yields the smallest bias at correlation 0.3, as well as reducing the MLE bias at all three correlation levels. This result can probably be attributed to short test length for the simulations, the sampling of the ability combinations, and the multidimensional structure of the item banks. As it has been mentioned in van der Linden (1997), MLE was strongly biased in the MCAT situations when the test length is not long enough. He has suggested an alternative method, i.e. the maximum modal method as used in Segall (1996), since by imposing a prior distribution to the likelihood function, the MAP estimator can be stabilize faster as a function of the test length. In our study the test length for each dimension is about 16 for Bank II. This test length seems not long enough for MLE to yield more accurate ability estimates than the other three methods. WLE shows the greatest discrepancies of the biases between the two item banks.

Table 7: Results of ANOVA of Bias for Each Correlation Level under Bank I

		Correlation zero		Correlation 0.3		Correlation 0.6	
Source	df	F	η^2	F	η^2	F	η^2
Main effect							
Method (M)	3	14.17*	0.0031	11.51*	0.0026	28.53*	0.0062
Ability (AC)	20	1.87*	0.0028	2.03*	0.0029	4.19*	0.0062
Combination							
Thetas (T)	2	0.57	0.0000	3.79	0.0005	3.06	0.0005
Interaction							
$M \times AC$	60	1.81*	0.0078	1.05*	0.0047	3.08*	0.0135
$M \times T$	6	1.47	0.0005	4.12	0.0017	2.43	0.0011
$AC \times T$	40	16.77*	0.0484	14.64*	0.0428	11.97*	0.0315
$M \times AC \times T$	120	5.09*	0.0440	4.64*	0.0407	2.77*	0.0243

* The F values are significant at .001 level.

Table 8: Cell Means and Mean Differences of Multiple Comparison of Bias at Each Inter-correlation Level For Bank I

		WLE	MLE	EAP	MAP
$cor = 0.0$					
	Cell Means	-0.0666	0.0030	-0.0007	0.0027
Mean Difference	WLE	—-	−0.0696*	−0.0659*	−0.0639*
	MLE		—-	0.0037	0.0003
	EAP			—-	-0.0034
	MAP				
$cor = 0.3$					
	Cell Means	-0.0555	0.0087	0.0013	0.0061
Mean Difference	WLE	—-	−0.0642*	−0.0568*	−0.0616*
	MLE		—-	0.0074	0.0026
	EAP			—-	-0.0048
	MAP				
$cor = 0.6$					
	Cell Means	-0.0924	0.0109	0.0038	0.0059
Mean Difference	WLE	—-	−0.1033*	−0.0962*	−0.0983*
	MLE		—-	0.0071	0.0050
	EAP			—-	-0.0021
	MAP				

* The mean differences is significant at the 0.001 level based on the Tukey test.

Table 9: Cell Means of Bias by Methods and Ability Levels at Intercorrelation Level 0.0 under Bank I

			BIAS	
Methods	Ability Levels	N	Mean	SD
WLE	$\theta < -1.5$	300	1.31205000	1.25408609
	$-1.5 < \theta < -.5$	600	0.71716667	1.04799834
	$-.5 < \theta < .5$	1200	-0.15570000	1.29883864
	$0.5 < \theta < 1.5$	900	-0.48993333	0.75713359
	$\theta > 1.5$	150	-1.15643333	1.08766231
MLE	$\theta < -1.5$	300	1.26838333	1.20491766
	$-1.5 < \theta < -.5$	600	0.72866667	0.85428161
	$-.5 < \theta < .5$	1200	-0.14403333	1.13916327
	$0.5 < \theta < 1.5$	900	-0.50337778	0.65241989
	$\theta > 1.5$	150	-1.23310000	1.05208394
EAP	$\theta < -1.5$	300	1.28414633	1.19615583
	$-1.5 < \theta < -.5$	600	0.70899933	0.83941359
	$-.5 < \theta < .5$	1200	-0.14975433	1.12669075
	$0.5 < \theta < 1.5$	900	-0.50076467	0.62385061
	$\theta > 1.5$	150	-1.20138933	1.02583150
MAP	$\theta < -1.5$	300	1.14205000	1.29518254
	$-1.5 < \theta < -.5$	600	0.67850000	1.11536092
	$-.5 < \theta < .5$	1200	-0.28195000	1.32523522
	$0.5 < \theta < 1.5$	900	-0.57893333	0.88622374
	$\theta > 1.5$	150	-1.28910000	1.25774917

Table 10: Results of ANOVA of Bias for Each Correlation Level under Bank II

		Correlation zero		Correlation 0.3		Correlation 0.6	
Source	df	F	η^2	F	η^2	F	η^2
Main effect							
Method (M)	3	1.26	0.0003	1.08	0.0002	2.27	0.0005
Ability (AC)	20	2.07	0.0033	3.83*	0.0058	4.40*	0.0068
Combination							
Thetas (T)	2	5.61	0.0008	0.07	0.0000	1.99	0.0003
Interaction							
$M \times AC$	60	0.79	0.0037	0.72	0.0033	1.27	0.0059
$M \times T$	6	1.17	0.0006	0.82	0.0004	0.54	0.0002
$AC \times T$	40	7.80*	0.0243	8.49*	0.0257	4.80*	0.0145
$M \times AC \times T$	120	3.49*	0.0326	3.15*	0.0280	2.01*	0.0180

* The F values are significant at .001 level.

Table 11: Cell Means and Mean Differences of Multiple Comparison of Bias at Each Inter-correlation Level For Bank II

		WLE	MLE	EAP	MAP
$cor = 0.0$					
	Cell Means	-0.0060	0.0086	-0.0037	0.0010
Mean Difference	WLE	—-	-0.0146	-0.0023	-0.0050
	MLE		—-	0.0123	0.0076
	EAP			—-	-0.0047
	MAP				
$cor = 0.3$					
	Cell Means	0.0028	0.0103	-0.0037	0.0058
Mean Difference	WLE	—-	-0.0075	0.0065	-0.0030
	MLE		—-	0.0145	0.0045
	EAP			—-	-0.0095
	MAP				
$cor = 0.6$					
	Cell Means	-0.0062	0.0119	-0.0010	0.0109
Mean Difference	WLE	—-	-0.0181	-0.0052	-0.0171
	MLE		—-	0.0129	0.0010
	EAP			—-	-0.0119
	MAP				

* The mean differences is significant at the 0.001 level based on the Tukey test.

Table 12: Cell Means of Bias by Methods and Ability Levels at Intercorrelation Level 0.0 under Bank II

			BIAS	
Methods	Ability Levels	N	Mean	SD
WLE	$\theta < -1.5$	150	-0.02793333	0.30605069
	$-1.5 < \theta < -.5$	1000	0.00239000	0.34418375
	$-.5 < \theta < .5$	850	-0.02237647	0.34689094
	$0.5 < \theta < 1.5$	950	-0.00346316	0.43349541
	$\theta > 1.5$	200	0.02630000	0.34719642
MLE	$\theta < -1.5$	150	-0.00526667	0.34730029
	$-1.5 < \theta < -.5$	1000	-0.00181000	0.33733618
	$-.5 < \theta < .5$	850	0.00550588	0.35725667
	$0.5 < \theta < 1.5$	950	0.01580000	0.31775416
	$\theta > 1.5$	200	0.05030000	0.36319555
EAP	$\theta < -1.5$	150	0.20959333	0.32463789
	$-1.5 < \theta < -.5$	1000	0.07155490	0.30328656
	$-.5 < \theta < .5$	850	-0.00915329	0.26828963
	$0.5 < \theta < 1.5$	950	-0.06577032	0.28124513
	$\theta > 1.5$	200	-0.22189100	0.25853068
MAP	$\theta < -1.5$	150	0.16873333	0.32821263
	$-1.5 < \theta < -.5$	1000	0.08679000	0.29686911
	$-.5 < \theta < .5$	850	-0.00308235	0.28905422
	$0.5 < \theta < 1.5$	950	-0.07430526	0.28031471
	$\theta > 1.5$	200	-0.17820000	0.27713575

Standard Errors

From Table 13 to Table 18, the average standard errors of estimates are listed for the four different methods at different correlation levels under each item bank. The average SE stands for the average of three SE's from three theta's for each ability combination. These standard errors are based on test length of 40 and 50 items, respectively. Since there are three dimensions estimated, about 15 items was administered per dimension. The small number of test items administered to each dimension should be considered when the results of SEs are interpreted. These tables clearly show that EAP yields consistently the smallest average SEs and MAP the second smallest for most of the ability combinations at all three correlation levels under both item banks. WLE yields comparable SEs as MLE under Bank II, but yields much larger SEs than the other three methods when Bank I was employed in the simulation. The tables also reveal that the average SEs after 50 items administered are smaller than the average SEs after 40 items administered in most of the ability combinations for all four methods at each correlation level under both item banks and the differences are most pronounced for MLE and WLE. Longer tests are resulting smaller average SEs especially under Bank II. Table 14 and Table 18 indicate that MLE yields consistently smaller average SEs for all ability combinations when 50 items are administered using Bank II. Even when Bank I is used, there are only a few exceptions when the shorter tests result in smaller average SEs for MLE under different correlation level.

Tables 13, 15, and 17 are presenting the average SEs for the four methods at the three different correlation levels under Bank I. The data indicate that the same estimation method yields similar average SEs at all three correlation levels. The correlation between dimensions did not help to increase estimation precision for all four methods. Tables 14, 16, and 18 show the average SEs for the four methods at each correlation level when Bank II is employed in the simulation. Again, the resulted data did not support higher intercorrelation between dimensions would increase estimation precision. However, the simulated data from both item banks clearly show that different multidimensional structures, equally dominant dimensions of Bank I and approximate simple structure of Bank II, might have an effect on estimation precision. Smaller SEs were obtained for all four estimation methods at the same correlation level when Bank II was used. The differences were especially pronounced for WLE. Since the two item banks are different only on the discriminating parameters, one speculated reason for the degraded performance of WLE is that the discriminating parameters in Bank I cause distorted estimation of the weighting function during the exhausting search procedure. The discriminating parameters were generated to be equally dominant in all three dimensions for Bank I. On the other hand, the discriminating parameters of each item in Bank II consist of one larger parameter for the dominant dimension and two smaller parameters for the other two secondary dimensions.

To summarize, the findings for SE indicate that random measurement error is more prevalent with the non-Bayesian ability estimates than with the Bayesian estimates at all three correlation levels under both item banks. The average SEs are most pronounced with WLE ability estimates when Bank I is used, in which the three dimensions are generated to be equally dominant. Longer tests usually reduces the average SEs especially for MLE ability estimates. For those exceptions, one possible reason is that the item pool does not have sufficient number of items for those particular ability combinations. Finally, the data fail to support that the correlation between dimensions has an effect on the precision of ability estimates.

Table 13: Average standard errors of estimates for WLE, MLE, EAP, and MAP ($r = 0.0$) under Bank I

Method of Estimation	WLE		MLE		EAP		MAP	
Termination Criterion	40	50	40	50	40	50	40	50
Ability Combination								
(-1.0608, -.6022, -.0613)	.7112	.6114	.5440	.4473	.3651	.3516	.3731	.3834
(-.1605, -.2593, -2.2625)	.5563	.5125	.6420	.5131	.3815	.3664	.3980	.3464
(-.6665, 1.5407, .3809)	.6519	.6576	.5499	.4708	.3575	.3673	.3328	.3136
(-1.2094, -.6288, -1.2545)	.5555	.4517	.6477	.5416	.3518	.3249	.3564	.3442
(.8678, 1.2602, -.6045)	.8710	.8060	.5803	.5885	.3996	.3870	.3132	.3437
(.3685, -1.3398, 2.0521)	.6502	.7309	.4745	.4717	.3485	.3429	.3072	.3236
(.2050, .2415, 1.3271)	.6940	.8789	.4892	.4774	.3353	.3408	.3912	.3867
(1.7038, -.6830, -1.1732)	.8175	.6114	.4650	.4509	.4437	.4257	.4056	.4055
(-1.7767, 1.7482, -.3762)	.6076	.5291	.6061	.5453	.3804	.3359	.4484	.4267
(-.7355, -1.3467, -.2333)	.6269	.4932	.6678	.6138	.3448	.3418	.3378	.3307
(.5683, -.7332, -.2076)	.6526	.6462	.5488	.5380	.3393	.3774	.4306	.4351
(-1.6270, -.8455, .6147)	.5895	.5303	.5925	.5524	.3802	.3915	.3677	.3614
(-.3926, -1.0253, -.5998)	.5450	.5125	.6623	.5391	.3930	.3545	.4029	.3912
(.8313, .8387, .5944)	.6997	.8715	.5412	.4862	.3635	.3437	.3380	.3579
(.7481, .3899, -1.0694)	.6290	.7048	.6084	.5695	.3691	.3844	.3845	.3841
(-1.2381, .8684, 1.0415)	.4885	.6721	.5298	.5151	.3454	.3703	.3474	.3601
(-.2178, -.8186, .7637)	.5893	.5375	.6332	.5472	.4070	.4009	.4381	.4278
(1.0263, -.1641, -.4467)	.5652	.5997	.5450	.4921	.3723	.3650	.3204	.2925
(.8255, .8322, .4964)	.8293	.8294	.5787	.4952	.3499	.3607	.3797	.3687
(.6907, -.7810, .9007)	.7510	.6607	.7006	.6047	.3142	.2952	.3764	.3487
(1.1956, 1.4806, .1036)	1.0482	1.2620	.5527	.5050	.3442	.3555	.3637	.3865
Mean	.6728	.6719	.5790	.5221	.3660	.3611	.3720	.3675
SD	.1318	.1846	.0647	.0480	.0290	.0285	.0401	.0380

* For each ability combination, 50 replications were made. The average SE stands for the average of three SE's from three theta's for each ability combination.

Table 14: Average standard errors of estimates for WLE, MLE, EAP, and MAP ($r = 0.0$) under Bank II

Method of Estimation	WLE		MLE		EAP		MAP	
Termination Criterion	40	50	40	50	40	50	40	50
Ability Combination								
(-1.0608, -0.6022, -0.0613)	.4092	.3254	.3709	.3154	.2733	.2669	.2734	.2737
(-0.1605, -0.2593, -2.2625)	.3365	.3243	.3684	.3177	.2827	.2533	.3736	.3229
(-0.6665, 1.5407, 0.3809)	.3899	.3980	.4174	.3813	.3325	.2886	.2992	.2906
(-1.2094, -0.6288, -1.2545)	.3865	.3548	.3636	.3517	.3401	.3291	.3428	.3162
(0.8678, 1.2602, -0.6045)	.3881	.3540	.3692	.3565	.2971	.2773	.2750	.2572
(0.3685, -1.3398, 2.0521)	.3440	.3121	.3330	.3077	.2920	.2771	.3172	.3113
(0.2050, 0.2415, 1.3271)	.3576	.3124	.4104	.3286	.2696	.2537	.2717	.2620
(1.7038, -0.6830, -1.1732)	.3415	.3077	.3546	.3392	.2978	.2853	.2890	.2604
(-1.7767, 1.7482, -0.3762)	.5419	.3306	.5061	.4347	.3598	.3069	.3855	.3266
(-0.7355, -1.3467, -0.2333)	.3314	.3328	.3699	.3497	.2735	.2681	.3290	.2926
(0.5683, -0.7332, -0.2076)	.3433	.3912	.3528	.3199	.2524	.2655	.3032	.2882
(-1.6270, -0.8455, 0.6147)	.3786	.3999	.4683	.3085	.3379	.3238	.2883	.2846
(-0.3926, -1.0253, -0.5998)	.3523	.3418	.3639	.3474	.2793	.2804	.3140	.3180
(0.8313, 0.8387, 0.5944)	.4282	.5659	.3760	.2867	.3177	.2918	.3169	.3074
(0.7481, 0.3899, -1.0694)	.3960	.5309	.3454	.3246	.3129	.2750	.2994	.2993
(-1.2381, 0.8684, 1.0415)	.3884	.3565	.3792	.3375	.3161	.2865	.2690	.2620
(-0.2178, -0.8186, 0.7637)	.3343	.3226	.3442	.3203	.2886	.2627	.2961	.2728
(1.0263, -0.1641, -0.4467)	.4525	.2675	.4131	.3206	.3067	.2912	.3038	.2691
(0.8255, 0.8322, 0.4964)	.4952	.3309	.3752	.3348	.2710	.2688	.2310	.2304
(0.6907, -0.7810, 0.9007)	.3901	.3135	.3374	.3235	.3043	.2765	.2844	.2593
(1.1956, 1.4806, 0.1036)	.3831	.3800	.3481	.3047	.2618	.2440	.2998	.2827
Mean	.3890	.3597	.3794	.3339	.2984	.2796	.3030	.2851
SD	.0541	.0710	.0430	.0312	.0285	.0214	.0349	.0258

* For each ability combination, 50 replications were made. The average SE stands for the average of three SE's from three theta's for each ability combination.

Table 15: Average standard errors of estimates for WLE, MLE, EAP, and MAP ($r = 0.3$) under Bank I

Method of Estimation	WLE		MLE		EAP		MAP	
Termination Criterion	40	50	40	50	40	50	40	50
Ability Combination								
(-.3625, .6203, .6061)	.7256	.5830	.4726	.4378	.3632	.3729	.3281	.3223
(.1090, .8991, .1424)	.7312	.7330	.5867	.5055	.3541	.3583	.3189	.3304
(-.5980, .3748, -1.2350)	.5047	.6062	.5439	.5324	.3502	.3364	.4106	.3741
(-.4047, -.9719, .1587)	.7112	.6994	.5513	.5459	.3947	.4005	.4042	.4289
(-.2161, .7989, .3751)	.6879	.5286	.5252	.4636	.3909	.3622	.3340	.3255
(1.5361, .7659, -.6542)	.9774	.9415	.6402	.4687	.3532	.3509	.3740	.3689
(.7892, -.1639, -1.8583)	.6401	.6276	.6281	.5978	.3938	.3952	.3611	.3727
(.3090, .5018, .6207)	.6917	.5872	.4715	.4278	.3569	.3223	.3448	.3568
(-1.1166, -.6423, -2.2183)	.5451	.4828	.4546	.4617	.3227	.3327	.3674	.3728
(2.0217, .4583, 1.7097)	1.1013	.9821	.5252	.5382	.3350	.3291	.4056	.3906
(-.5561, .6430, .3503)	.7019	.5177	.5697	.5605	.3382	.3401	.3717	.3613
(-1.5828, -.1699, -.6448)	.5835	.5002	.4543	.4726	.4081	.3972	.3829	.3727
(-1.9975, -1.9115, .3195)	.6111	.5366	.5771	.5206	.3548	.3477	.3364	.3052
(.1941, -.7866, -.7366)	.5998	.6205	.5900	.5905	.3750	.3623	.4290	.4515
(1.4661, -.1238, .6985)	.8463	.8777	.5969	.4929	.3547	.3418	.4025	.3774
(.0944, -.3065, 1.4019)	.6700	.5719	.5859	.5678	.3106	.3138	.3265	.3135
(.2711, -1.6905, .1297)	.4974	.5696	.6013	.5415	.3782	.3799	.3954	.4037
(-.3313, -.9797, -.1959)	.5342	.6496	.4915	.4943	.3670	.3552	.3528	.3677
(1.0842, -.0607, .2823)	.7916	.6014	.6147	.5212	.3417	.3444	.3260	.3193
(-.7425, .1764, -.3848)	.7717	.6696	.6105	.5480	.3987	.3764	.4298	.4132
(.0436, 2.6403, 1.0670)	1.0985	1.0631	.4264	.4088	.3432	.3253	.3924	.3514
Mean	.7153	.6643	.5485	.5094	.3612	.3545	.3711	.3657
SD	.1725	.1651	.0643	.0528	.0260	.0252	.0354	.0385

* For each ability combination, 50 replications were made. The average SE stands for the average of three SE's from three theta's for each ability combination.

Table 16: Average standard errors of estimates for WLE, MLE, EAP, and MAP ($r = 0.3$) under Bank II

Method of Estimation	WLE		MLE		EAP		MAP	
Termination Criterion	40	50	40	50	40	50	40	50
Ability Combination								
(-.3625, .6203, .6061)	.3700	.3851	.4087	.3377	.3350	.3088	.2771	.2701
(.1090, .8991, .1424)	.3995	.3268	.4012	.3032	.2928	.2823	.2501	.2416
(-.5980, .3748, -1.2350)	.3207	.3447	.3783	.3627	.2875	.2938	.3023	.2860
(-.4047, -.9719, .1587)	.4213	.3895	.4186	.3564	.2987	.2656	.2859	.2701
(-.2161, .7989, .3751)	.3638	.3239	.3473	.3600	.2957	.2980	.3300	.3274
(1.5361, .7659, -.6542)	.3862	.3288	.3549	.3157	.2674	.2654	.2817	.2703
(.7892, -.1639, -1.8583)	.3230	.3092	.3737	.3438	.3046	.2850	.3279	.2963
(.3090, .5018, .6207)	.3507	.3152	.3925	.3094	.2632	.2459	.2928	.2589
(-1.1166, -.6423, -2.2183)	.3708	.3878	.5563	.4063	.3207	.2819	.3127	.2725
(2.0217, .4583, 1.7097)	.6563	.3387	.3824	.3462	.2691	.2641	.2920	.2844
(-.5561, .6430, .3503)	.3327	.4180	.3815	.3101	.3089	.2984	.2931	.2739
(-1.5828, -.1699, -.6448)	.3464	.3246	.3313	.3027	.2347	.2482	.2848	.2714
(-1.9975, -1.9115, .3195)	.3910	.3677	.3800	.3434	.3051	.2774	.3608	.3054
(.1941, -.7866, -.7366)	.3648	.3748	.5409	.3921	.3184	.3107	.3181	.3090
(1.4661, -.1238, .6985)	.5937	.3372	.3445	.3470	.3298	.3043	.3400	.3185
(.0944, -.3065, 1.4019)	.3532	.3031	.3021	.2870	.3070	.2989	.3186	.2933
(.2711, -1.6905, .1297)	.3438	.3117	.3451	.3134	.3076	.2837	.3242	.2866
(-.3313, -.9797, -.1959)	.3626	.3252	.3462	.3054	.2455	.2524	.2724	.2877
(1.0842, -.0607, .2823)	.3875	.3310	.4207	.3319	.2986	.2735	.2682	.2766
(-.7425, .1764, -.3848)	.4387	.3333	.4738	.3462	.2993	.2779	.2975	.2834
(.0436, 2.6403, 1.0670)	.3569	.3494	.3415	.3328	.2619	.2640	.2836	.2796
Mean	.3921	.3441	.3915	.3359	.2929	.2800	.3007	.2839
SD	.0834	.0312	.0644	.0301	.0268	.0194	.0268	.0199

* For each ability combination, 50 replications were made. The average SE stands for the average of three SE's from three theta's for each ability combination.

Table 17: Average standard errors of estimates for WLE, MLE, EAP, and MAP ($r = 0.6$) under Bank I

Method of Estimation	WLE		MLE		EAP		MAP	
Termination Criterion	40	50	40	50	40	50	40	50
Ability Combination								
(.7456, .5141, .2030)	.9378	.8154	.5561	.5019	.3382	.3136	.3368	.3607
(-.2206, .5439, .5845)	.7071	.5781	.5987	.5152	.3605	.3571	.3661	.3533
(-.1898, .4468, -1.1742)	.6178	.6051	.5604	.5819	.3754	.3885	.3863	.4006
(.9927, .4856, .6723)	.8213	.9409	.5940	.4853	.3453	.3338	.3966	.3794
(-.0339, -.9653, -1.2353)	.6794	.6350	.6751	.5398	.3508	.3571	.3765	.3537
(-.5380, -1.0996, -.6165)	.6793	.5708	.6027	.4950	.3584	.3570	.4265	.4030
(-1.5095, -1.4189, -2.0371)	.4929	.4634	.5718	.4946	.3752	.3929	.3306	.3316
(.5468, .2746, -.2876)	.6079	.6843	.5215	.4590	.3805	.3543	.3834	.3844
(.6452, 1.1306, .1933)	1.0925	1.2586	.4808	.4762	.4328	.3990	.3571	.3348
(1.5993, .9850, 1.2278)	.6752	.8973	.6020	.5814	.3520	.3444	.3665	.3620
(1.2558, -.4750, .7035)	1.0865	.7018	.5485	.4824	.4209	.4243	.3457	.3190
(-1.2861, -.5675, 1.1683)	.5961	.5835	.6751	.5976	.4648	.4567	.4186	.3897
(-.8293, -.8132, -.4131)	.5752	.4820	.7124	.6072	.3493	.3276	.3954	.3420
(.5755, -.6714, -.0504)	.6617	.5337	.6327	.5140	.4287	.3924	.3880	.3703
(.0120, -.0264, -.6960)	.5384	.6165	.5248	.5114	.3533	.3441	.3919	.4114
(.7440, 1.0081, 1.1386)	.8274	.6212	.5525	.5628	.3222	.3244	.3820	.3733
(.1961, 1.4592, 1.1999)	1.0314	1.1857	.5248	.5036	.3497	.3269	.4088	.3872
(-1.0055, .8955, -.9847)	.6000	.4844	.5622	.4958	.3268	.3195	.3149	.3282
(-2.4622, -2.3222, -1.0200)	.3779	.4182	.4905	.4228	.3197	.3151	.3638	.3499
(.4408, 1.2551, 1.4407)	.8795	1.1681	.5653	.4832	.3257	.3359	.3519	.3536
(.2422, -.7353, -.1502)	.5668	.5153	.5741	.5123	.3416	.3456	.3868	.3921
Mean	.7168	.7028	.5774	.5154	.3653	.3576	.3750	.3657
SD	.1965	.2490	.0594	.0473	.0399	.0381	.0287	.0265

* For each ability combination, 50 replications were made. The average SE stands for the average of three SE's from three theta's for each ability combination.

Table 18: Average standard errors of estimates for WLE, MLE, EAP, and MAP ($r = 0.6$) for Bank II

Method of Estimation	WLE		MLE		EAP		MAP	
Termination Criterion	40	50	40	50	40	50	40	50
Ability Combination								
(.7456, .5141, .2030)	.4007	.5467	.3806	.3461	.3046	.2923	.2755	.2650
(-.2206, .5439, .5845)	.3720	.3606	.3582	.2819	.2753	.2743	.2733	.2697
(-.1898, .4468, -1.1742)	.3640	.3457	.3571	.3441	.3029	.3129	.3236	.3018
(.9927, .4856, .6723)	.3643	.3334	.3948	.3323	.2968	.2745	.2980	.2623
(-.0339, -.9653, -1.2353)	.4108	.3827	.4087	.3564	.2986	.2940	.3055	.2996
(-.5380, -1.0996, -.6165)	.3775	.3334	.3919	.3705	.2979	.2760	.2932	.2973
(-1.5095, -1.4189, -2.0371)	.3373	.3061	.4264	.3349	.3133	.2825	.2973	.2815
(.5468, .2746, -.2876)	.4392	.3822	.4351	.4062	.2711	.2634	.3338	.3023
(.6452, 1.1306, .1933)	.3867	.4064	.3575	.3185	.2955	.2912	.2929	.2946
(1.5993, .9850, 1.2278)	.4200	.4526	.3904	.3818	.3064	.3056	.2842	.2692
(1.2558, -.4750, .7035)	.5863	.4866	.3510	.3191	.2652	.2470	.2825	.2765
(-1.2861, -.5675, 1.1683)	.3681	.3514	.5275	.3899	.3065	.2780	.2829	.2580
(-.8293, -.8132, -.4131)	.3880	.3548	.3363	.3069	.3214	.3070	.2661	.2859
(.5755, -.6714, -.0504)	.4150	.3637	.3846	.3329	.2677	.2748	.3056	.2799
(.0120, -.0264, -.6960)	.3706	.3322	.3350	.3240	.2594	.2452	.3228	.3072
(.7440, 1.0081, 1.1386)	.4008	.4820	.4236	.3706	.2929	.3039	.2905	.2819
(.1961, 1.4592, 1.1999)	.6387	.4498	.3599	.3190	.2697	.2661	.3016	.2757
(-1.0055, .8955, -.9847)	.3583	.3371	.3467	.3414	.3243	.2864	.3030	.2867
(-2.4622, -2.3222, -1.0200)	.3557	.3111	.4679	.3311	.2904	.2800	.4175	.3544
(.4408, 1.2551, 1.4407)	.7018	.4044	.3528	.3232	.3173	.2995	.2748	.2687
(.2422, -.7353, -.1502)	.4043	.3755	.3781	.3171	.3023	.2919	.3115	.2978
Mean	.4219	.3856	.3888	.3404	.2943	.2832	.3017	.2865
SD	.0971	.0641	.0473	.0299	.0192	.0184	.0319	.0213

* For each ability combination, 50 MCAT simulations were made. The average SE stands for the average of three SE's from three theta's for each ability combination.

RMSE

RMSE is a function of both SE and bias. Consequently, the results for RMSE are related to those already discussed for SE and bias. RMSE results are given in Appendix C. Figures 19, 20, and 21 show the total RMSE for each ability combination at each correlation level under Bank I at a test length of 50 items. Figures 22, 23, and 24 show the total RMSE for each ability combination at each correlation level under Bank II at a test length of 50 items. The total RMSE stands for the summation of RMSE for each dimension at each ability combination. Figures 19, 20, and 21 indicate clearly that WLE tends to have the largest total RMSEs, followed by MLE, MAP, and EAP. The differences between the two Bayesian methods are negligible and nonsignificant based on Bonferroni tests of differences between means for all main effects. When Bank I was used, based on Bonferroni grouping, the four methods are divided into three groups, WLE in the highest, MLE in the second, and MAP and EAP in the third groups. The same pattern of grouping occurs at each correlation level. As mentioned in the bias section, the two Bayesian methods yield comparable estimation bias across the 21 ability combinations at each correlation level when Bank I was used. These plots also indicate that the two Bayesian methods yield highly comparable total RMSE for the 21 ability combinations at the three correlation levels.

Ability combinations with large RMSE's (=> 2.5) for WLE were combinations 5,7,14,19, and 21 for correlation 0.0; combinations 6, 10, 15, and 21 for correlation 0.3; and combinations 1, 4, 9, 10, 17 and 20 for correlation 0.6. Further examinations of the combinations of thetas failed to reveal any meaningful pattern, suggesting that these larger RMSE's were produced by ability combinations that are most likely idiosyncratic.

Figures 22, 23, and 24 also indicate that WLE tends to have the largest total RMSEs, followed by MLE, MAP, and EAP when Bank II is used. However, based on Bonferroni grouping, the four methods are divided into two groups when correlation levels are 0.0 and 0.3 and divided into three groups when the correlation level is 0.6. At correlation 0.6, the RMSEs of WLE and MLE significantly differ from each other. The result of Bonferroni grouping is also evident by the three plots for Bank II. In Figure 24, total RMSEs of WLE for almost half of the 21 ability combinations are much higher than those of MLE. However, in the other two plots, the total RMSEs of WLE are more comparable with those of MLE. The difference between the the Bayesian methods is also negligible, which is evident by the two comparable solid lines.

A comparison of the total RMSEs based on Bank I and Bank II reveals that the magnitude of total RMSE is smaller when Bank II is used. WLE estimates yield the largest discrepancies between the two banks, followed by MLE. The total RMSEs for the two Bayesian methods are smaller when Bank II is used, but the discrepancies are not as great as the non-Bayesian methods. This finding is also consistent with the results from both bias and SE analysis in the previous sections, suggesting that item banks with item structure of one primary dimension and some secondary dimensions are more desirable for a MCAT system.

Figures 19 to 24 also reveal that each method yields similar total RMSE under different correlation levels, meaning that for any of the four methods, comparable size of total RMSEs will be resulted no matter at what correlation levels the ability combinations are. The plots do show, however, that there may be an interaction effect between the methods and the ability combinations, meaning that some method yields a smaller RMSE at a certain ability combination and a larger RMSE at another ability combination, compared with the other method. For example, if we look at Figure 21, it clearly shows that at the 13th $(-0.8293, -0.8132, -0.4131)$ and the 17th $(0.1961, 1.4592, 1.1999)$ ability combinations, WLE and MLE perform differently in the opposite direction. By connecting the two WLE RMSEs and the two MLE RMSEs, we obtain a disordinal interaction plot, meaning the rank order of the levels of estimation methods changes at the different levels of ability combination. Basically, the interaction effect occurs mainly because the two non-Bayesian methods perform differently at different ability combinations.

To summarize, the results for RMSE mirror the combined effects on SEs and bias discussed earlier. The RMSE plots indicate that there may be an interaction effect of estimation methods by ability combinations when Bank I is used and that the possible interaction effect occurs mainly because the two non-Bayesian methods perform differently for some ability combinations. In general, the two Bayesian yield smaller total RMSEs , followed by MLE, and WLE yield the largest under both banks. The four methods result in similar magnitude of RMSE at different correlation levels, which is also consistent with the findings based on the analyzed results of bias and SE.

Test Information

Test information plots are presented in Appendix D. Figures 25 and 26 present the test information plots for the simulated test at correlation 0.0 using four different estimation methods at test length of 40 and 50 items when Bank I was used.

It is clear that WLE yield the most test information, followed by MLE, and then the two Bayesian in both plots. Test information for 50 items are greater than that for 40 items, since there are additional information from ten more items administered. MAP and EAP have almost identical test information for all the 21 ability combinations.

Figures 27 and 28 present the test information plots for the simulated test at correlation 0.0 using four different estimation methods at test length of 40 and 50 items when Bank II was used. WLE still yield the most test information, followed by MLE, and then the two Bayesian for both test lengths. It is noticed that test information based on Bank II is much smaller than that based on Bank I. This is due to the higher discriminating parameters in Bank I which affect the item information.

It is a well known fact that there is an inverse relationship between standard error of estimation and test information. Findings from the present study seem to contradict this common knowledge because not only WLE and MLE produced larger standard errors, they also generated larger test information. Possible factors affecting the size of SEs and test information are further explored below.

Relationships between SEs and Test Information in MCAT

It has been mentioned in the review chapter that an *unbiased* MLE estimator $\widehat{\vec{\theta}}$ of estimating $\vec{\theta}$ itself is a minimum variance bound estimator (Kendall & Stuart, 1977) *i.e.*,

$$\mathrm{Var}(\widehat{\vec{\theta}}|\vec{\theta}) = (I(\vec{\theta}))^{-1} \tag{14}$$

The above equation establishes when the MLE is unbiased. However, for short tests, it has been shown that MLE is strongly biased (Wang & Vispoel, 1998; van der Linden, 1997).

Warm (1989) has shown that WLE estimator is asymptotically normally distributed with variance equal asymptotically to the variance of MLE estimator, that is, $VAR(WLE(\theta)) \simeq VAR(MLE(\theta)) \simeq I^{-1}$. Warm also demonstrated with simulated data that such information-based approximations for error variance systematically underestimate actual error variances when CATs are terminated at short test lengths. This result has important practical implications, because test users unaware of this problem may overestimate the accuracy of MLE ability estimates derived from short-length CATs. This finding also underscores the importance of determining test lengths for operational CATs for which information-based variance errors yield sufficiently accurate results.

The results from the MCAT simulation also showed that test information-based SEs underestimate actual SEs of MLE and WLE estimators when the test lengths are 40 and 50 items for a three-dimensional test. Figures 29 and 30 presents the plots for the actual average SEs and the test information-based SEs for WLE at test lengths of 40 items and 50 items. It is rather clear that the test information-based SEs are much smaller than the actual SEs at each ability combination. The test information-based SEs are all around 0.1.

If we look at the test information plots for Bank II in Figure 27, the information is ranging between 80 to 100 for MLE and WLE. If we take square root of 80 or 100 and then taking reciprocal, we obtain .1 for test information of 100 and .111 for test information of 81. Actually, for test information of 36, the estimated SE will be .167. The actual average SEs are also listed on the tables in section 5.2. The average SE for each combination were obtained by averaging the three SEs from each three dimensions. For the test length of 40 items, the discrepancies between the actual and estimated are even larger. It implies that test information-based SEs underestimate the actual SEs for WLE and that the magnitude of underestimation is increased when the test length gets shorter. The same underestimation occurred for the MLE estimators, which is evident by Figures 31 and 32. The great discrepancies between the actual and estimated SEs may imply that a lot more items are needed to be administered in order to get a sufficiently accurate results.

In Wang & Vispoel (1998), they also pointed out that when item discrimination is high, the test information-based SEs also underestimated the actual SEs for MLE estimators. This is due to the test information is the summation of item information, which is a function of item parameters. The higher discriminating parameter is for an item, the higher item information is. The three discriminating parameters in Bank I of our study are distributed uniformly between 2 and .4. Compared with the discriminating parameters in Bank II, two of the three a_i are larger than those for the two secondary dimensions. Therefore, it can be expected that the test information-based SEs for MLE and WLE estimators are underestimating the actual SEs even more. This is evident by Figures 33 to 36. As it has been shown in the SE section, the SEs of ability estimators for the four estimation methods in Bank I were much higher than those in Bank II. However, the test information based on the simulations under Bank I is higher due to the high discriminating parameters. If test information-based SEs were used to approximate the SEs for the MLE and WLE estimators, there would be a hugh underestimation occurred and the results would be misleading.

Lord (1980) also showed that the information function is an upper bound to the information that can be obtained by any method of scoring the test. Therefore, $I(\vec{\theta})$ can be used as the information function in MAP though it is an upper bound. In Wang & Vispoel (1998), posterior distribution-based SEs were used to approximate the actual SEs for the MAP and EAP estimators. They concluded that the posterior distribution-based SEs are good estimators for the SEs of the Bayesian estimators.

Summary and Conclusions

The purpose of this study was to extend Warm's (1989) weighted likelihood estimation to a multidimensional adaptive test setting. WLE was compared with the other three scoring methods(MLE, MAP and EAP) under various CAT conditions. The goals of this investigation were to evaluate the accuracy and the properties of ability estimates for WLE, MLE, MAP and EAP under various MCAT situations, to examine the effects of the degree of the intercorrelation between underlying abilities on ability estimation, and hopefully to provide guidelines for the selection of a particular ability estimation method for the multidimensional IRT models in the CAT environment. In this section, the significant findings are summarized.

This is the first empirical study about Warm's weighted likelihood method for estimation of ability and comparisons of the accuracy of ability estimation methods between Warm's method and other commonly used methods under the multidimensional 3PL IRT model in the CAT environment. The results from the MCAT simulations did show some advantages of WLE in reducing the bias of MLE under the approximate simple structure of Bank II with a fixed test length of 50 items, which was consistent with the previous research findings based on different unidimensional models (Warm, 1989; Wang, Hanson, & Lau, 1999).

In the following, the findings of this Monte Carlo study are summarized according to the order of the research questions.

(1) The main advantage of the Bayesian methods over MLE and WLE is that they yield lower SEs (i.e., less random estimation error) and smaller RMSEs under both multidimensional structures. WLE performs the most differently when different multidimensional structures were used. Under the multiple dominant structure of Bank I, WLE yields large SEs and large bias at the high end of ability scale. But when the approximate simple structure of Bank II was used, the SEs for WLE are reduced greatly and WLE yields the smallest bias at correlation level 0.3 and the second smallest bias at correlations 0.0 and 0.6 based on the ANOVA results. All four methods yield smaller bias and smaller SEs when the approximate simple structure was employed in the MCAT simulations.

The ANOVA on the bias indicates that MLE yields the largest bias when Bank II was employed. This probably can be attributed to the test length and the sampling of the 21 ability combinations. For Bank II, a test length of 50 items means that the average items administered for each dimension are only 16. The MLE estimates can not be stabilized as fast as the two Bayesian methods at this short test length. van der Linden (1997) also pointed out that MLE is strongly biased when the test length is not long enough. Another reason for the MLE bias to stand out is that for some dimensions of the generated 21 ability combinations, the extreme theta values were not sampled due to the low probability in the extreme of the distribution. The disadvantage of the Bayesian methods is most due to the large bias in the extreme end of the distribution. Since some of the extreme points were not sampled, the disadvantage of the two Bayesian is diminished. Compared with MLE, WLE performs better under Bank II, even though no statistically significance was found among the four estimation methods on the bias when Bank II was used.

In terms of test information, WLE yields the most test information, followed by MLE, and then the two Bayesian methods. This finding seems to contradict the findings of SE, which is inversely correlated with test information. A possible explanation is the short test lengths employed. In a short test, not only WLE and MLE are extremely biased, test information-based SEs are poor estimators of actual SEs. Thus the relationship between test information and actual SEs may not turn out as expected.

(2) The criteria for estimation accuracy (bias, SE, and RMSE) all do not indicate any significant effect of intercorrelation between dimensions on the ability estimation for any of the four methods. With the 21 ability combinations at each correlation level, we are able to examine the effect of intercorrelation between dimensions on the ability estimation. The results from all the criteria do not support that there is any different effect on the accuracy of ability estimates from different levels of correlation. One possible explanation is that the IRT model that was used in the investigation does not provide the information that is related to the relationship between dimensions.

(3) Based on the results from the ANOVA on the bias at each correlation level under Bank II, it is clear that the same main effect and interaction effects occurred at each correlation level. Also at each correlation, MLE yields the highest bias, followed by MAP. Even though EAP yields the smallest bias at correlation level 0.0 and 0.6 and WLE yields the smallest bias at correlation 0.3. There is no significant difference on bias found among the four estimation methods; however there is an interaction between ability levels and methods. The two Bayesian methods resulted in overestimations at the low end of the ability scale and underestimations at the high end of the ability scale under Bank II. WLE and MLE are more consistent and unbiased along the ability scale under Bank II.

Appendix A

Multidimensional Weighted Likelihood Estimation (MWLE)

Let $V = \{v_1, v_2, \cdots\}$ be a set containing the identifiers of the adaptively administered items. The likelihood of a vector of observed responses $\vec{u}$ given ability $\vec{\theta}$ is expressed by

$$L(\vec{u}|\vec{\theta}) \stackrel{\text{def}}{=} L(u_{v_1}, u_{v_2}, \cdots, |\vec{\theta}) = \prod_{i \in V} P_i(\vec{\theta})^{u_i} Q_i(\vec{\theta})^{1-u_i} \tag{15}$$

where $P_i(\vec{\theta})$ is defined by (5).

The maximum likelihood (ML) estimates are the solution to the set of p simultaneous equations given by

$$\frac{\partial}{\partial \vec{\theta}} \ln L(\vec{u}|\vec{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(\vec{u}|\vec{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln L(\vec{u}|\vec{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(\vec{u}|\vec{\theta}) \end{bmatrix} = 0 \tag{16}$$

The derivative of the log likelihood with respect to θ_k, $k = 1, 2, \cdots, p$, takes on a form similar to the univariate 3PL model (Lord, 1980):

$$\frac{\partial}{\partial \theta_k} \ln L(\vec{u}|\vec{\theta}) = \sum_{i \in V} (u_i - P_i(\vec{\theta})) \frac{P_i'(\vec{\theta})}{P_i(\vec{\theta}) Q_i(\vec{\theta})} \tag{17}$$

By extending Warm's weighted likelihood estimator (1989), a class of estimators, $\vec{\theta}^*$, could be defined as the value of θ that maximizes

$$w(\vec{\theta}) L(\vec{u}|\vec{\theta}) = w(\vec{\theta}) \prod_{i \in V} P_i(\vec{\theta})^{u_i} Q_i(\vec{\theta})^{1-u_i}, \tag{18}$$

for a suitably chosen function w. $\vec{\theta}^*$ was found at the zero of the following estimation equations

$$\begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(\vec{u}|\vec{\theta}) + \frac{\partial}{\partial \theta_1} \ln w(\vec{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln L(\vec{u}|\vec{\theta}) + \frac{\partial}{\partial \theta_2} \ln w(\vec{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(\vec{u}|\vec{\theta}) + \frac{\partial}{\partial \theta_p} \ln w(\vec{\theta}) \end{bmatrix} = 0 \tag{19}$$

If $w(\vec{\theta})$ is a positive constant, $\vec{\theta}^*$ is a maximum likelihood estimate of $\vec{\theta}$, MLE($\vec{\theta}$). If $w(\vec{\theta})$ is an assumed prior density function of $\vec{\theta}$, then (19) is regarded as the posterior density function, and $\vec{\theta}^*$ is a Bayesian modal estimate of $\vec{\theta}$, BME($\vec{\theta}$) (Lord, 1983; 1984).

For the univariate 3PL model, Lord (1983) derived the asymptotic expression for the bias of MLE(θ), denoted by BIAS(MLE(θ)). For the multivariate 3PL model, we derived the asymptotic expression for the bias of MLE(θ_k,) for any $\theta_k, k = 1, 2, \cdots, p$,

$$\text{BIAS}(\text{MLE}(\theta_k)) = \frac{-J_k}{2I_k^2} \tag{20}$$

where

$$I_k = \sum_{i \in V} \frac{(P_i'(\vec{\theta}))^2}{P_i(\vec{\theta})Q_i(\vec{\theta})} \tag{21}$$

$$J_k = -3.4 \sum_{i \in V} a_{ki} \frac{(P_i'(\vec{\theta}))^2}{P_i(\vec{\theta})Q_i(\vec{\theta})} [\frac{P_i(\vec{\theta}) - c_i}{1 - c_i} - \frac{1}{2}] \tag{22}$$

Lord (1984) also gave the bias of BME(θ) with a standard normal prior. In the multivariate 3PL model, we could extend his result by simply replacing θ by θ_k for $k = 1, 2, \cdots, p$, which is

$$\text{BIAS}(\text{BME}(\theta_k)) = \text{BIAS}(\text{MLE}(\theta_k)) - \frac{\theta_k}{I_k} \tag{23}$$

The last term on the right hand side of (23) is the derivative, with respect to θ_k, of the log of the standard normal density divided by test information. From this observation, we could conjecture that the bias of the estimator defined by (19) is

$$\text{BIAS}(\text{WLE}(\theta_k^*)) = \text{BIAS}(\text{MLE}(\theta_k)) - \frac{\frac{\partial \ln w(\vec{\theta})}{\partial \theta_k}}{I_k} \tag{24}$$

Thus, in order to find the estimator that is unbiased, we set the right hand side of (24) to zero and obtained

$$\frac{\partial \ln w(\vec{\theta})}{\partial \theta_k} = \frac{J_k}{I_k} \tag{25}$$

Substituting (25) for $k = 1, 2, \cdots p$ into (19) yields

$$\begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(\vec{u}|\vec{\theta}) + \frac{J_1}{I_1} \\ \frac{\partial}{\partial \theta_2} \ln L(\vec{u}|\vec{\theta}) + \frac{J_2}{I_2} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(\vec{u}|\vec{\theta}) + \frac{J_p}{I_p} \end{bmatrix} = 0 \tag{26}$$

where I_k and J_k are defined in (22). An estimate satisfying (26) was called a multidimensional weighted likelihood estimate (MWLE).

Appendix B

Figure 1: Bias plot for theta1 estimates at cor=0.0 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

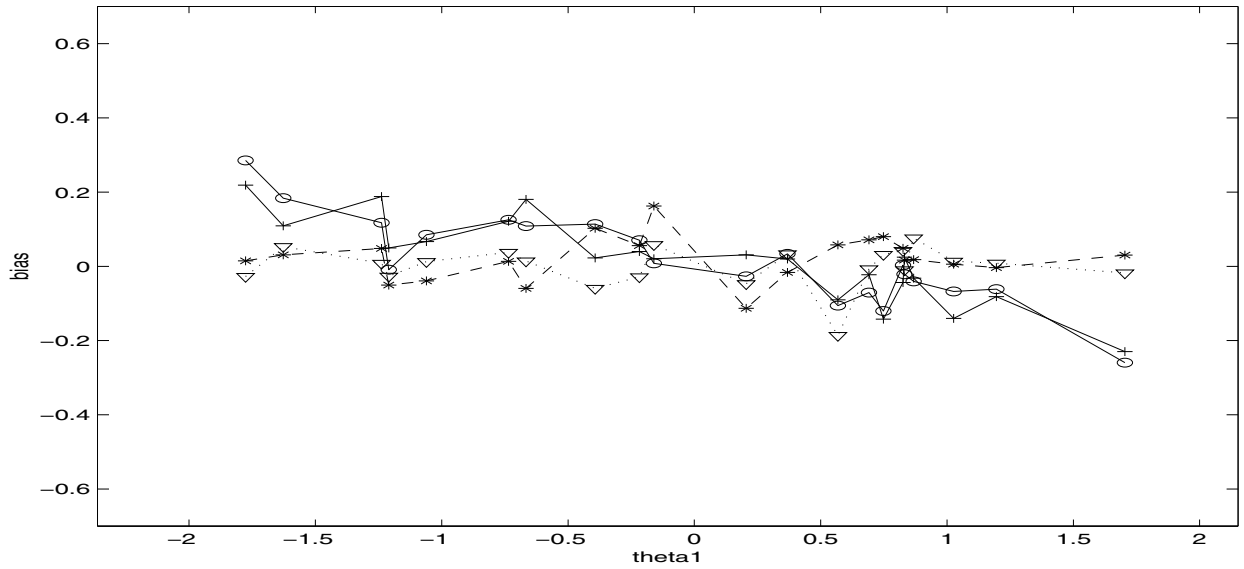


Figure 2: Bias plot for theta1 estimates at cor=0.0 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

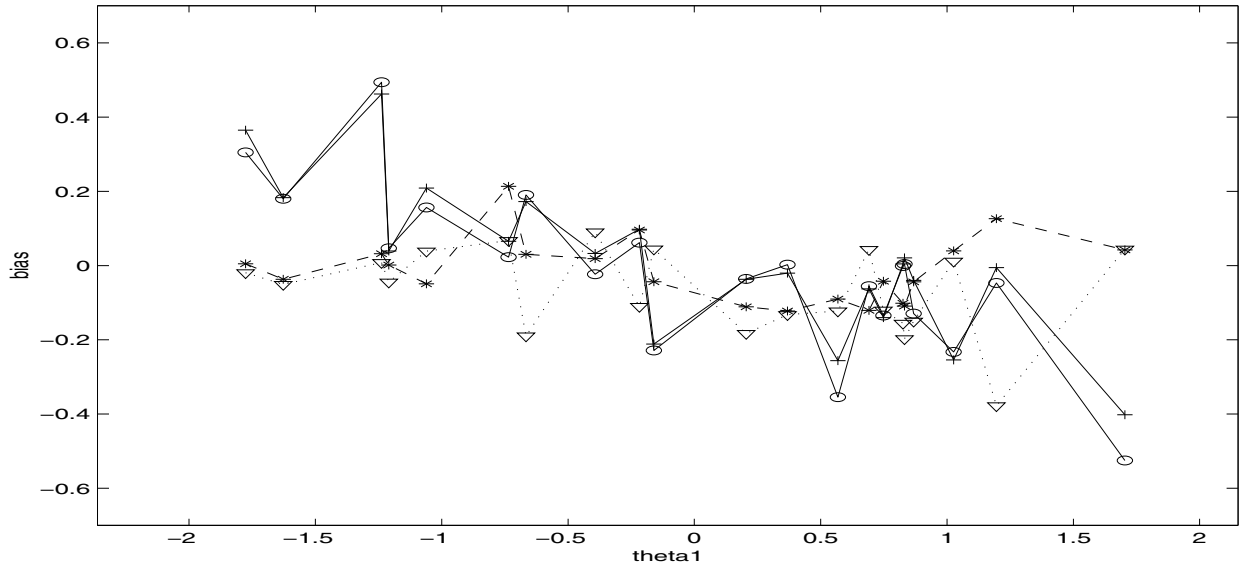


Figure 3: Bias plot for theta2 estimates at cor=0.0 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

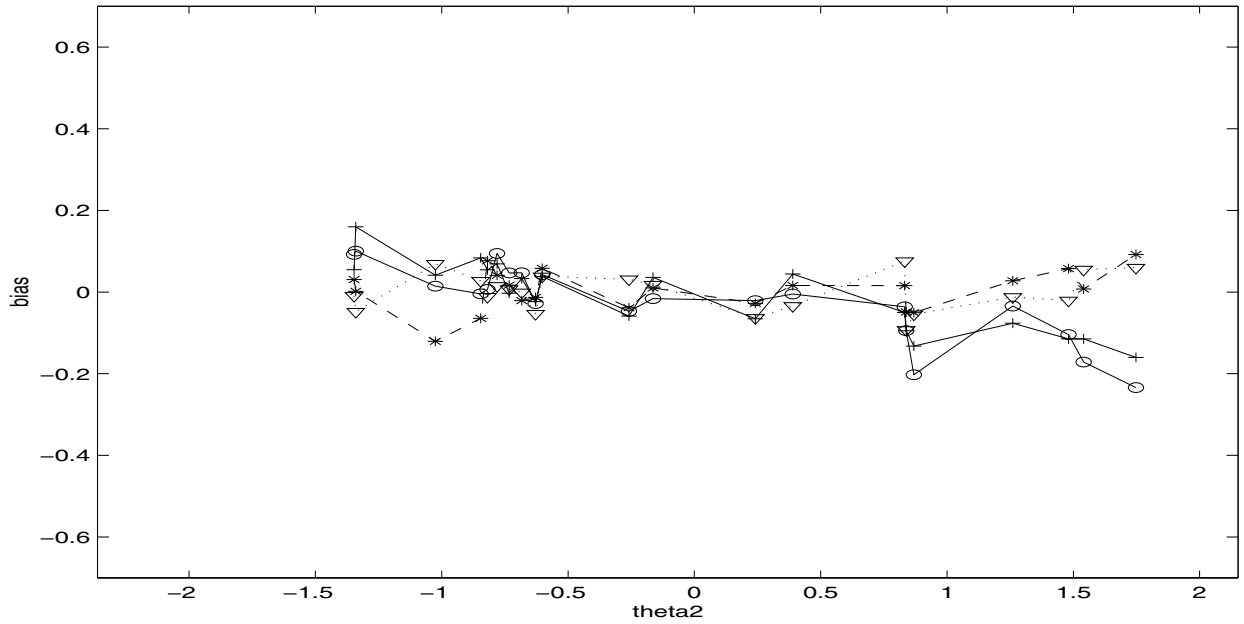


Figure 4: Bias plot for theta2 estimates at cor=0.0 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

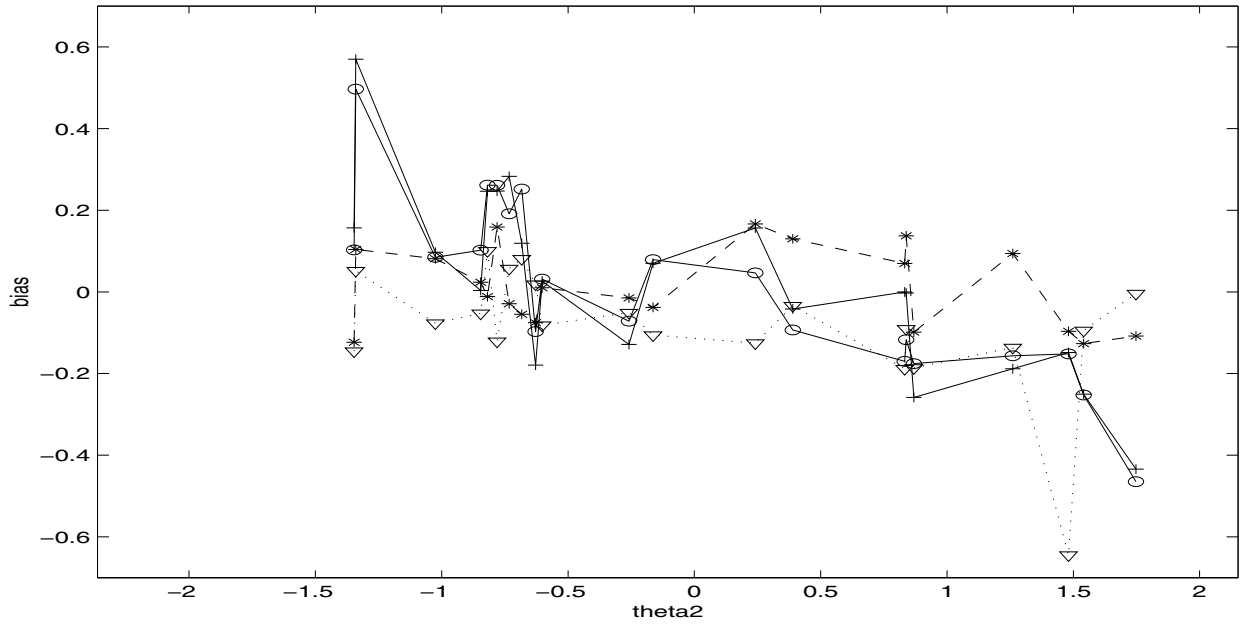


Figure 5: Bias plot for theta3 estimates at cor=0.0 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

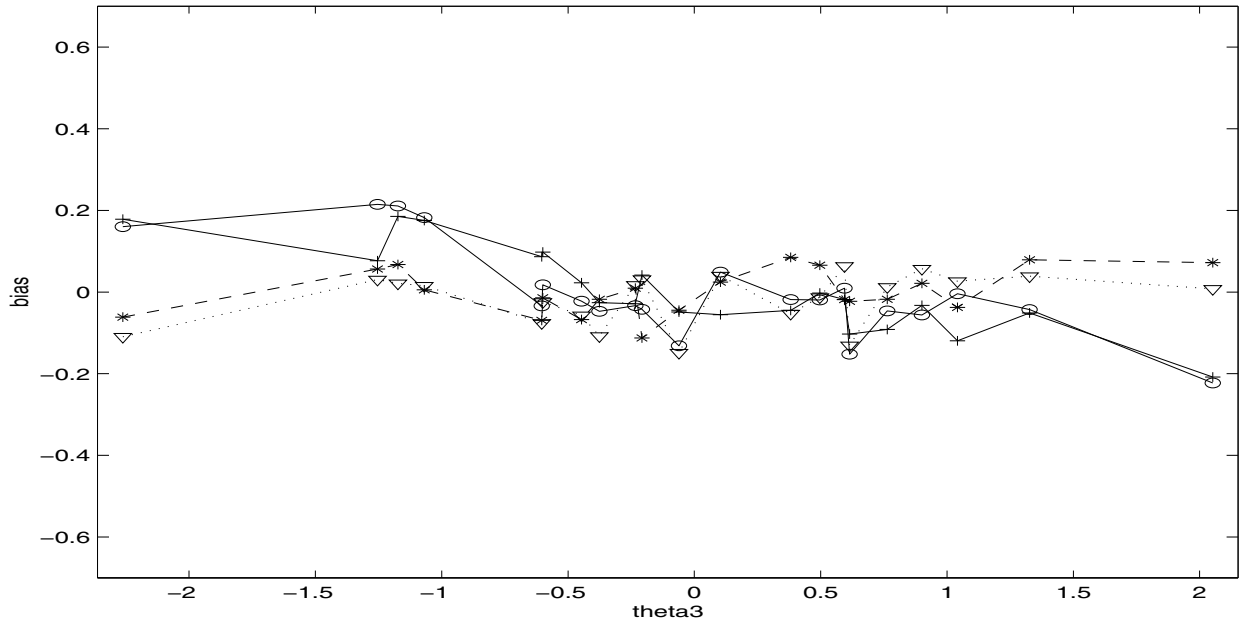


Figure 6: Bias plot for theta3 estimates at cor=0.0 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

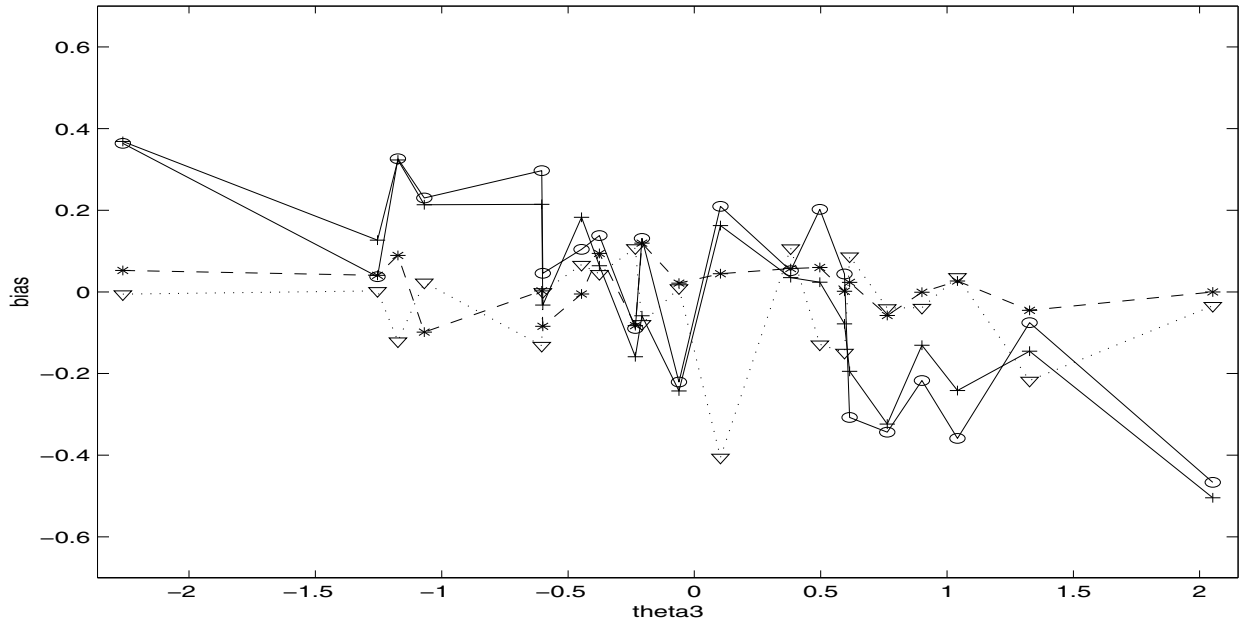


Figure 7: Bias plot for theta1 estimates at cor=0.3 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

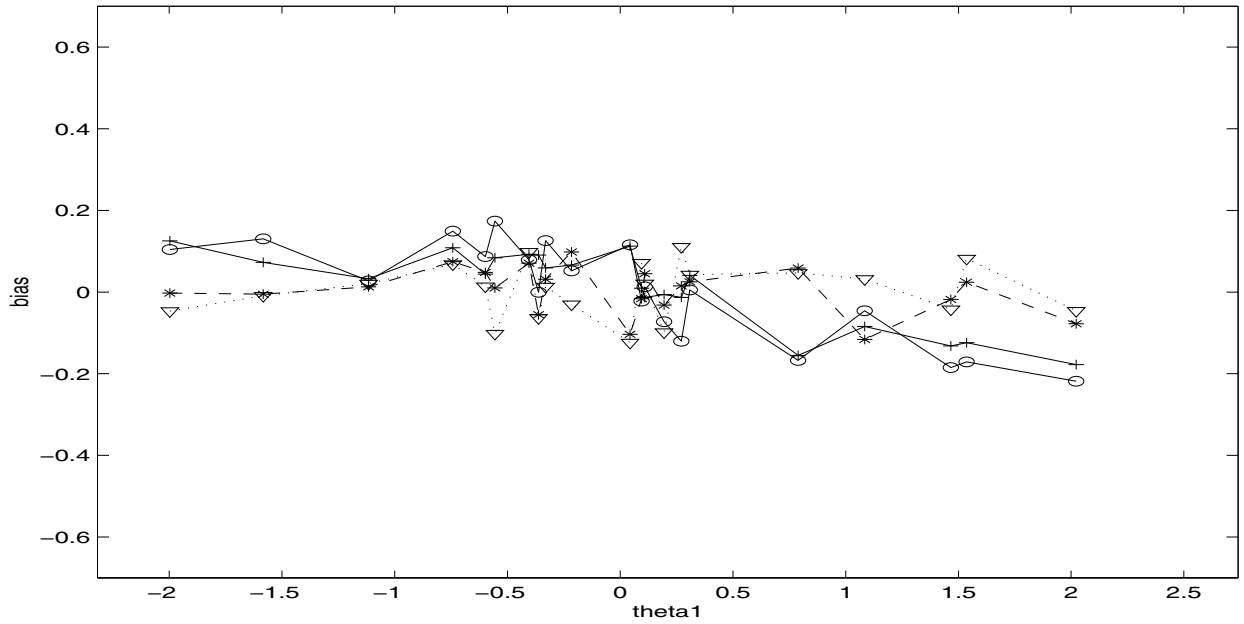


Figure 8: Bias plot for theta1 estimates at cor=0.3 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

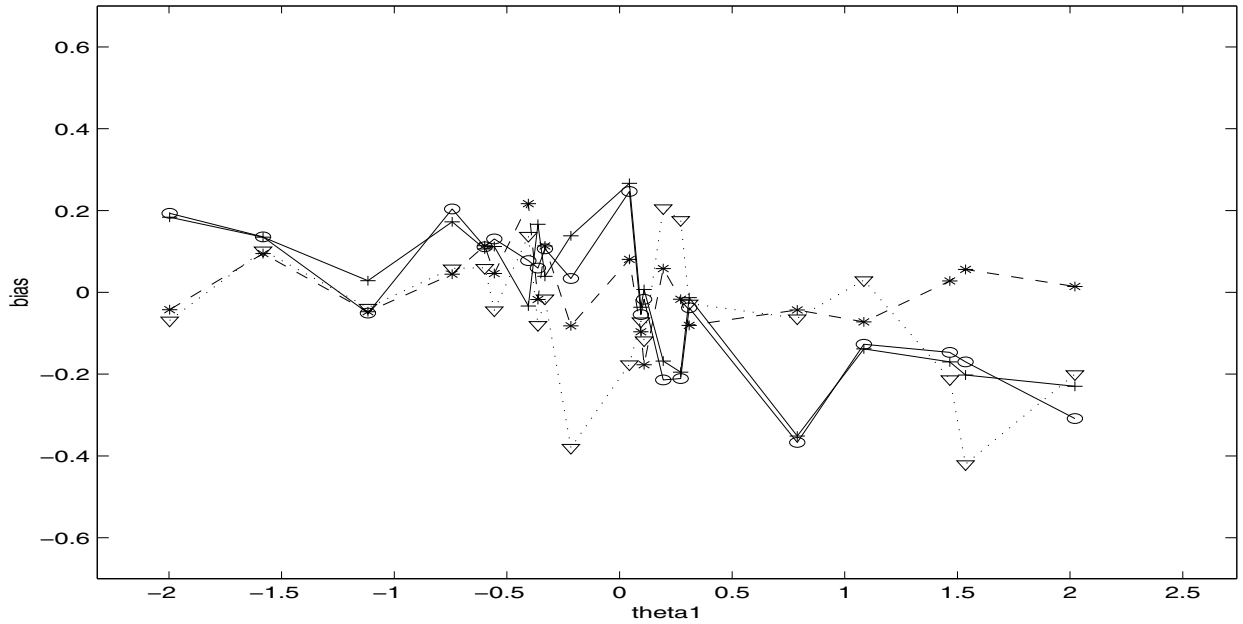


Figure 9: Bias plot for theta2 estimates at cor=0.3 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

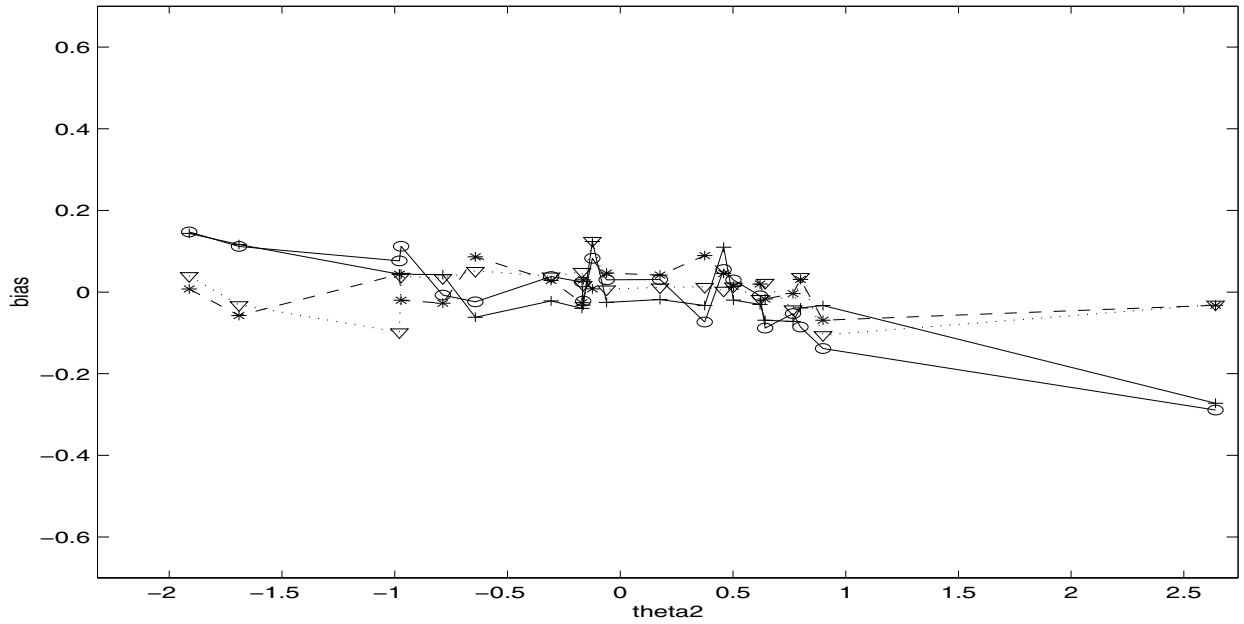


Figure 10: Bias plot for theta2 estimates at cor=0.3 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

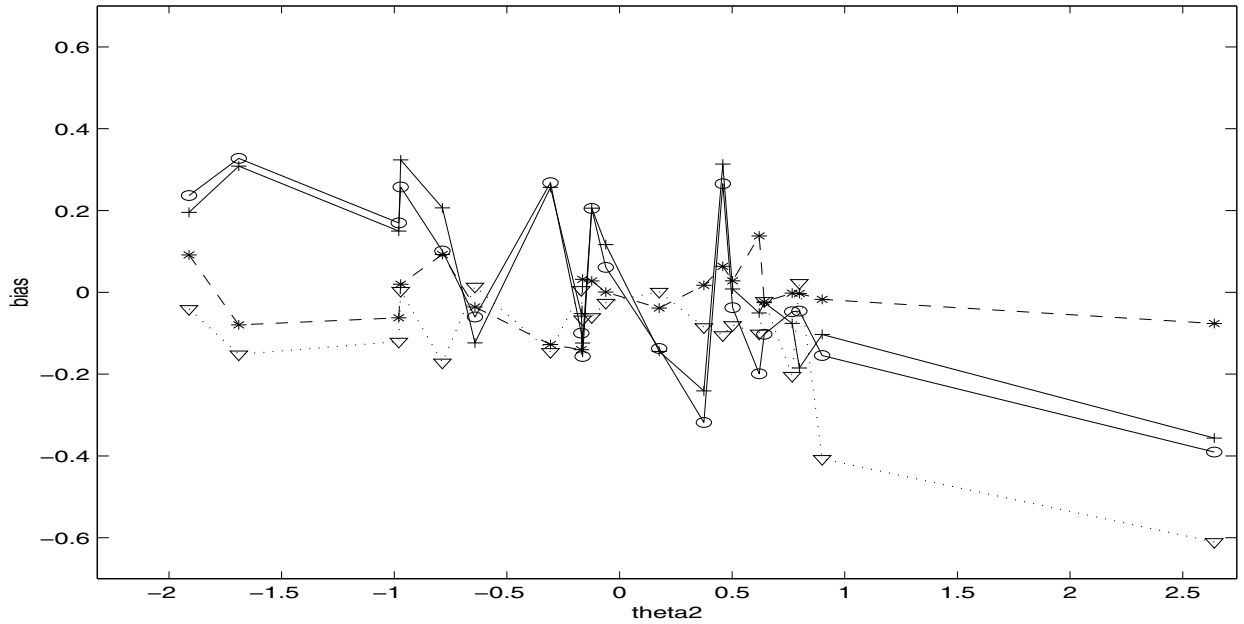


Figure 11: Bias plot for theta3 estimates at cor=0.3 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

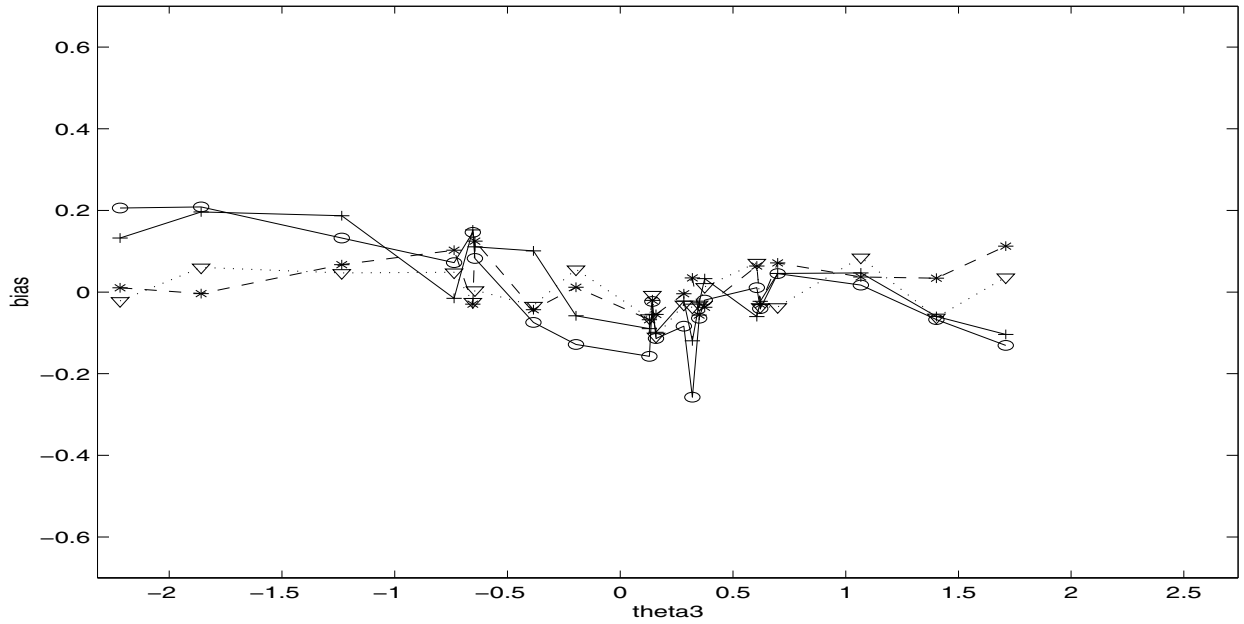


Figure 12: Bias plot for theta3 estimates at cor=0.3 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

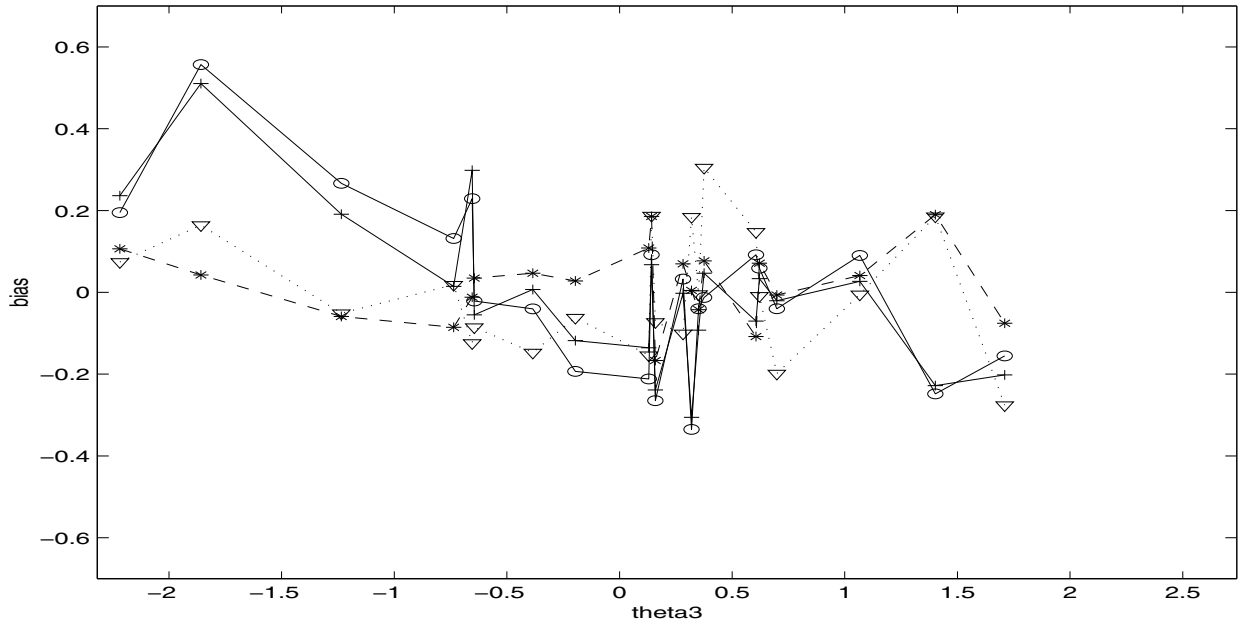


Figure 13: Bias plot for theta1 estimates at cor=0.6 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

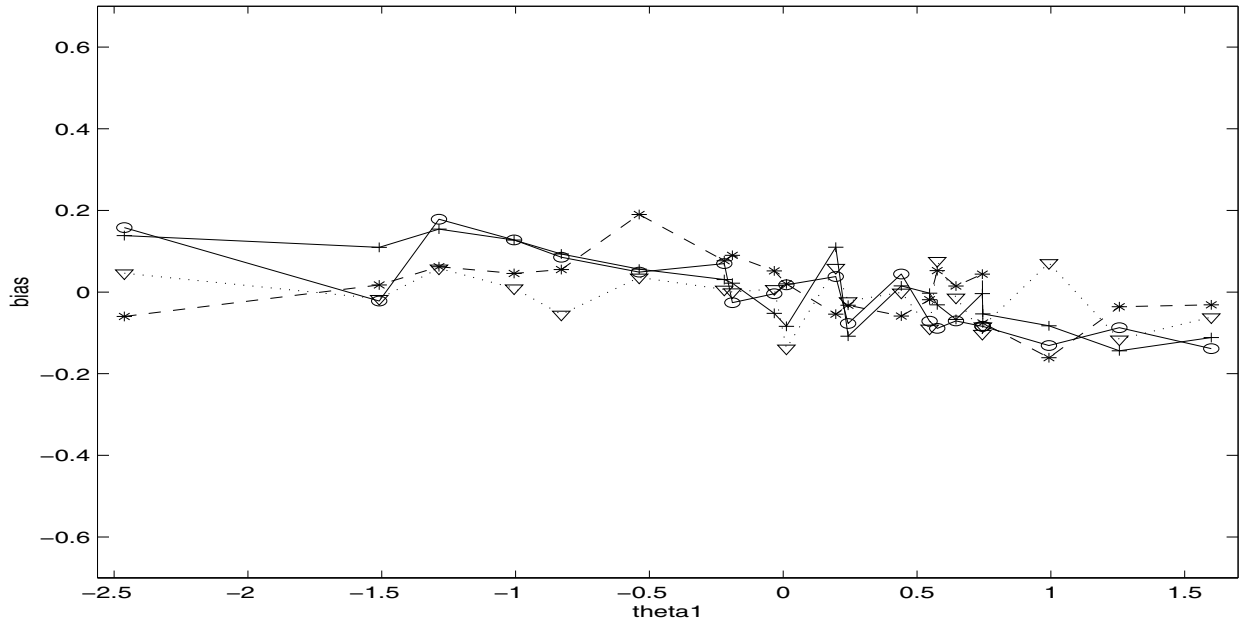


Figure 14: Bias plot for theta1 estimates at cor=0.6 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

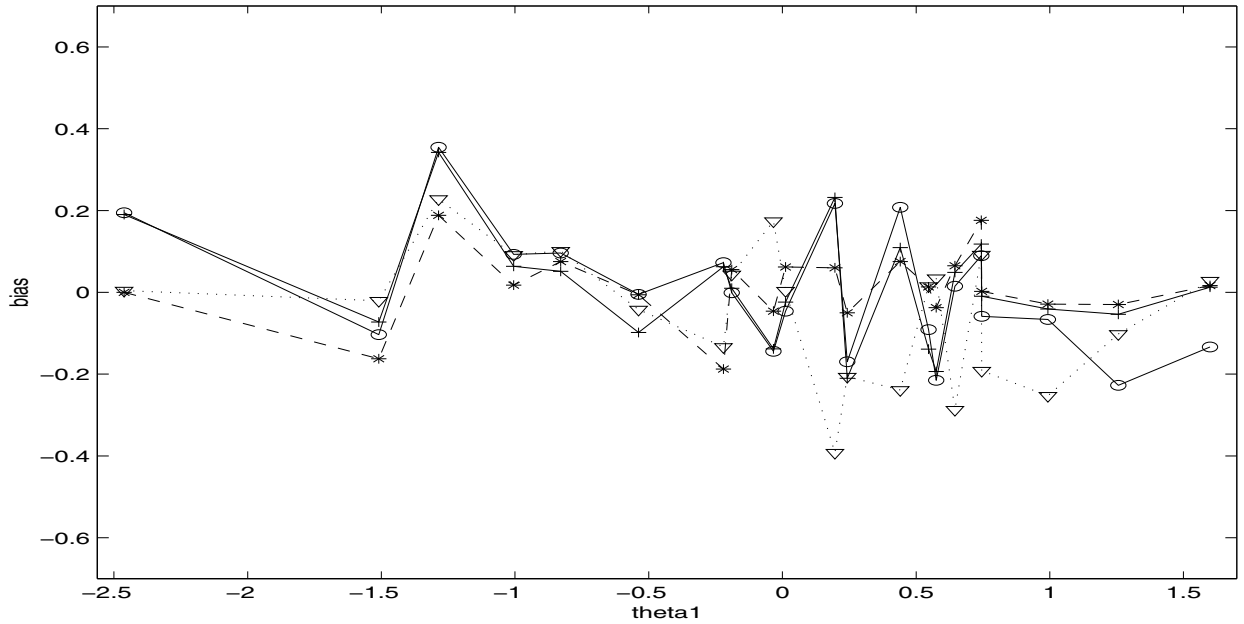


Figure 15: Bias plot for theta2 estimates at cor=0.6 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

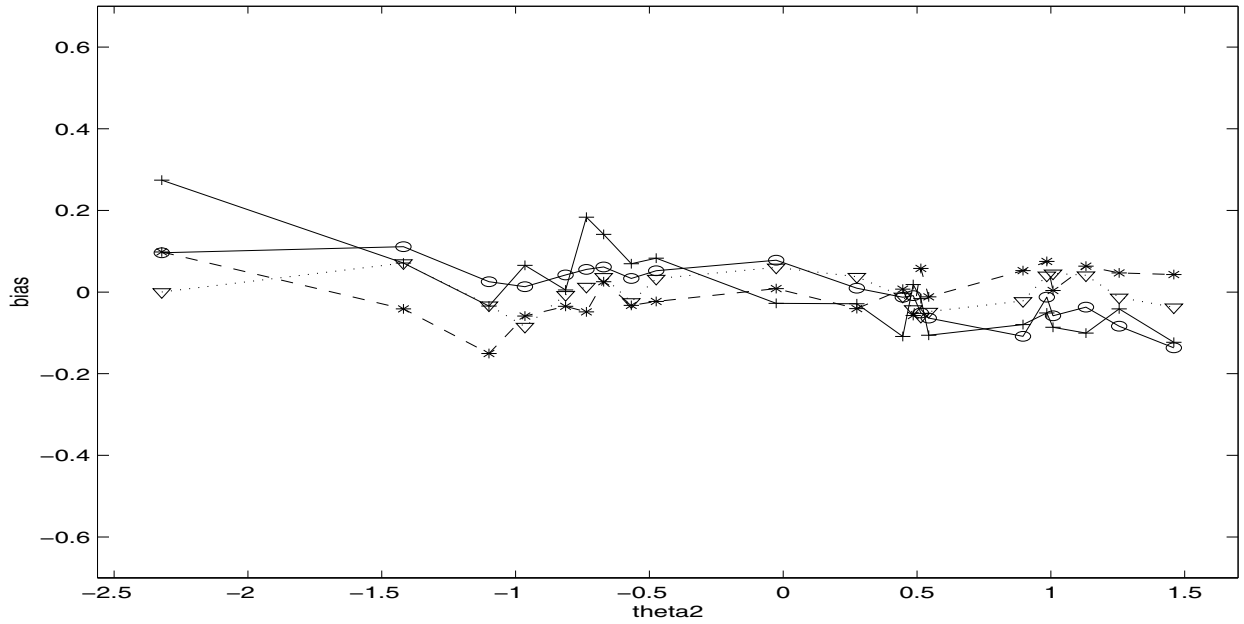


Figure 16: Bias plot for theta2 estimates at cor=0.6 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

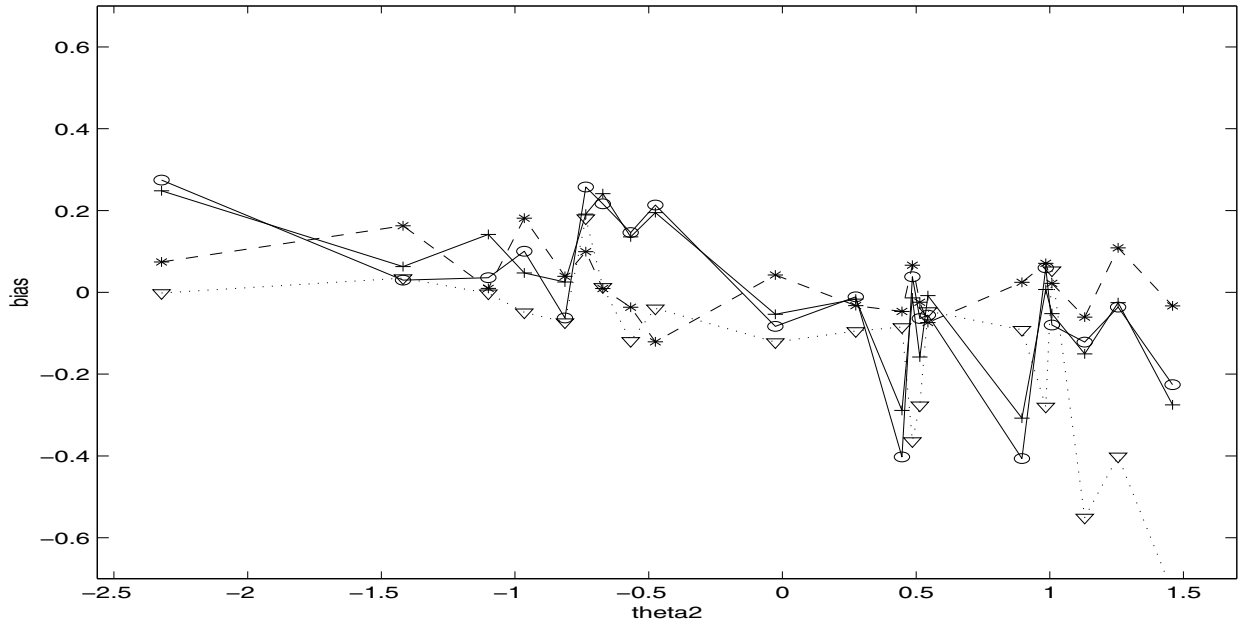


Figure 17: Bias plot for theta3 estimates at cor=0.6 under Bank II, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

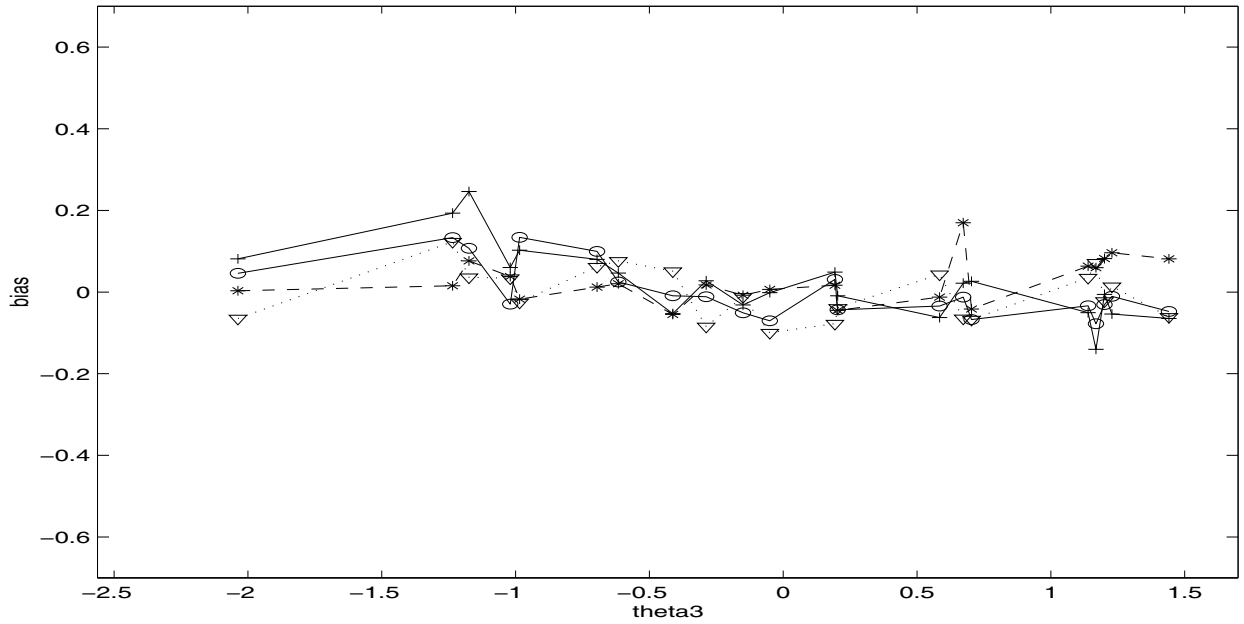
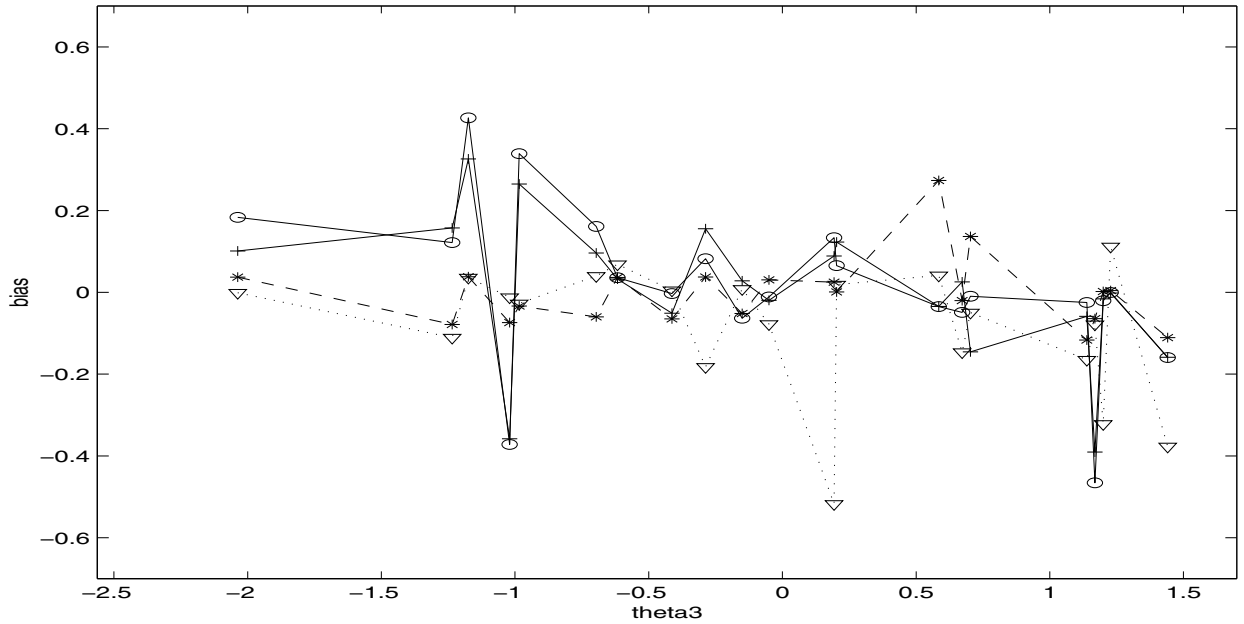


Figure 18: Bias plot for theta3 estimates at cor=0.6 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)



Appendix C

Figure 19: Total RMSE for each ability combination at cor=0.0 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

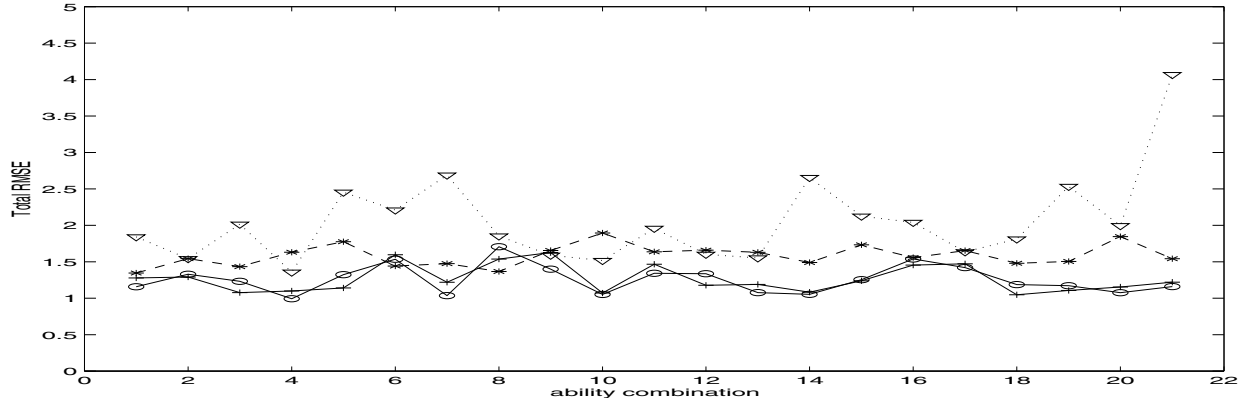


Figure 20: Total RMSE for each ability combination at cor=0.3 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

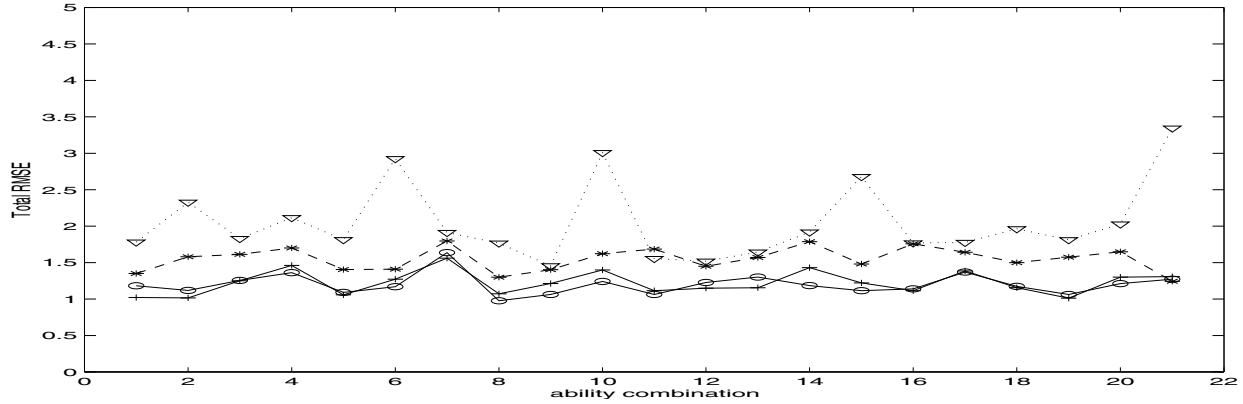


Figure 21: Total RMSE for each ability combination at cor=0.6 under Bank I, MLE(*), WLE($\bigtriangledown$), EAP(o), and MAP(+)

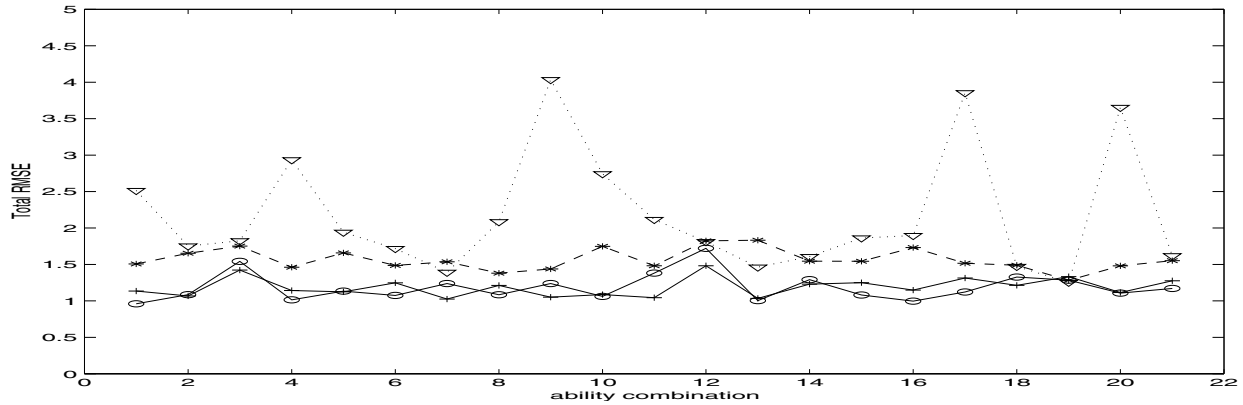


Figure 22: Total RMSE for each ability combination at cor=0.0 under Bank II, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)

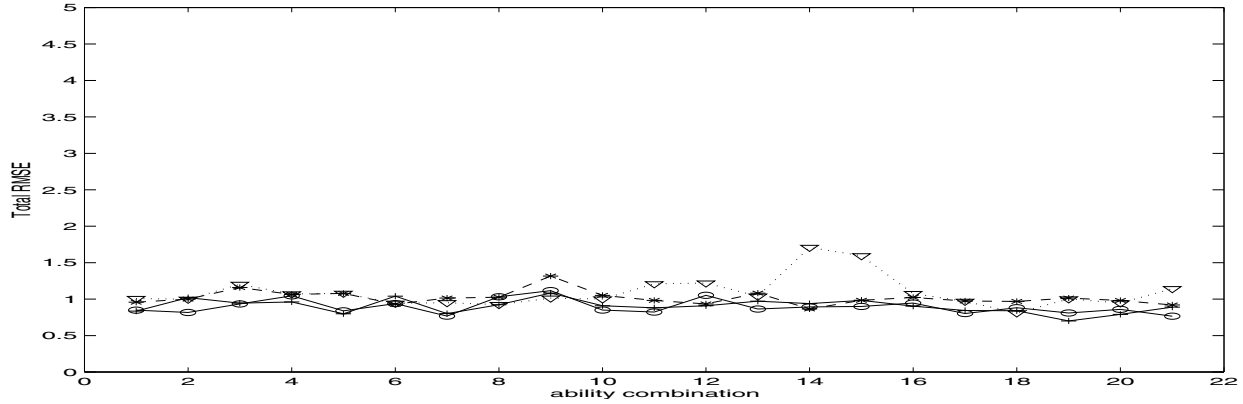


Figure 23: Total RMSE for each ability combination at cor=0.3 under Bank II, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)

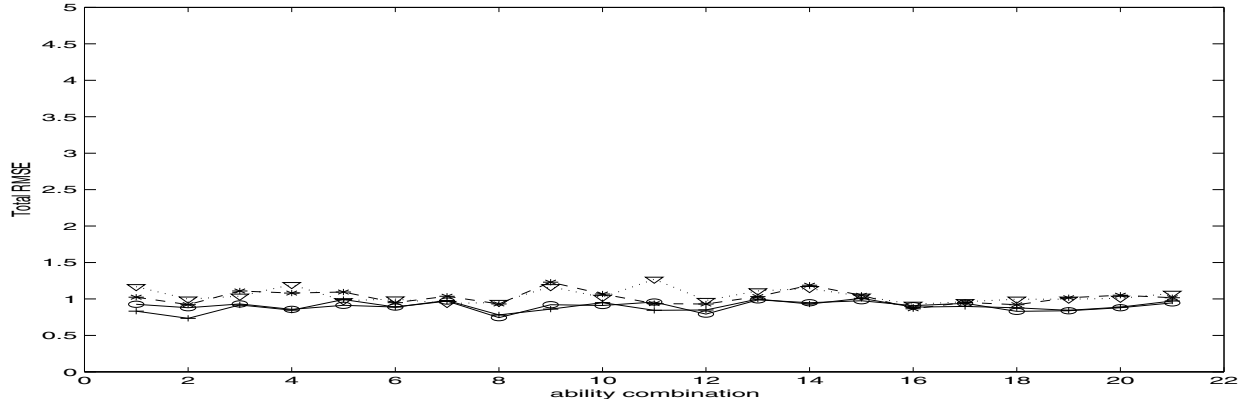
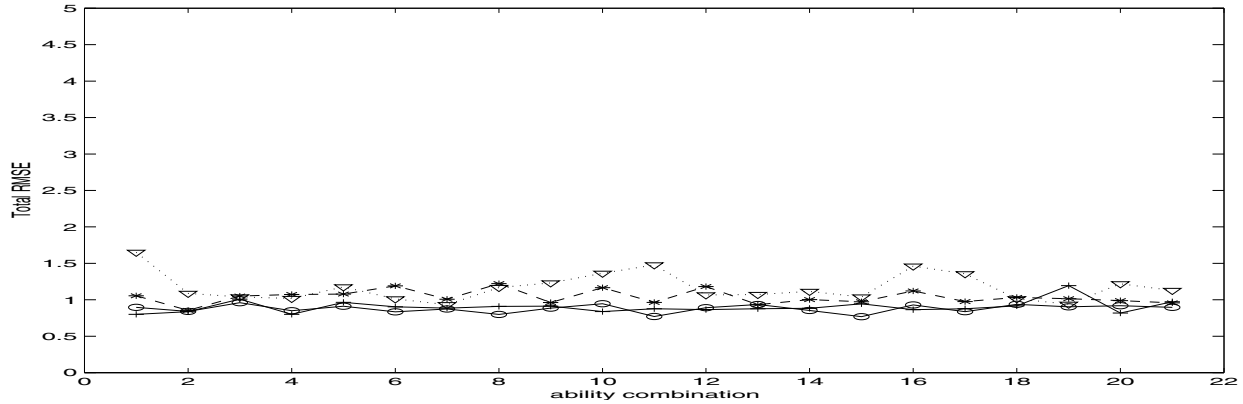


Figure 24: Total RMSE for each ability combination at cor=0.6 under Bank II, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)



Appendix D

Figure 25: Test Information for each ability combination at cor=0.0 at test length of 50 items under Bank I, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)

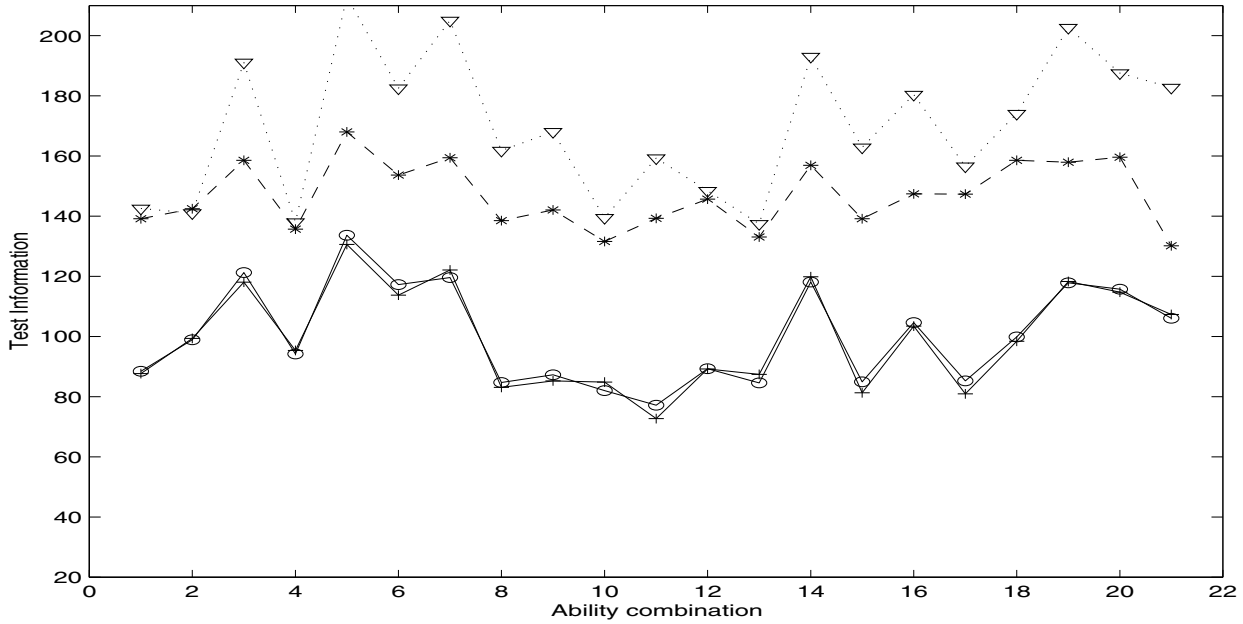


Figure 26: Test Information for each ability combination at cor=0.0 at test length of 40 items Under Bank I, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)

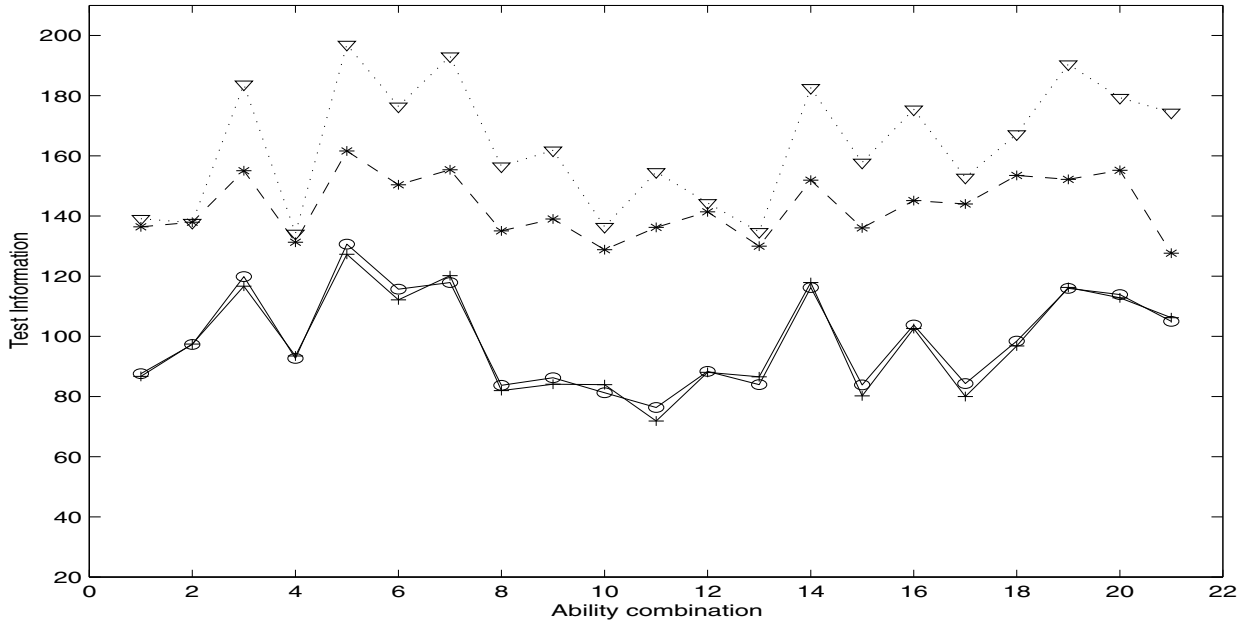


Figure 27: Test Information for each ability combination at cor=0.0 at test length of 50 items under Bank II, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)

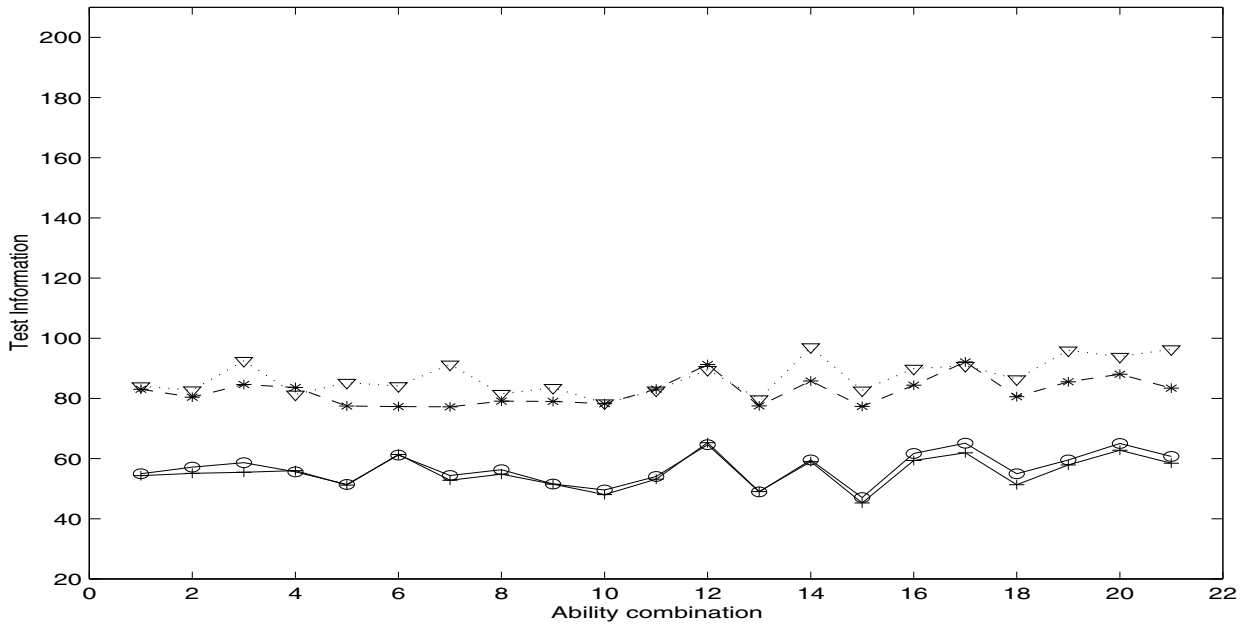


Figure 28: Test Information for each ability combination at cor=0.0 at test length of 40 items Under Bank II, MLE(*),WLE($\bigtriangledown$), EAP(o), and MAP(+)

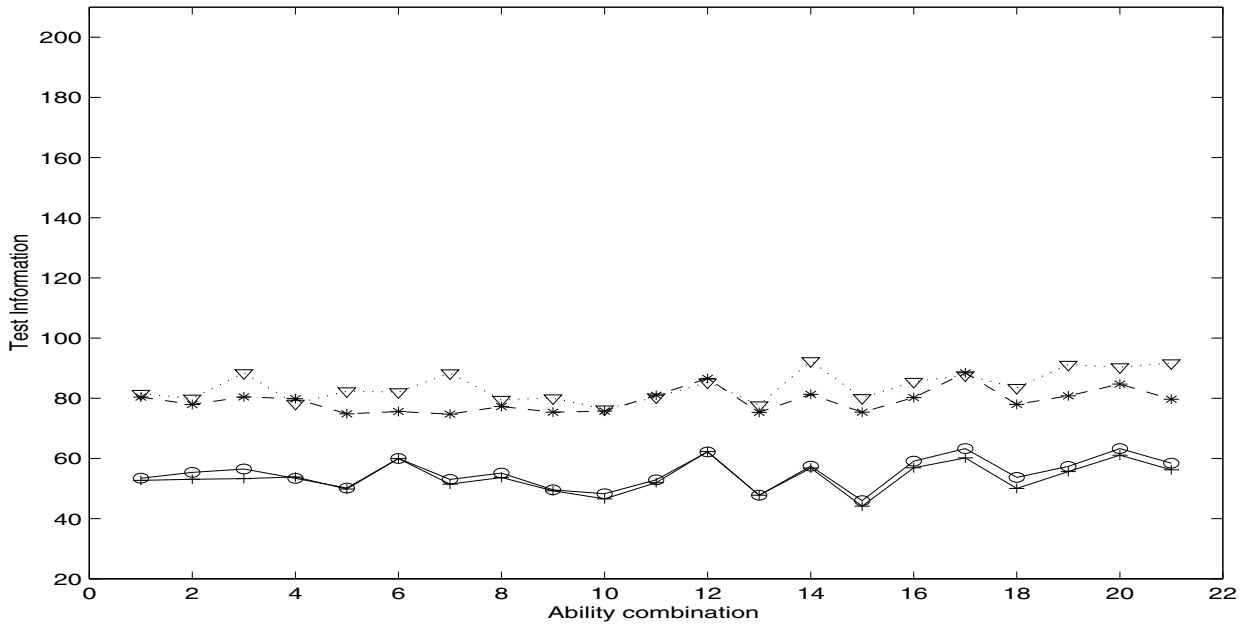


Figure 29: Actual average SEs($\bigtriangledown$) vs. test information-based SEs(*) for WLE at cor=0.0 at test length of 50 items under Bank II

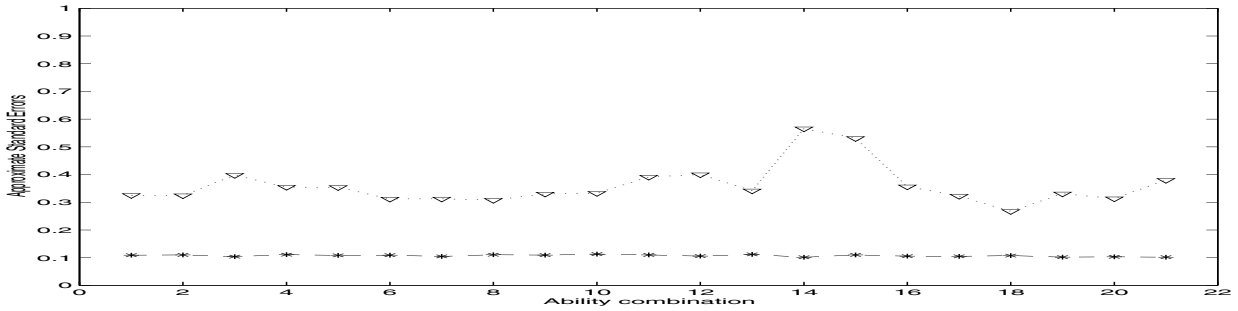


Figure 30: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs(*) for WLE at cor=0.0 at test length of 40 items under Bank II

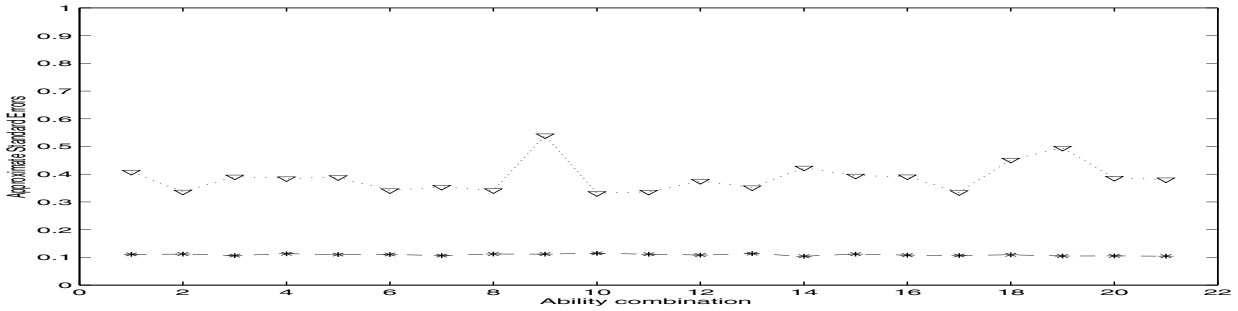


Figure 31: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs (*) for MLE at cor=0.0 at test length of 50 items under Bank II

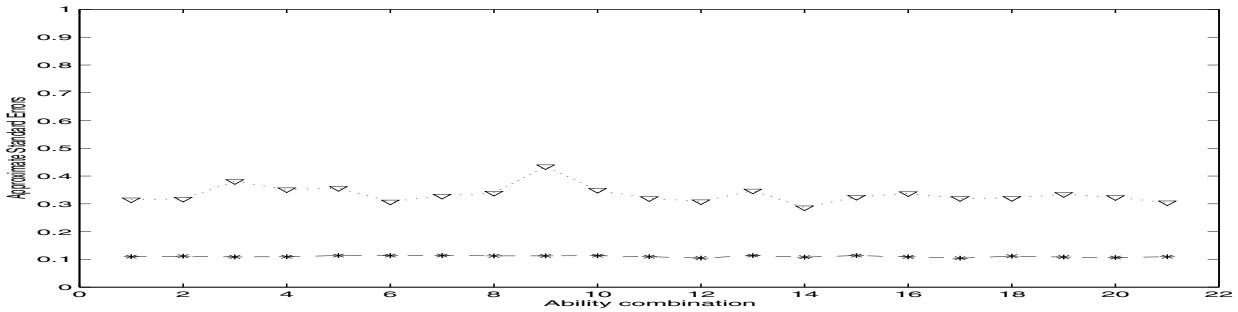


Figure 32: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs (*) for MLE at cor=0.0 at test length of 40 items under Bank II

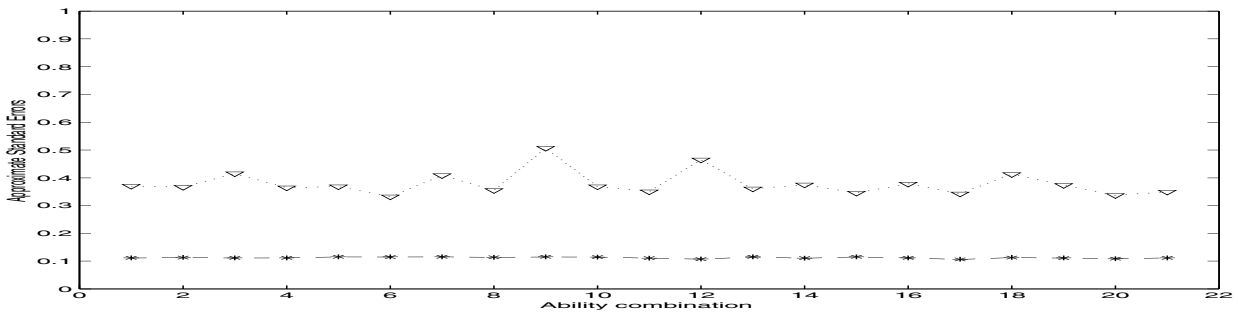


Figure 33: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs (*) for WLE at cor=0.0 at test length of 50 items under Bank I

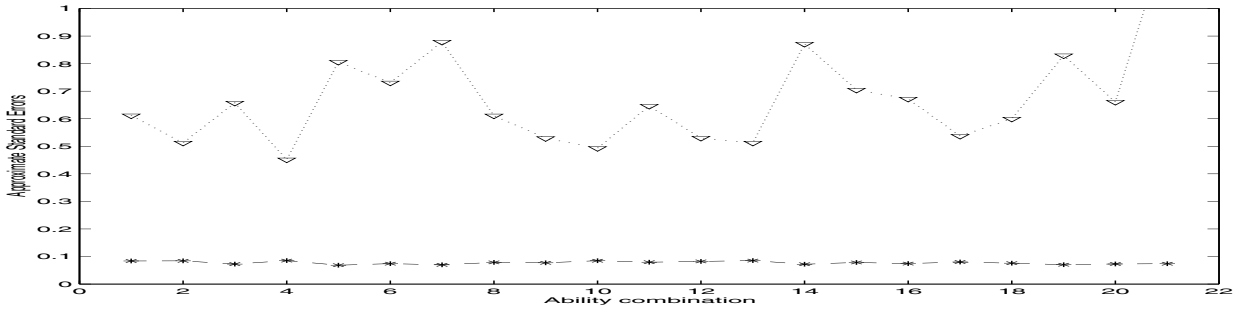


Figure 34: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs (*) for WLE at cor=0.0 at test length of 40 items under Bank I

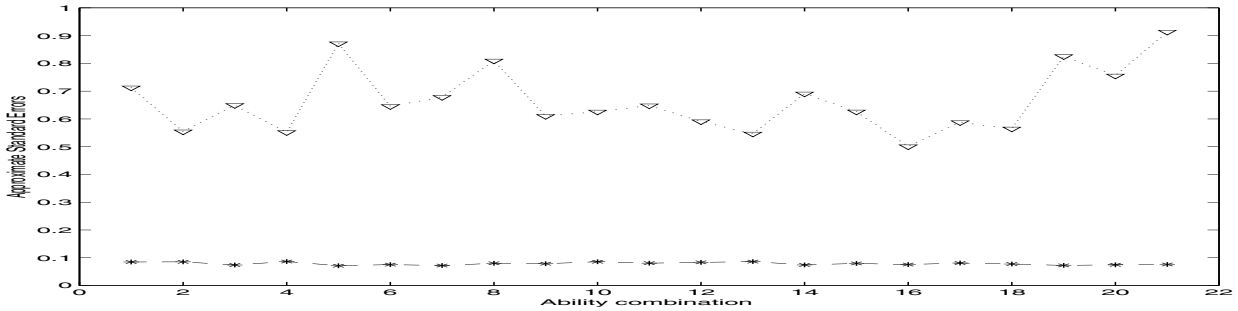


Figure 35: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs (*) for MLE at cor=0.0 at test length of 50 items under Bank I

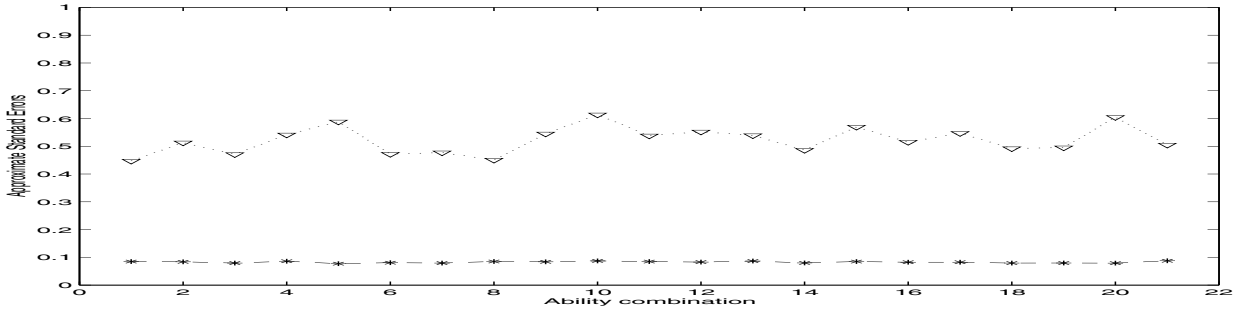
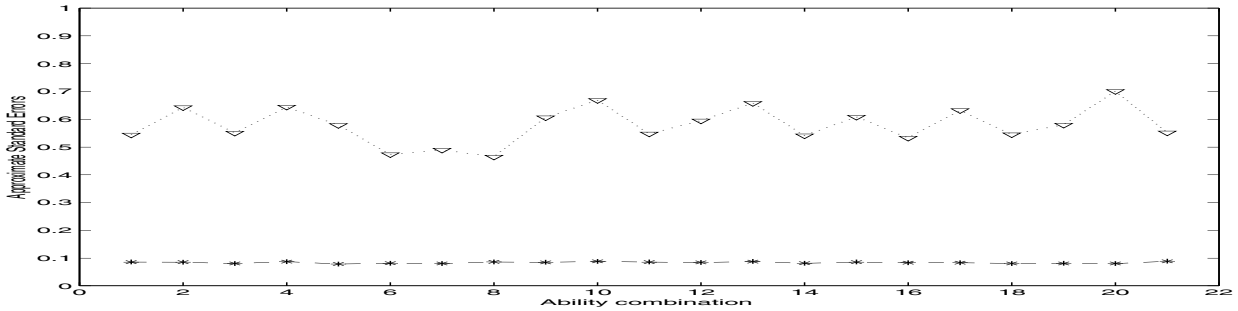


Figure 36: Actual average SEs ($\bigtriangledown$) vs. test information-based SEs (*) for MLE at cor=0.0 at test length of 40 items under Bank I



References

[1] Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement,* 15, 13-24.

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

[3] Cohen, J. (1988). *Statistical power analysis for the behavior sciences (2nd ed.).* Hilldale, NJ: Lawrence Erlbaum Associates.

[4] De Ayala, R. J. (1989). A comparison of the nominal response model and the three-parameter logistic model in computerized adaptive testing. *Educational and Psvcholoeical Measurement,* 49,789-805.

[5] Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement,* 20, 101-125.

[6] Ho, R., & Hsu, T. C. (1989). *A comparison of three adaptive testing strategies using MicroCAT.* Paper presented at the annual meeting of the American Educational research Association, San Francisco.

[7] Kendall, M.G. & Stuart, A. (1977). *The Advanced Theory of Statistics*, Griffin, London, U.K..

[8] Lord, F. M. (1977). A broad-range tailored test of verbal ability . *Applied Psychological Measurement,* 1, 95-100.

[9] Lord, F. M. (1980). *Application of item response theory to practical testing problems.* Hillsdale, NJ:Erlbaum.

[10] Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, 48, 233-245.

[11] Lord, F. M. (1984). *Maximum likelihood and Bayesian parameter estimation in item response theory* (Research Rep. No. RR-84-30-ONR). Princeton, NJ: Educational Testing Service.

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

[13] Luenberger, D. G. (1984).*Linear and Nonlinear Programming* . Reading, MA: Addison-Wesley.

[14] McBride, J. R. , & Martin, J. T. (1983) . Reliability and validity of adaptive ability tests in a military setting. In Weiss, D. J. (Ed.), *New horizons in testing.* New York: Academic Press.

[15] McKinley, R. L., & Reckase, M. D. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space.* (Research Rep. No. ONR83-2). Iowa City, Iowa: American College Testing, Program resident Programs Department.

[16] Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association,* 70, 351-356.

[17] Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401-412.

[18] Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement,* 25, 193-203.

[19] Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one ability. *Applied Psychological Measurement*, 15, 361-373.

[20] Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometricka*, 61, 331-354.

[21] Stone, C. A. (1993). *The use of multiple replications in IRT based Monte Carlo research.* Paper presented at European Meeting of the Psychometric Society , Barcelon.

[22] Urry, V. W. (1977). Tailor testing: A successful application of item response theory. *Journal of Educational Measurement,* 14, 181-196.

[23] van der Linden, W. J. (1997). *Multidimensional adaptive testing with a minimum error-variance criterion*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

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

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

[27] Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement,* 6, 473-492.

[28] Weiss, D. J., & Kingsley, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement,* 2l, 361-375.

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

[30] Zhang, J., & Stout, W. (1999). The theoretical detect of dimensionality and its application to approximate simple structure. *Pschometrika,* 64, 213-249.