Precision of Warm's Weighted Likelihood Estimation of Ability for a Polytomous Model in CAT

Shudong Wang
University of Pittsburgh

Tianyou Wang
Measurement and Research Department
ACT

Introduction

Currently, CAT is of considerable interest to the measurement and research community because of its advantages over the traditional paper-pencil tests (Lord, 1977; Kingsbury & Weiss, 1983; McBride & Martin, 1983; Urry, 1977; Wainer et al., 1990). A major feature of CAT is its ability to select a unique set of items from an existing item bank to match the current estimate of the ability level of an examinee. A equal or greater measurement accuracy can be achieved with fewer items than a paper-pencil test. However, the advantages of CAT cannot be fully realized without the application of item response theory (IRT). IRT is a mathematical model describing the relationship between the probability of an examinee's correct response on a test item and his or her underlying ability. The estimation of ability is one of the major components in CAT systems. The accuracy of ability estimation methods used in CAT has significant impacts on the quality of CAT testing because it affects not only the final score reported, but also the item selection and test termination. The purpose of this investigation is to assess the relative accuracy of four CAT ability estimation methods: Warm's weighted likelihood estimate (WLE, 1982), maximum likelihood estimate (MLE; Lord, 1980), expected a posteriori estimate (EAP; Bock & Mislevy, 1982) and maximum a posteriori estimate (MAP; Samejima, 1969) using the generalized partial credit model (Muraki, 1992), in various conditions of CAT. Special attention has been paid to the Warm's (1989) weighted likelihood estimation of ability in polytomous IRT models because no empirical CAT study has been done on this procedure using polytomous IRT models.

In computerized adaptive testing (CAT), an examinee's ability is estimated after each item response is given. There are several ability estimation methods available, such as the maximum likelihood estimation (MLE; lord, 1980) and the Bayesian estimation

methods (OWEN, 1975; EAP, 1982; MAP, 1983). Wang (1995) provided guidelines for selecting an appropriate CAT ability estimation method in different decision contexts for three-parameter IRT model. All of these estimation methods produce estimates that are biased to some degree, and are shown to have the first-order bias $O(n^{-1})$ and higher-order bias, in other words, bias is inversely proportional to n, n^2, n^3, ... (Lord, 1983a, 1983b, 1984, Wang, 1995). In general, the asymptotic bias of the MLE $\hat{\theta}$ may be written as

$$\text{Bias}(\text{MLE}(\theta)) = \text{Bias}_1(\text{MLE}(\hat{\theta})) + \text{Bias}_2(\text{MLE}(\hat{\theta})) + \ldots \quad , \tag{1}$$

where Bias_1, Bias_2 stand for first-order and second-order bias, etc. Firth (1993) stated that there are two approaches that may reduce the MLE bias, especially reducing the first-order bias term, one is a corrective approach and the other is a preventive approach.

The corrective approach includes the two methods that have been extensively studied in the literature, one is the computationally intensive methods, such as jackknife method and bootstrap method (Quenouille, 1949,1956), the other is simply to subtract as estimate of the first-order bias $\text{Bias}_1(\text{MLE}(\hat{\theta}))$ from the MLE estimate; the bias-corrected estimate is then expressed as

$$\hat{\theta}_{\text{CorrectBias}} = \hat{\theta}_{\text{MLE}} - \text{Bias}_1(\text{MLE}(\hat{\theta})). \tag{2}$$

Both of these methods may succeed in removing term $\text{Bias}_1(\text{MLE}(\hat{\theta}))$ from the asymptotic bias (Warm, 1989). A common feature of the two methods is that they are 'corrective' in nature, that is, the MLE $\hat{\theta}$ is first calculated, and then corrected. However, both methods require the existence of a finite $\hat{\theta}$.

The preventive approach (Firth, 1993) to reduce MLE estimate bias, on the other hand, modifies the score function before the MLE estimate is calculated. In general, the MLE is derived as a solution to the score equation $S(\theta)$:

$$S(\theta) = \frac{\partial}{\partial\theta} l(\theta) = 0, \tag{3}$$

where $l(\theta)=\ln L(\theta)$ is the log likelihood function for any given model.

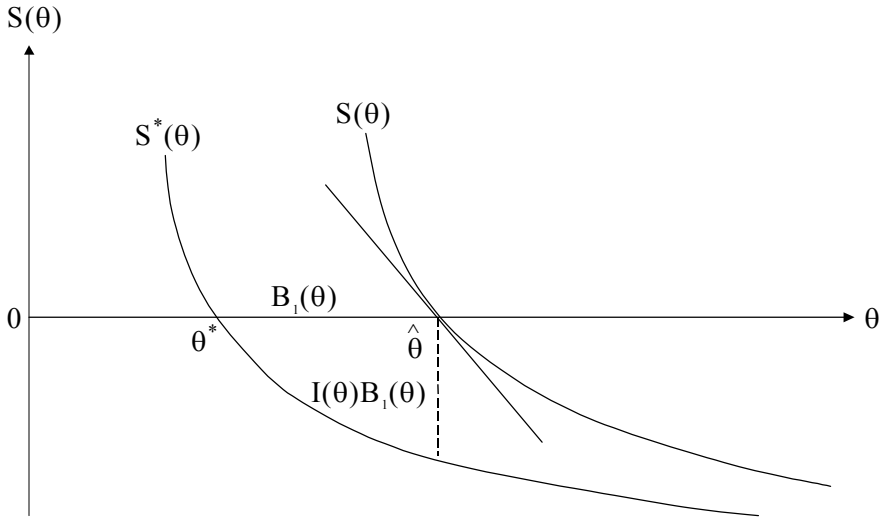

Figure 1. Modification of the Score Equation

The bias in $\hat{\theta}$ can be reduced by introducing a small bias (first-order) into the score function (Firth, 1993). For given bias B(θ), score function S(θ) can be corrected to the score function $S^*(\theta)$ by simple triangle geometry as illustrated in Figure 1. Since the Fisher Information function is defined as

$$I = I(\theta) = -E\left[\frac{\partial^2}{\partial\theta^2}\ln L(\mathbf{u}\,|\,\theta)\right] = -E\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right] = -E\left[\frac{\partial}{\partial\theta}S(\theta)\right].$$

Which is the expectation of the negative value of S(θ) tangent at $\hat{\theta}$,or $-I(\theta) = S'(\theta)$ then for any given bias B(θ), the score function can be shifted by the amount of I(θ)B(θ).

The direction of "corrected" bias depends on the sign of bias. If $\hat{\theta}$ is subjected to a positive bias B(θ), the score function is shifted downward at each point θ by an amount I(θ)B(θ); otherwise, the score function is shifted upward. This defines a modified score function

$$S^*(\theta) = S(\theta) - I(\theta)B(\theta) \qquad (4)$$

and hence a modified estimate θ^*, is given by solving $S^*(\theta) = 0$. It is can be seen that, in general, the $O(n^{-1})$ bias may be removed from the maximum likelihood estimator by introduction of an appropriate bias term into score function. It is not an assumption of this procedure that bias reduction is always desirable. The merits of bias reduction in any particular problem will depend on a number of factors (Copas, 1988).

Warm (1989) proposed a weighted likelihood estimation (WLE) method for the 3-parameter IRT model based on the relationship between the bias functions of the maximum a posteriori (MAP) and MLE methods. The WLE method not only reserves the MLE method's attractive asymptotic properties, but also overcome the MLE method's unbounded nature. The WLE method happens to be a special case of Firth (1993) preventive approach. The focus of Warm's paper is a method for reducing the bias, especially on the removing first-order term. Warm derived so-called weighted likelihood estimation (WLE) and proved that WLE has less bias than MLE with same asymptotic variance and normal distribution (Warm,1989). Warm had proved that the WLE is unbiased to order $o(n^{-1})$, that is,

$$\text{Bias}(\text{WLE}(\theta)) = 0 + o(n^{-1}).$$

Warm concluded that WLE is less biased than MLE estimation method in every condition investigated in his study. Samejima (1998) expanded Warm's procedure to Graded Response Model.

Warm's Weighted Likelihood Estimation for Dichotomously Scored Responses

For a test with n items, the MLE of θ is the value of θ that maximizes the likelihood function L(u|θ),

$$L(\mathbf{u} \mid \theta) = \prod_{j=1}^{n} P_j(\theta)^{u_j} (1 - P_j(\theta))^{1-u_j}, \tag{5}$$

where **u** is the vector of n scored item responses, u_j=1 if item j is answered correctly, and u_j=0 if item j is answered incorrectly. $P_j(\theta)$ is the probability answering item j correctly for examinee with θ ability modeled by 3PL IRT model. The weighted likelihood estimate of θ, WLE(θ)=θ^*, is defined as the value of θ, such that the weighted likelihood function:

$$w(\theta) \bullet L(\mathbf{u}|\theta), \tag{6}$$

is maximized. θ^* is the solution of

$$\sum_{j=1}^{n} \frac{(u_j - P_j)P_j'}{P_j(1-P_j)} + \frac{\partial \ln w(\theta)}{\partial \theta} = 0. \tag{7}$$

Lord (1983a) gives the asymptotic bias in the MLE(θ), which is $O(n^{-1})$:

$$\text{Bias}(\text{MLE}(\theta))=\frac{-J}{2I^2}, \tag{8}$$

where I is test information, $I=\frac{\sum_{j=1}^{n} P_j'^2(\theta)}{P_j(\theta)Q_j(\theta)}$, $Q_j(\theta)=1-P_j(\theta)$; $J=\frac{\sum_{j=1}^{n} P_j'(\theta)P_j''(\theta)}{P_j(\theta)Q_j(\theta)}$.

Lord (1984) also gives the bias of Bayesian modal estimate of θ, BME(θ), with a standard normal prior:

$$\text{Bias}(\text{BME}(\theta))=\text{Bias}(\text{MLE}(\theta))-\frac{\theta}{I}, \tag{9}$$

based on Equation 8 and 9, Warm conjectured that the bias of the estimator defined by (7) is

$$\text{Bias}(\theta^*)=\text{Bias}(\text{WLE}(\theta))=\text{Bias}(\text{MLE}(\theta))+\frac{\frac{\partial \ln w(\theta)}{\partial \theta}}{I(\theta)}. \tag{10}$$

In order to find the estimator that is unbiased, setting Equation 10 to zero and substituting Equation 8, Warm obtains

$$\frac{\partial \ln w(\theta)}{\partial \theta}=\frac{J}{2I}=-\left(\frac{-J}{2I^2}\right)*I=-\text{Bias}(\text{MLE}(\theta))*I(\theta). \tag{11}$$

An estimate satisfying Equation 7 and 11 is called a weighted likelihood estimate. As Warm pointed out, the WLE is not in any sense Bayesian, because no assumptions have been made about the distribution of θ, and w(θ) is a function of the item parameters of the test. Warm (1989) proved that Bias(WLE(θ)) is only $o(n^{-1})$ and showed WLE(θ) is asymptotically normally distributed with variance equal to the variance of MLE(θ).

Expansion of WLE for Polytomously Scored Responses

Samejima (1998) expanded the Warm's weighted likelihood estimation method to the polytomously scored responses. First, Samejima (1993a, 1993b) generalized Lord's MLE bias function in 3PL for polytomous responses, then based on the bias function for polytomous scored items, she applied Warm's WLE to the polytomous responses.

The MLE bias function for general discrete responses is given by

$$\text{Bias}(\text{MLE}(\theta;\hat{\theta}_v)) \cong E[\hat{\theta}_v - \theta \mid \theta] \cong -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k=0}^{m_j}\frac{\frac{\partial}{\partial\theta}P_{jk}(\theta)\frac{\partial^2}{\partial\theta^2}P_{jk}}{P_{jk}(\theta)}\ , \qquad (12)$$

where $P_{jk}(\theta)$ is the probability of a response in category k to item j with the assumption that $P_{jk}(\theta)$ is, at least, five times differentiable with respect to θ. When $P_{jk}(\theta)$ is modeled by the Samejima's Graded Response Model (1969), Equation 12 becomes

$$B(\theta;\hat{\theta}_v) \cong E[\hat{\theta}_v - \theta \mid \theta] \cong -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}\frac{\frac{\partial}{\partial\theta}P_{jk}(\theta)\frac{\partial^2}{\partial\theta^2}P_{jk}(\theta)}{P_{jk}(\theta)}$$

$$= -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}\left\{\frac{Da_j\exp[Da_j(\theta-b_{jk})]}{\{1+\exp[Da_j(\theta-b_{jk})]\}^2} - \frac{Da_j\exp[Da_j(\theta-b_{jk+1})]}{\{1+\exp[Da_j(\theta-b_{jk+1})]\}^2}\right\} *$$

$$\left\{\frac{D^2a_j^2\left[\frac{\exp[Da_j(\theta-b_{jk})]-\exp(2[Da_j(\theta-b_{jk})]\}}{\{1+\exp[Da_j(\theta-b_{jk})]\}^3} - \frac{\exp[Da_j(\theta-b_{jk})]-\exp(2[Da_j(\theta-b_{jk+1})]\}}{\{1+\exp[Da_j(\theta-b_{jk+1})]\}^3}\right]}{P_{jk}(\theta)}\right\} \qquad (13)$$

When $P_{jk}(\theta)$ modeled by the Muraki's Generalized Partial Credit Model, Equation 12 takes the form

$$\text{Bias}(\text{MLE}(\theta;\hat{\theta}_v))$$

$$\cong -\frac{1}{2[I(\theta)]^2}\sum_{j=1}^{n}\sum_{k_j}D^3a_j^3P_{jk}\left[k-\sum_{c=0}^{m}cP_{jk}\right]\left[k^2-2k\sum_{c=0}^{m}cP_{jk}+2\left(\sum_{c=0}^{m}cP_{jk}\right)^2-\sum_{c=0}^{m}c^2P_{jk}\right]\ . \qquad (14)$$

Based on the general discrete response, a straightforward expansion of Warm's WLE for 3PL to general discrete response can be expressed as

$$\frac{\partial}{\partial\theta}\ln L(v \mid \theta) + \frac{\partial}{\partial\theta}\ln w(\theta) = \sum_{k_j \in v}\frac{\partial}{\partial\theta}\ln P_{jk}(\theta) - \text{Bias}(\text{MLE}(\theta,\hat{\theta}))I(\theta) \equiv 0\,. \qquad (15)$$

Where $I(\theta)$ is the test information.

Method and Data

The primary goal of this design is to answer the stated research questions and to maximize generalizability and replicability of research results. Both descriptive methods

and inferential procedures will be used in this MC study. Although the descriptive methods provide global summaries of the study results, the deficiencies of descriptive methods are (a) masking of complex effects, (b) failing to provide estimates of the magnitudes of the effects, and (c) failing to systematically take into account the sampling error associated with the random generation of data (Harwell, 1997; Hoaglin and Andres, 1975; Timm, 1976). The inferential procedure, on the other hand, can overcome all of these deficiencies by conceptualizing this MC study as statistical sampling experiments (Harwell, 1997; Spence, 1983). Additional advantages of using this conceptualization are that the threats to internal and external validity can be evaluated and that the similarities and differences among the results of different studies on the same problems can be compared (Campbell & Stanley, 1963).

1 Independent Variables

(1) Ability Estimation Methods

The primary discrete independent variable examined in this study is the ability estimation methods. Four ability estimation methods for the generalized partial credit model and the graded response model are MLE, WLE, EAP, and MAP and the relationship between this variable and other independent variables (described below) are considered. For each method, the values of the true ability parameter used in this study are equally spaced across a fixed range on the θ scale. Specifically, 21 true ability values are used ranging from –4.0 to 4.0 by increments of 0.4 are used for the GPCM. CAT will be simulated for 500 examinees at each of the 21 true ability parameter points.

(2) Test Termination Rules

Two types of test termination rules to be investigated for the GPCM and the GRM are:

(a) Fixed test length: The CAT test will be terminated after certain number of items has been administered. Four test lengths will be used in this study are 5, 10, 15, and 20 items. Previous studies (Koch & Dodd, 1985; DeAyala, 1992; Dodd et al., 1989) employed the fixed test lengths from 10 to 30 items. To search for the smallest acceptable test length, a test length of 5 items is also included in this investigation.

(b) Fixed test reliability: The testing will be terminated when certain values of estimated reliability are reached.

Since the relationship between reliability (ρ) and the standard error of ability estimation (SE) can be expressed as $\rho = 1 - SE^2$ given the θ variance is 1, estimated reliability and estimated SE have the same effect on the CAT termination. Both MLE and WLE use the square root of the reciprocal of test information as SE. Both EAP and MAP Bayesian methods use the standard deviation of the posterior distribution as SE. The three values of reliability to be employed are 0.7, 0.8, and 0.9, which correspond to SE of 0.55, 0.45, and 0.32 and test information values of 3.33, 5, and 10. A maximum test length of 33 items will be used to terminate the test for a given exaimee if the prespecified levels of reliability cannot be reached. This maximum test length is used because the smallest bank size is 33 items.

(3) Sizes of Item Bank

Although some researches has found that item banks with 30 items may be sufficient for accurate θ estimation with few nonconvergence problems (Dodd, 1987, 1993; Dodd & DeAyala, 1994; Dodd et al., 1989; Koch & Dodd, 1989). However, these findings do not imply that any item bank composed of 30 or more items will be sufficent for polytomous CAT (Dodd et al., 1995).

The real-life item bank 1GP, which consisted of 263 polytomous scored 1996 NEAP' science items that have 3 to 5 categories for grade 4, 8, and 12, will be used for this study. The item parameters for the generalized partial credit model were calibrated by the Educational Testing Service. There are three sizes of item banks (1GP, 2GP, and 3GP) for the GPCM. Table 1 describes the sizes of the three item banks and the number of different category items in each bank. The 66 items in bank named 2GP and 33 items in the bank named 3GP are randomly drawn from bank named 1GP using the proportional stratified random sampling method (Gall et al., 1996), based on which the same proportional items with different numbers of category are randomly drawn from the 1GP bank. Similarly, the 66 items in the bank 2GR and 33 items in the bank 3GR are randomly drawn from the 1GR bank using the same technique.

The proportions (BN/BS) in Table 1 represent the percentage of different category items in the item bank in which the items are sampled. All three banks have the

Table 1

The Sizes of the Item Banks and the Number of Different Category Items in Each Bank.

IRT Model	Bank Name	Bank Size (BS)	3 Categories Item Number (BN)	3 Categories Item BN/BS	4 Categories Item Number (BN)	4 Categories Item BN/BS	5 Categories Item Number (BN)	5 Categories Item BN/BS
GPCM	1GP	263	208	0.79	47	0.18	8	0.03
	2GP	66	52	0.79	12	0.18	2	0.03
	3GP	33	26	0.79	6	0.18	1	0.03

Table 2

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

Name/ Size	Parameter	Mean	Median	S.D.	Min.	Max.
1GP (263)	a	0.549	0.522	0.229	0.105	1.871
	b_1	0.713	0.720	2.011	-6.972	11.746
	b_2	1.270	1.264	2.640	-17.381	13.926
	b_3	1.034	1.004	2.371	-6.369	7.187
	b_4	0.822	0.822	2.546	-3.159	4.924
2GP (66)	a	0.539	0.527	0.171	0.171	1.200
	b_1	1.066	1.000	1.728	-3.204	7.399
	b_2	1.679	1.491	2.519	-2.665	13.926
	b_3	1.832	1.412	1.656	-0.856	5.506
	b_4	4.270	4.270	0.535	0.535	4.925
3GP (33)	a	0.560	0.523	0.190	1.90	1.055
	b_1	0.752	0.631	1.384	-2.738	3.437
	b_2	1.695	1.684	2.495	-3.638	7.293
	b_3	1.467	1.680	3.480	-6.369	7.187
	b_4	2.000	2.000	0.000	2.000	2.000

same proportions of items with different numbers of categories. Table 2 shows the summary of descriptive statistics for the item parameters of these three item banks.

2 Dependent Variables

There are a variety of statistics that can be used to evaluate the accuracy of ability estimation in CAT. Based on the consideration of consistence with previous CAT ability estimation studies, five criteria variables (or their log transformation) were used in this study: the biases, standard errors (SEs), root mean square errors (RMSEs), fidelity (correlation of the estimated and true parameters, Wang & Vispoel, 1998), and administrative efficiency (the mean numbers of items needed to reach a criterion SE level). These criterions are used to examine the effects of the manipulated independent variables described in the last subsection because they can provide complementary evidence. The bias in IRT ability estimate can cause several problems: (a) difficulty to maintain comparability of CAT and paper-and-pencil versions of a test (Eignor & Schaeffer, 1995; Segall, 1995; Segall & Carter 1995; Wang & Kolen, 1997), (b) problematic in the test with a domain-referenced cut-score (Wang et al., 1998). In general, bias has little effect in the situations where only the relative orders of ability estimates are important. In this kind of the situation, SE or RMSE may play an important rule. The bias, SE, and RMSE can be assessed both conditionally or overall (average) across an entire ability distribution. The conditional indices are computed at each θ point, and overall indices are computed by taking the absolute values of the conditional indices and integrating them over a normally distributed θ for a population of examinees using the numerical integration method (Wang & Vispoel, 1988). These formulas are:

Conditional indexes:

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

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

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\frac{1}{N}\sum_{r=1}^{N}(\hat{\theta}_r - \theta)^2}\,, \qquad (3)$$

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

The RMSE can be separated into two components, Bias and SE ($\mathrm{RMSE}^2 = \mathrm{Bias}^2 + \mathrm{SE}^2$). With respect to θ, this can be expressed as

$$\frac{\sum_{r=1}^{N}(\hat{\theta}_j - \theta)^2}{N} = \sum_{r=1}^{N}(\hat{\theta}_j - \theta)^2 + \frac{1}{N}\sum_{r=1}^{N}\left(\hat{\theta}_r - \frac{\sum_{t=1}^{N}\hat{\theta}_t}{N}\right)^2. \qquad (4)$$

The Bias used in this study is averaged across the replications (or examinees),

$$\mathrm{Bias}(\hat{\theta}) = \frac{1}{N}\sum_{r=1}^{N}(\hat{\theta}_r - \theta)\,. \qquad (5)$$

Overall indexes:

$$\mathrm{AVERAGE}_{\mathrm{Bias}} = \sum_{i=1}^{21}|\,\mathrm{Bias}(\hat{\theta})\,|\,\big|\theta_i * \mathrm{weight}(\theta_i), \qquad (6)$$

$$\mathrm{AVERAGE}_{\mathrm{SE}} = \sqrt{\frac{\sum_{i=1}^{21}\mathrm{SE}^2(\hat{\theta})\big|\theta_i * \mathrm{weight}(\theta_i)}{21}}, \qquad (7)$$

$$\mathrm{AVERAGE}_{\mathrm{RMSE}} = \sum_{i=1}^{21}\mathrm{RMSE}(\hat{\theta})\big|\theta_i * \mathrm{weight}(\theta_i), \qquad (8)$$

where the weight(θ_i) are quadrature weights based on the standard normal distribution, and the θ_i are the 21 equally spaced true ability levels that ranged from –4 to 4 in an increment of 0.4. On the basis of the assumption, $\hat{\theta} = c\theta + e$, where c is a constant

depending on bias, and e is the random error (If no bias, c=1). Wang (1995) defined the fidelity correlation as

$$r_{\theta\hat{\theta}} = \frac{\mathrm{cov}(\theta,\hat{\theta})}{\sigma_\theta \sigma_{\hat{\theta}}} = \frac{c\sigma_\theta^2}{\sigma_\theta \sigma_{\hat{\theta}}} = \frac{c\sigma_\theta}{\sigma_{\hat{\theta}}} = \frac{c\sigma_\theta}{\sqrt{c^2\sigma_\theta^2 + \sigma_e^2}} = \frac{\sigma_\theta}{\sqrt{\sigma_\theta^2 + \frac{\sigma_e^2}{c^2}}}. \qquad (3.9)$$

If there is no linear relationship between bias and θ, then this equation will provide only an approximation to the fidelity index. Not all dependent variables defined here will be used in the ANOVA. Descriptive statistics will be provided for all conditional indices and overall indices, and inferential statistics will be used only for the overall indices.

3 Experimental Design

In addition to obtain the descriptive statistics, for the overall indices, one 4 x 3 x 4 and one 4 x 3 x 3 completely crossed design ANOVA (Table 3) will be used to investigate the effects of the four ability estimation methods by

Table 3

Experimental Design A and B for the Fixed Test Length (Items) and Fixed Test Reliability Termination Rules

Design	Estimation Methods	Bank Sizes	Termination Rules: Fixed Test Length (Items)
Design A	MLE	263, 66, 33	5, 10, 15, 20
	WLE	263, 66, 33	5, 10, 15, 20
	EAP	263, 66, 33	5, 10, 15, 20
	MAP	263, 66, 33	5, 10, 15, 20
Design B	MLE	263, 66, 33	0.7, 0.8, 0.9
	WLE	263, 66, 33	0.7, 0.8, 0.9
	EAP	263, 66, 33	0.7, 0.8, 0.9
	MAP	263, 66, 33	0.7, 0.8, 0.9

A. The four test termination rules (fixed test length), the three sizes of item, and the four ability estimation methods will be used in design A.

B. The three test termination rules (fixed reliability), the three sizes of item, and the four ability estimation methods will be used in design B.

The four nominal levels of the ability estimation methods manipulated are MLE, WLE, EAP, and MAP for both designs. The three levels of item bank sizes manipulated are 261, 66, and 33 for both designs. The four levels of the test termination rules manipulated are 5,10,15, and 20 using fixed length for design A. The three levels of the test termination rules manipulated are 0.7, 0.8, and 0.9 using fixed reliability for design B.

4 Computerized Adaptive Testing Simulation Procedures

The C program CATMASTER will be used to carry out this CAT simulation study. The known item parameters for the GPCM will be used throughout the simulation process. The procedure for polytomous IRT models CAT related to this simulation study include the following: First, conditioned on each of the 21 equally spaced true ability θ levels, 500 simulees are assigned, the range of 21 true θ values is from -4 to 4 in an increment of 0.4. Second, for each simulee, a CAT is simulated which takes the following steps:

Step 1. To start the test, an initial ability estimate of 0.0 is assumed. The *maximum information* item selection algorithm is used to select all the items, including the first one.

Step 2. After an item is selected, a response is generated based on the simulee's true θ and item parameter estimates. To generate a response, the probabilities $P_{jk}(\theta_i)$ of obtaining each of the k response categories are computed using either the GPCM or GRM model. Then the cumulative probabilities of getting response category k or higher are computed and compared to a random number between 0 and 1 generated from a uniform distribution. If the random number falls between the cumulative probabilities of k-1 and k category, a response score k is assigned for this simulee.

Step 3. After a response is generated, the provisional ability level is estimated using one of four ability estimation methods (MLE, WLE, EAP, MAP). Based on this provisional estimate, the next item is selected using the maximum information procedure and one of the stopping rules is checked using this information.

Step 4. Step 2 and step 3 are repeated until a termination criterion is researched. If the testing is terminated, the final ability estimate and error variance estimates are recorded.

For all CAT simulations in this study, the provisional estimate after the first item is based on the EAP with a normal prior.

5 Analyses of Results

Both descriptive statistics and inferential statistics will be used to analyze the results. The descriptive statistics, such as tabular summaries, and graphical presentations, will be used to present the conditional and overall indices. The results based on the conditional and overall indices are presented in following ways:

(1) The conditional indices Bias, SE, and RMSE of the four ability estimation methods (MLE, WLE, EAP, and MAP) for different simulated conditions will be plotted against θ to investigate how different the bias, SE, and RMSE are by using different ability estimation methods.

(2) The mean and SD of the number items required for fixed reliabilities of four ability estimation methods will be plotted against θ to see how efficient of each ability estimation method is at different ability levels.

(3) The overall indices of bias, SE, and RMSE will be tabulated for different CAT simulation conditions.

At the same time, the inferential analyses will be carried out. Because the independent variables of ability estimation method and test termination rule are nominal factors rather than metric factors, the ANOVA is preferred over a regression analysis for this study. Several separate ANOVAs for each of the overall dependent variables are run to detect the effects of independent variables.

Results

1. Results Based on the Conditional Indices

Fixed test length

Figure 2 through 12 show the bias, SE, and RMSE of four ability estimation methods (WLE, MLE, EAP, and MAP) with different ability levels for four fixed test lengths based on bank size 263. It is very clear from figures 2 and 5 that the WLE has the smallest bias over entire range of ability among all four estimation methods for all test

---

Insert Figures 2 to 5 about here

---

lengths, this result also matches the dichotomous model case (Wang, Hanson, & Lau, 1998). Both WLE and MLE have considerably less bias than two Bayesian methods, it can be seen that the MLE has "outward bias", which means the bias of MLE is positively correlated with θ; while WLE almost has no bias. The biases of EAP and MAP have "inward bias", which means the bias is negatively correlated with θ.

From Figures 6 and 9, we can see that the WLE has less SE than MLE at almost all ability levels for both fixed test lengths, which means WLE not only reduce MLE's bias, but also reduce MLE's SE. It can be also seen that for both ability extremes, the Bayesian methods of EAP and MLE have far less variability than WLE and MLE, this reduction of variability is at the expense of increased bias. This finding also confirms previous finding that Bayesian methods had lower SE but higher bias when compared with MLE and WLE for dichotomous models (Wang, Hanson, & Lau, 1998; Warm, 1989).

Insert Figures 6 to 9 about here

The figures 10 and 13 show that for all test lengths, the RMSE for all four estimate methods is similar along all ability levels, and WLE had less RMSEs than MLE. All bias, SE, and RMSE decrease as the test length increases.

Insert Figures 10 to 13 about here

Fixed reliability

Figures 14 to 16 show the effects of fixing reliability at 0.7, 0.8, and 0.9 on the biases of the four ability estimation methods based on item bank 263. The fixed reliability changed the bias direction of MLE from "outward" to the "inward". The WLE

further increased that "inward" bias, which is opposite with the situation of fixed test length. This means under fixed reliability rule, WLE fails to reduce the bias of MLE. Both MLE and WLE have remarkable smaller bias than Bayesian methods. As indicated in Figures 17 to 19, both WLE and MLE have larger SE than Bayesian methods and WLE has less SE than MLE. Figures 20 through 22 show that non-Bayesian methods have lower RMSE than Bayesian methods at extreme ability levels, but have higher RMSE than Bayesian methods at middle level of ability. WLE perform slightly well than MLE. All bias, SE, and RMSE decrease as the test reliability increase.

Insert Figures 14 to 22 about here

The Figure 23 to 25 shows the mean numbers of items required for achieved the same reliability 0.7, 0.8, and 0.9 by using different ability estimation methods. For average ability level examinees, the number of items required for MLE and WLE are approximately the double of those for EAP and MAP; for extreme ability level examinees, the numbers of items required for MLE and WLE are tripled.
The sizes of item bank have slight effects on the all conditional indexes.

Insert Figures 23 to 25 about here

---

1. Results Based on the Overall Indexes

Table 4 though 6 show the results of three-way ANOVA of absolute bias, average SE, and average RMSE for fixed test length termination rule and Table 7 to 9 show the results of three-way ANOVA of absolute bias, average SE, and average RMSE for fixed test length reliability rule.

In general, the results for overall indexes match well with the results of conditional indexes. The factor of ability estimation methods has the most influence on absolute bias because it accounted for 54.0% total variance of absolute bias for fixed test length termination rule and 74.0% total variance of absolute bias for fixed reliability termination rule across the three sizes of item banks. The test termination rule also play a important part of estimated absolute bias, it take 15.4% total variance for fixed test length and

Table 4
Results of ANOVA of Absolute Bias for Fixed Test Length Termination Rules

Source	df	F	*p*	η^2
Main Effects				
M (Method)	3	1039.952	0.000	0.540
S (Size)	2	59.298	0.000	0.021
L (Length)	3	613.137	0.000	0.318
Interaction Effects				
M x S	6	37.496	0.000	0.039
M x L	9	38.191	0.000	0.059
S x L	6	2.336	0.046	0.002
M x S x L	18	4.076	0.000	0.013
Error	48			

Table 5
Results of ANOVA of Average SE for Fixed Test Length Termination Rules

Source	df	F	p	η^2
Main Effects				
M (Method)	3	122.258	0.000	0.167
S (Size)	2	66.079	0.000	0.060
L (Length)	3	483.836	0.000	0.659
Interaction Effects				
M x S	6	0.605	0.725	0.002
M x L	9	18.696	0.000	0.076
S x L	6	1.904	0.099	0.005
M x S x L	18	1.059	0.419	0.009
Error	48			

Table 6
Results of ANOVA of Average ln(RMSE) for Fixed Test Length Termination Rules

Source	df	F	p	η^2
Main Effects				
M (Method)	3	3.361	0.026	0.052
S (Size)	2	1.473	0.238	0.015
L (Length)	3	20.135	0.000	0.312
Interaction Effects				
M x S	6	1.731	0.134	0.054
M x L	9	1.559	0.155	0.072
S x L	6	2.650	0.020	0.082
M x S x L	18	1.733	0.066	0.161
Error	48			

Table 7
Results of ANOVA of Absolute Bias for Fixed Test Reliability Termination Rules

Source	df	F	p	η^2
Main Effects				
M (Method)	3	5774.783	0.000	0.740
S (Size)	2	7.463	0.002	0.001
R (Reliability)	2	1805.544	0.000	0.154
Interaction Effects				
M x S	6	33.933	0.000	0.009
M x R	6	341.405	0.000	0.087
S x R	4	13.018	0.000	0.002
M x S x R	12	11.723	0.000	0.006
Error	36			

Table 8
Results of ANOVA of Average SE for Fixed Test Reliability Termination Rules

Source	df	F	p	η^2
Main Effects				
M (Method)	3	60.737	0.000	0.021
S (Size)	2	590.280	0.000	0.139
R (Reliability)	2	2518.213	0.000	0.591
Interaction Effects				
M x S	6	29.867	0.000	0.021
M x R	6	37.894	0.000	0.027
S x R	4	304.547	0.000	0.143
M x S x R	12	37.996	0.000	0.054
Error	36			

Table 9
Results of ANOVA of Average ln(RMSE) for Fixed Test Reliability Termination Rules

Source	df	F	p	η^2
Main Effects				
M (Method)	3	124.574	0.000	0.062
S (Size)	2	8.993	0.000	0.003
R (Reliability)	2	2632.117	0.000	0.876
Interaction Effects				
M x S	6	6.910	0.001	0.007
M x R	6	9.523	0.000	0.010
S x R	4	7.438	0.000	0.005
M x S x R	12	3.564	0.002	0.007
Error	36			

15.4% total variance for fixed test reliability test. Both WLE and MLE show significant less absolute bias than Bayesian methods and WLE perform best. On the other hand, the SE was affected mostly by the factor of test termination rule, the fixed test length take 65.9% of total variance in average SE and the fixed test reliability take 59.1% total variance in average SE. The second most influence factor for SE is different for different termination rule. For fixed test length termination rule, the ability estimation method accounted for 16.7% variance in SE; and for fixed test reliability termination rule, the size of item bank accounted for 13.9% variance in SE. SE from both WLE and MLE are significant larger than the SE of Bayesian methods, WLE has significant less amount of SE than MLE and EAP has significant less amount of SE than MAP. RMSE (log of RMSE) was affected mostly by the factor of test termination rule. The factor of fixed test length accounted for 31.2% total variance of RMSE and three-way interaction of M x S x L takes 16.1% total variance of RMSE. The factor of fixed test reliability accounted for 87.6% total variance of RMSE and second largest factor of ability estimation takes 6.3% total variance of RMSE. There are no significant differences among different ability estimation methods except the difference between WLE and MLE.

Conclusion

The precision of ability estimation method used in CAT has significant impact on the quality of CAT testing, because it not only affects the final score reported, but also affects which items are selected for particular examinee. The bias index is one of the most important measures of the precision of ability estimation in CAT. In general, for all four ability estimation methods (WLE, MLE, EAP, and MAP), both conditional and overall indexes (bias, SE, RMSE, fidelity) of dependent variables decrease as the values of independent variables (test length, test reliability, and item bank size) increase. The magnitudes of the differences among those dependent variables decrease as the values of independent variables increase.

WLE is superior to MEL in terms of all those dependent variables and WLE perform better than Bayesian methods in terms of bias. MLE has less bias than both Bayesian methods. Both EAP and MAP show more favorable results of SE and fidelity than the results of both WLE and MLE; EAP did better job than MAP for almost all

conditions. Different test termination rules have significant impact on those dependent variables for given ability estimation methods, especially for WLE and MLE methods. Although the quality of item banks has vast effects on the conditional distribution of bias, SE, RMSE, and test efficiency, the factor of size of item bank has less impact on the differences among those dependent variables than the factor of test termination rules. This study confirms the Warm's conclusions of that (a) WLE biased to $o(n^{-1})$, while MLE, EAP, and MAP are biased to $O(n^{-1})$, (b) WLE method has small variance over entire range of θ for fixed test length CAT testing.

References

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

Bock, R. D. (1983). The discrete Bayesian. In H. Wainer & S. Messick (Eds*.), Principles of modern psychological measurement. A festschrift for Frederick M. Lord* (pp. 103-115). NJ: Lawrence Erlbaum.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-459.

Copas, J. B. (1988). Binary regression models for contaminated data. *J. R. Statist. Soc. B 50*, 225-65.

De Ayala, R. J. (1992). The nominal response model in computerized adaptive testing. *Applied Psychological Measurement*, *16*, 327-343.

Dodd, B. G., & De Ayala, R. J. (1994). Item information as a function of threshold values in the rating scale model. In M. Wilson(Ed*.), Objective measurement: Theory into practice* (Vol. 2; pp. 301-317). Norwood NJ: Ablex.

Dodd, B. G., Kock, W. R., & De Ayala, R. J. (1989). Operational characteristics of adaptive testing procedures using the Graded Response Model. . *Applied Psychological Measurement, 13*, 129-143.

Eignor, D. R., & Schaeffer, G. A. (April, 1995). *Comparability studies for the GRE General CAT and the NCLEX using CAT*. Paper presented at the meeting of the National Council on Measurement in Education, San Francisco.

Firth, D. (1993). Bias reduction of maximum likelihood estimates [Correction: 95V82 p667]. *Biometrika, 80*, 27-38.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational Research: An introduction*. New York: Longman

Harwell, M. R. (1997). Analyzing the results of Monte Carlo studies in item response theory. *Educational and Psychological Measurement, 57*, 266-279.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *American Statistician, 29*, 122-126.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed*.), New*

horizons in testing: Latent trait test theory and computerized adaptive testing (pp. 257-238). New York: Academic Press.

Kock, W. R., & Dodd, B. G. (April, 1985). *Computerized adaptive attitude measurement*. Paper presented at the annual meeting of the American Educational Research Association, Chiago.

Kock, W. R., & Dodd, B. G. (1989). An investigation of procedures for computerized adaptive testing using partial credit scoring. *Applied Measurement in Education*, 2, 335-357.

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

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

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

Lord, F. M. (1983b). *Memorandum for: Ms. Stocking, Ms. M. Wang, Ms. Wingersky. Subject: Sampling variance and bias for MLE and Bayesian estimation of θ.* August 26, 1983 (Internal Memorandum). Princeton, NJ: Educational Testing Service.

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.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed*.), New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 224-236). New York: Academic Press.

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

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.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *J. R. Statist. Soc. B 11*, 68-84.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika 43*, 353-360.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34*, (4, Pt. 2).

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item response are discrete. *Psychometrika, 58*, 119-138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika, 58*, 195-209.

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

Segall, D. O. (1995). Equating the CAT-ASVAB: *Experiences and lessons learned*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement, 7*, 405-425.

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

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterey CA: Brooks/Cole.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., and Thissen, D. (1990). *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, T. (1995). *The precision of ability estimation methods in computerized adaptive testing*. Unpublished doctoral dissertation. Iowa City, IA: University of Iowa.

Wang, T., Hanson, B. A., & Lau, C. M. (1998). Reducing bias in CAT ability estimation: A comparison of approaches. *Applied Psychological Measurement, to appear*.

Wang, T., & Kolen, M. J. (1997). *Evaluating comparability in computerized adaptive testing: a theoretical framework with an example*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

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

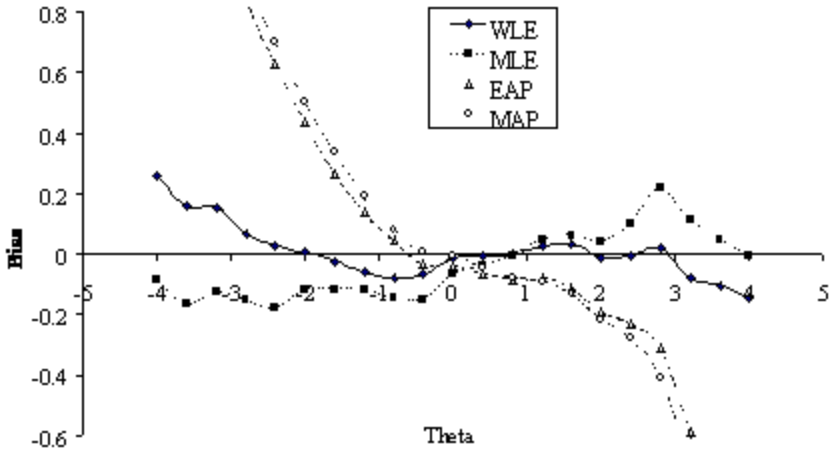

Figure 2. Bias comparison of the ability estimation methods, test length= 5, bank size=263.

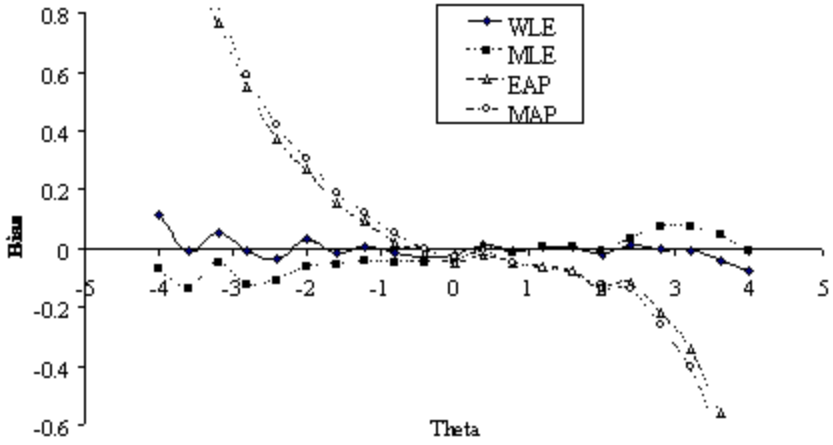

Figure 3. Bias comparison of the ability estimation methods, test length= 10, bank size=263.

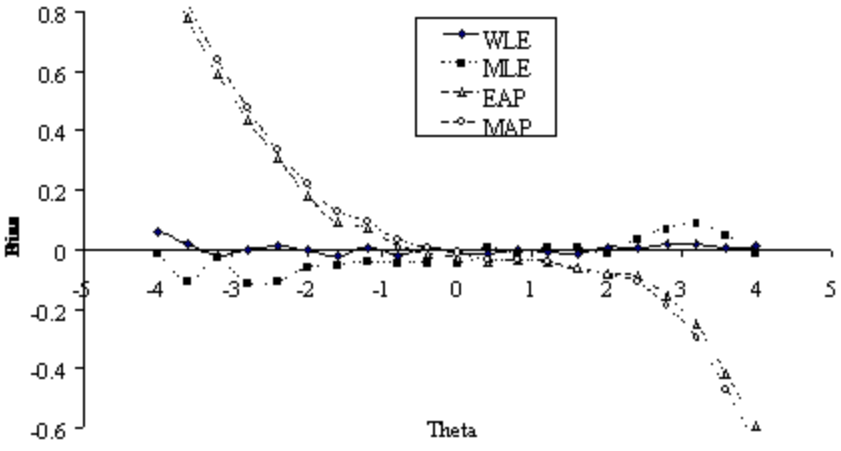

Figure 4. Bias comparison of the ability estimation methods, test length= 15, bank size=263.

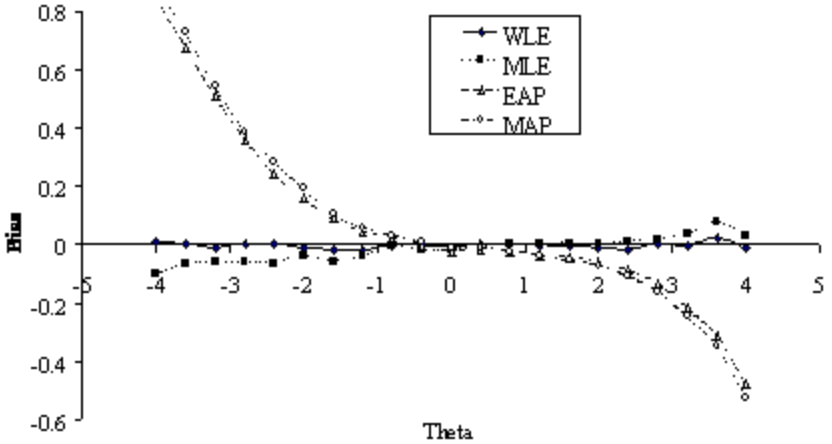

Figure 5. Bias comparison of the ability estimation methods, test length=20 , bank size=263.

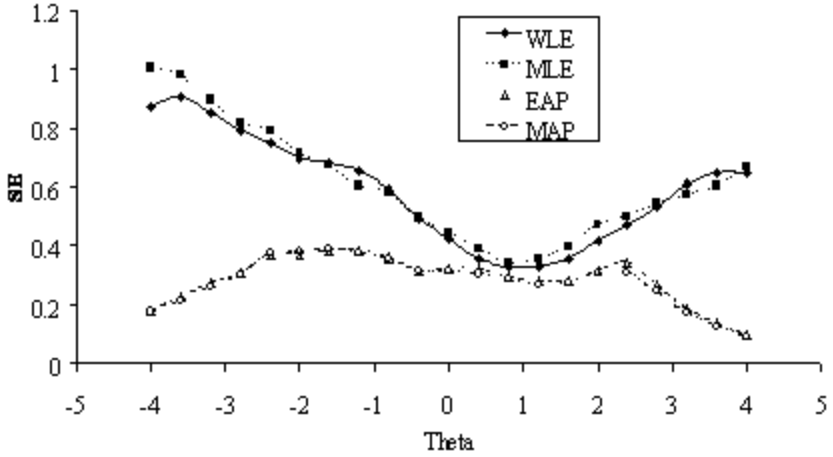

Figure 6. SE comparison of the ability estimation methods, test length=5, bank size=263.

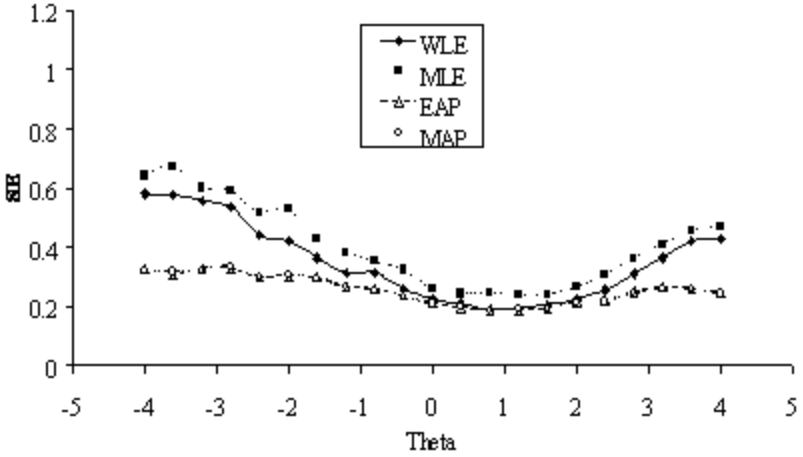

Figure 8. SE comparison of the ability estimation methods, test length=15, bank size=263.

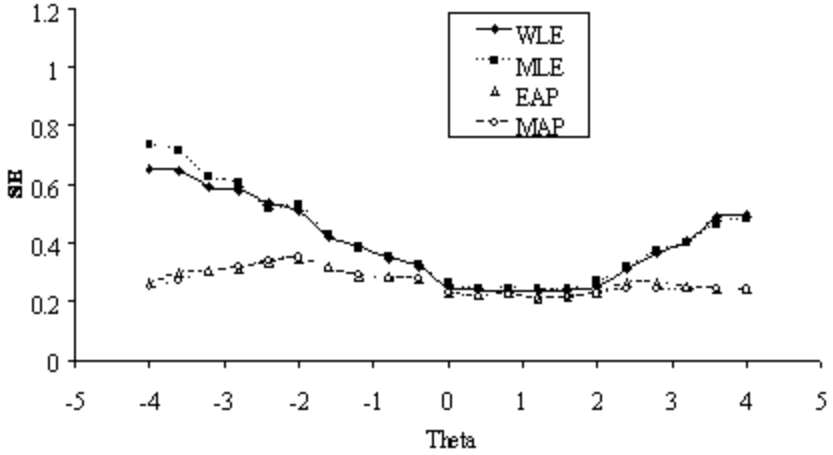

Figure 7. SE comparison of the ability estimation methods, test length=10, bank size=263.

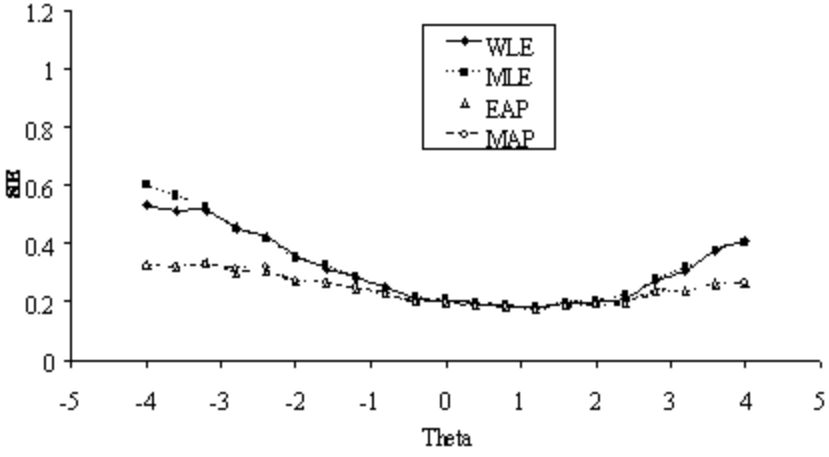

Figure 9. SE comparison of the ability estimation methods, test length=20, bank size=263.

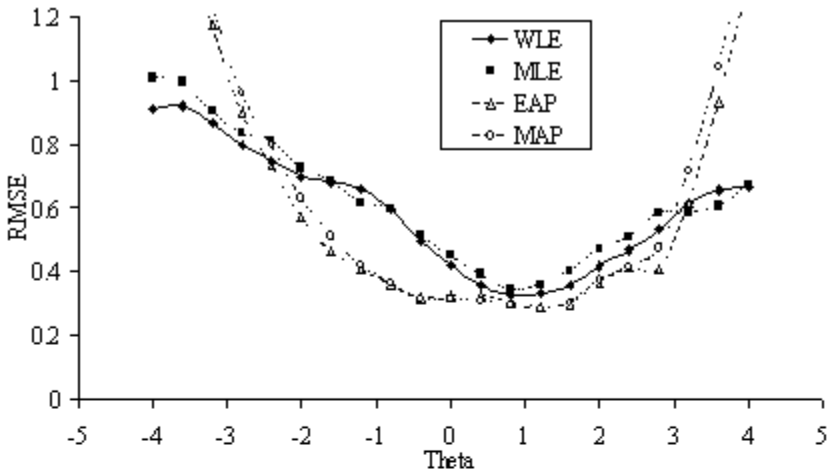

Figure 10. RMSE comparison of the ability estimation methods, test length=5, bank size=263.

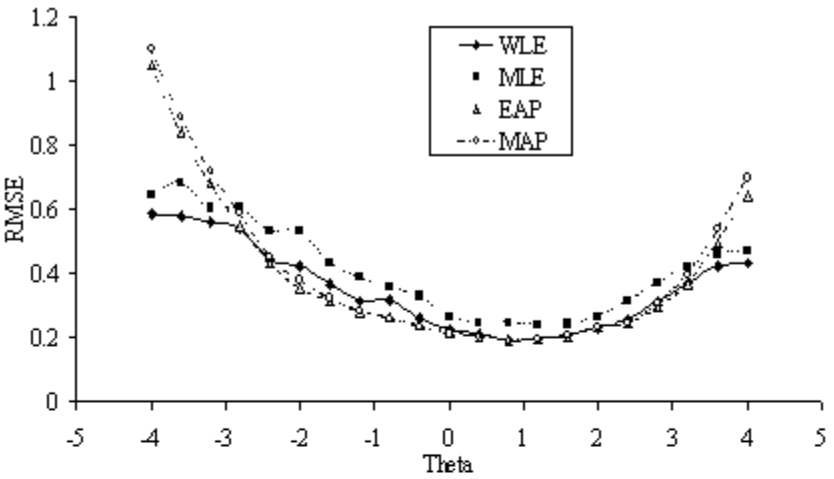

Figure 12. RMSE comparison of the ability estimation methods, test length=15, bank size=263.

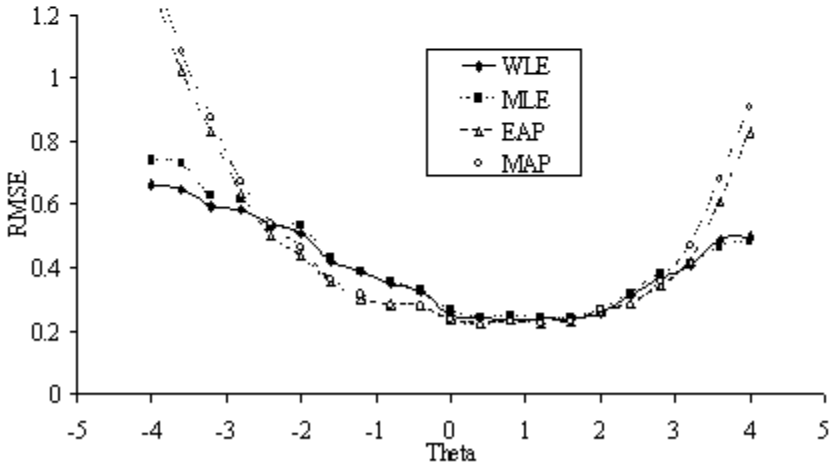

Figure 11. RMSE comparison of the ability estimation methods, test length=10, bank size=263.

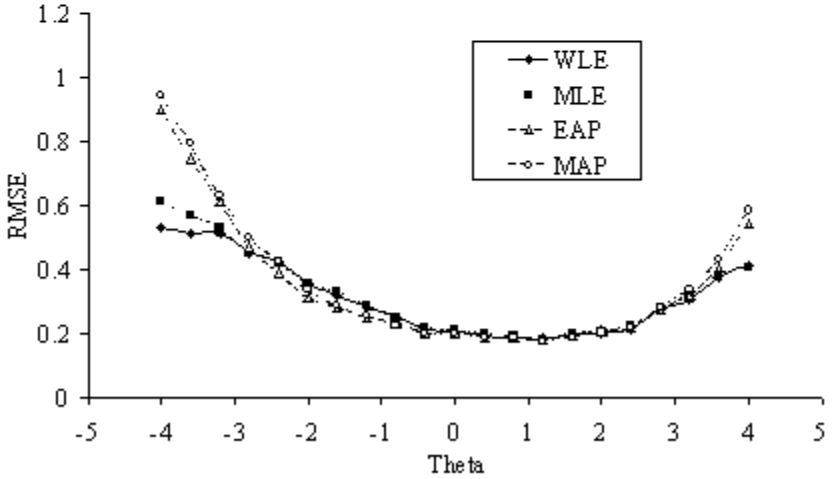

Figure 13. RMSE comparison of the ability estimation methods, test length=20, bank size=263.

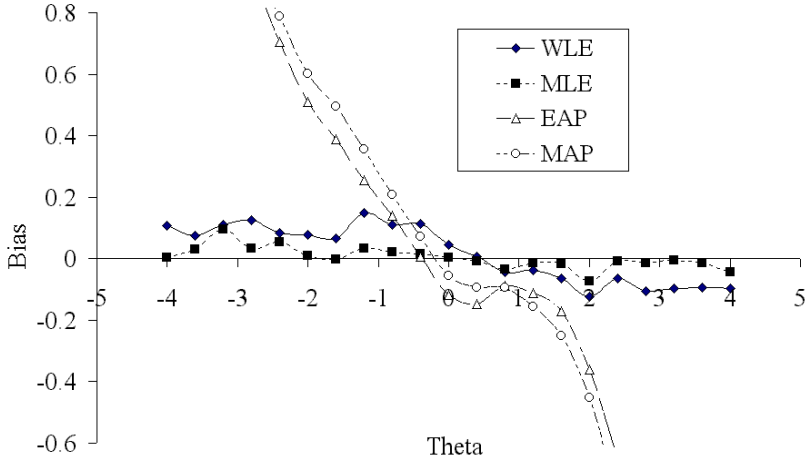

Figure 14. Bias curves of the ability estimation methods, reliability=0.7, bank size=263.

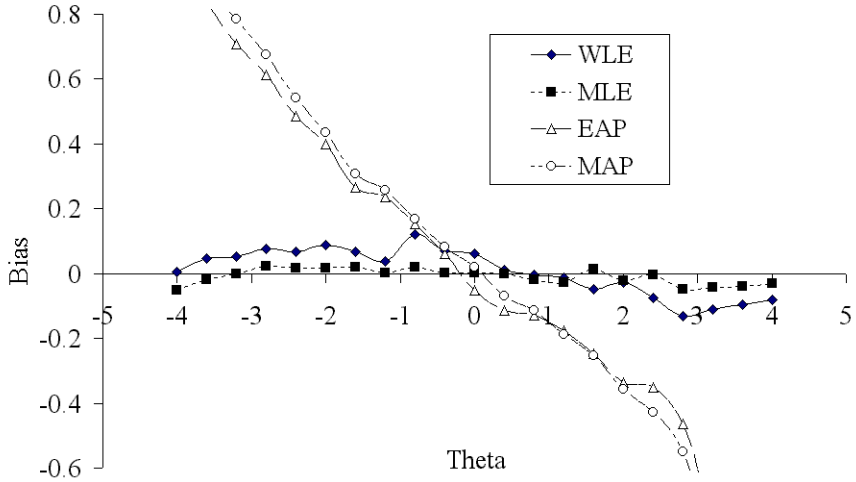

Figure 15. Bias curves of the ability estimation methods, reliability=0.8, bank size=263.

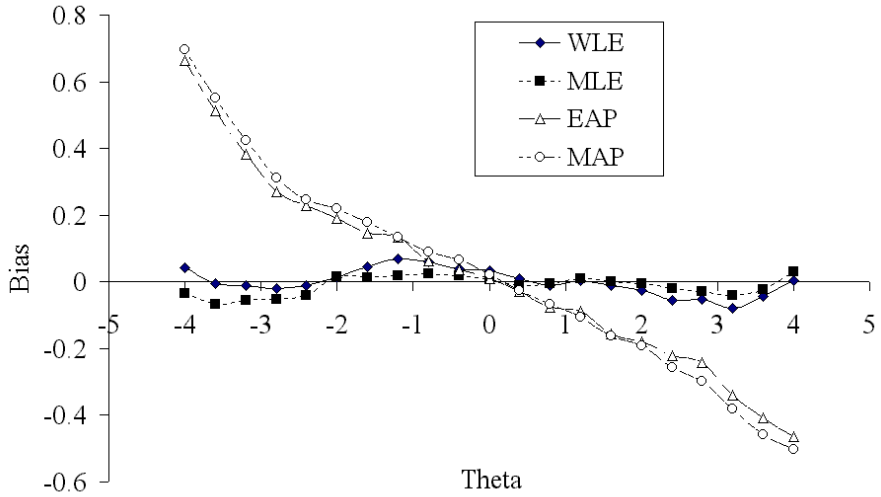

Figure 16. Bias curves of the ability estimation methods, reliability=0.9, bank size=263.

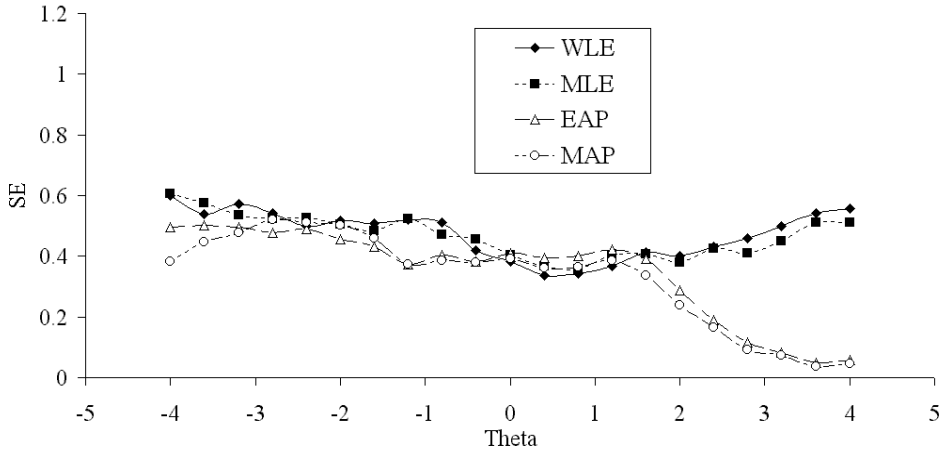

Figure 17. SE curves of the ability estimation methods, reliability=0.7, bank size=263.

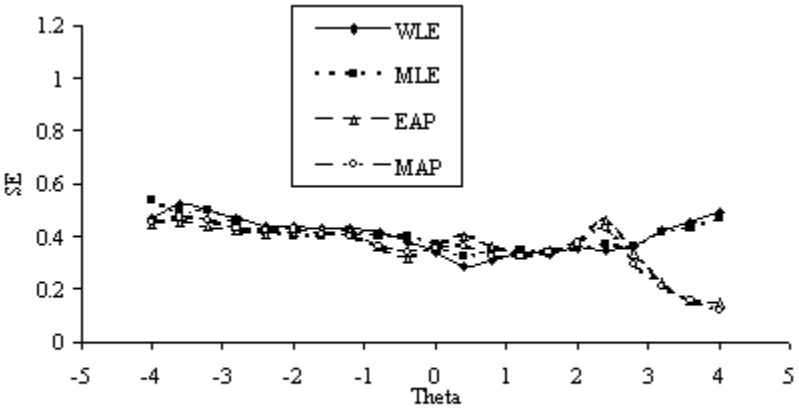

Figure 18. SE curves of the ability estimation methods, reliability=0.8, bank size=263.

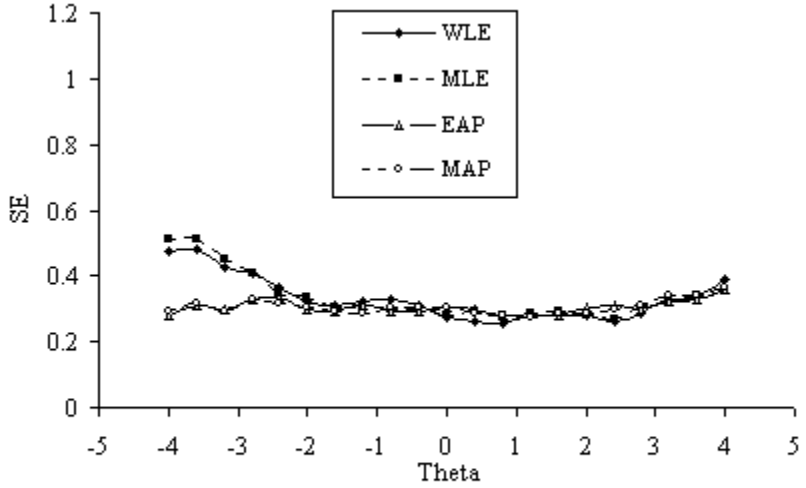

Figure 19 SE curves of the ability estimation methods, reliability=0.9, bank size=263.

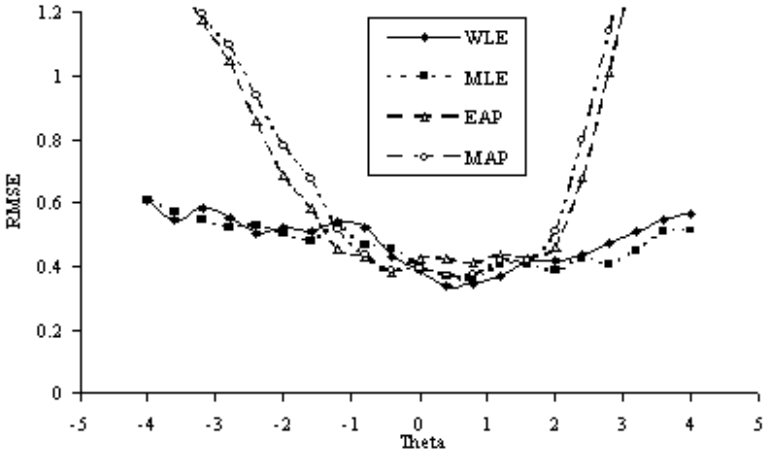

Figure 20. RMSE curves of the ability estimation methods, reliability=0.7, bank size=263.

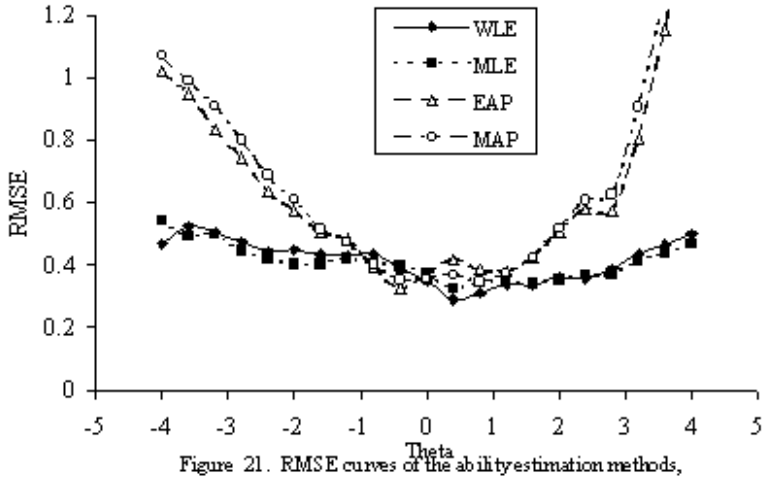

Figure 21. RMSE curves of the ability estimation methods, reliability=0.8, bank size=263.

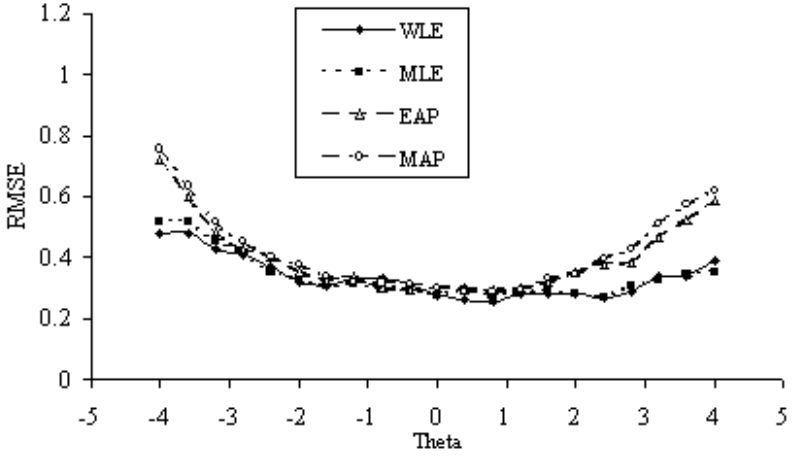

Figure 22. RMSE curves of the ability estimation methods, reliability=0.9, bank size=263.

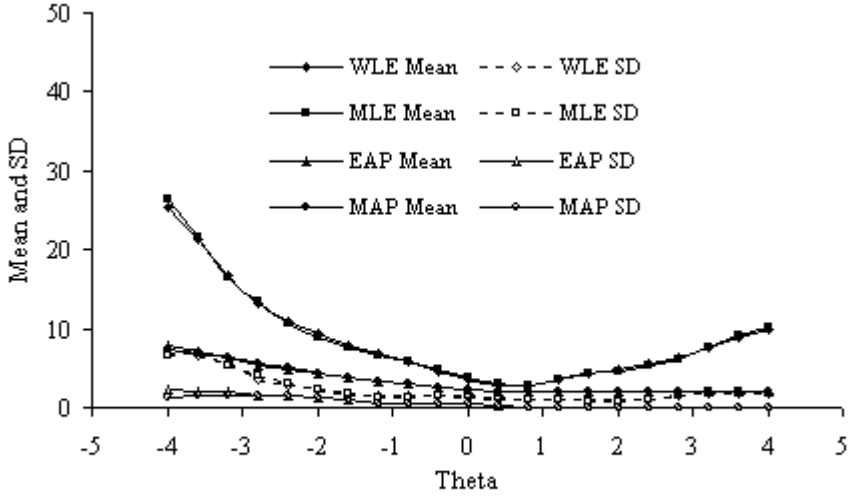

Figure 23. Mean and SD of number items, fixed reliability=0.7, bank size=263.

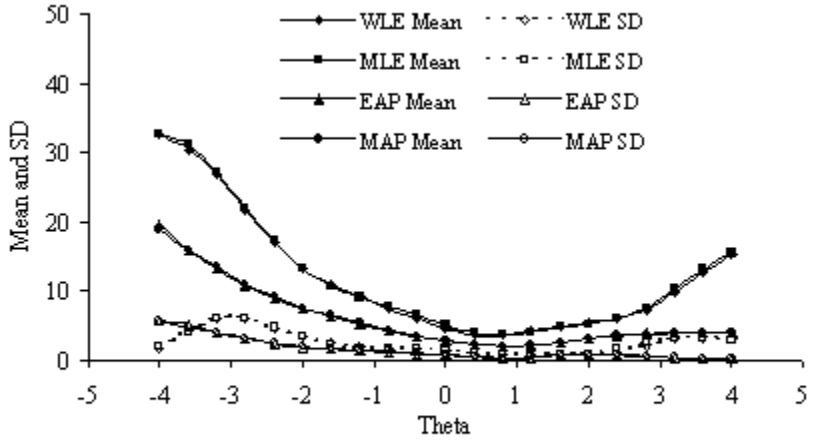

Figure 24. Mean and SD of number items, fixed reliability=0.8, bank size=263.

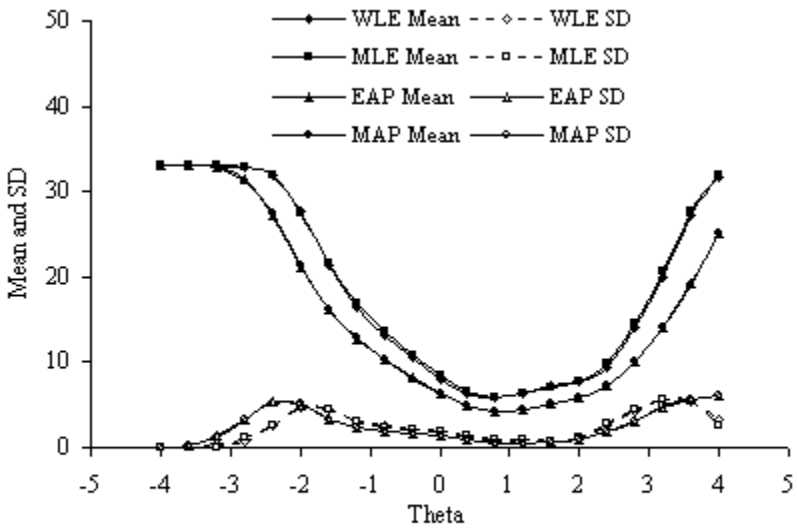

Figure 25. Mean and SD of number items, fixed reliability=0.9, bank size=236.