The Effect of Item Selection Methods on the Accuracy of CAT's Ability Estimates When Item Parameters Are Contaminated with Measurement Errors

by

Yuan H. Li

Prince Georges County Public Schools, Maryland

William D. Schafer

University of Maryland at College Park

Address:
Yuan H. Li
Prince George's County Public Schools
Test Admin. Department, Room, 202E
Upper Marlboro, MD 20772
e-mail:yuanhwangli@juno.com

Paper presented at the annual meeting of the National Council on Measurement in Education, April, 22-24, 2003, Chicago, IL.

The Effect of Item Selection Methods on the Accuracy of CAT's Ability Estimates When Item Parameters Are Contaminated with Measurement Errors

Abstract

For decades, most IRT studies that have involved ability estimates assumed that the item parameters are true values without measurement errors. The impact of this assumption on the accuracy of CAT ability estimates was addressed in this study for the three item selection methods: maximizing information function (MIF), matching item difficulty (MID) or optimal difficulty (MOD) using the test-taker's current ability estimate.

When CAT abilities were estimated from item estimates rather than their corresponding true parameters and when the data-model fit condition was met, their accuracy was slightly overestimated on one side of abilities ($\theta > -.5$) and was slightly underestimated on the other side ($\theta < -.5$). This mixed result was consistently found for all three item selection methods. This finding was encouraging, but also were consistent with our initial speculation that when item estimates were used in CAT, CAT might not result in as accurate ability estimates as those having been reported in literature.

Key Words: Computerized Adaptive Testing (CAT), Item Response Theory (IRT), Measurement Errors, Item Information, Optimal Item Difficulty

I. Introduction

A. Background

In the ideal testing condition, an examinee is given a test that is adapted to his/her ability level. Without the aid of computers, the earliest application of adaptive testing was in the work of Binet on intelligence testing in 1908. The rationale behind his adaptive testing was that an examiner should select items from an item pool so that each individual is examined using items at his/her appropriate difficulty level.

Recent advances in computer technologies as well as item response theory (IRT, Lord, 1980) have led to the development of computerized adaptive testing (CAT). In the past three decades, researchers have strived to seek promising methods in ability estimation and in item selection for CAT. For example, Warm (1989) proposed the weighted likelihood estimator (WLE) ability estimation that weights the maximum likelihood estimation (MLE) function in order to correct MLE biased trait (θ) estimates, especially when they are estimated from small numbers of items. Additionally, Chang and Ying (1996) recommended using the global Kullback-Leibler (KL) information instead of the most commonly used maximum item-information function (MIF) for the earlier-stage of CAT's item selection (see a comprehensive study by Chen, Ankenmann and Chang, 2000).

The MIF item selection method is based on the rationale that seeking items with the maximum information in the pool for an examinee can rapidly improve the examinee's ability estimate. Currently, growing interest is focused on whether or not the MIF method is appropriately employed from the beginning throughout the end of CAT (e.g., Chang, Qian & Ying, 2001; Hau & Chang, 2001; Leung, Chang & Hau, 2002; Veerkamp & Berger, 1999). The reasons for this speculation are explained below.

Theoretically, an item's information is dependent on the value of an examinee's ability parameter. The information value is most accurate when it is computed from an examinee's "true" rather than "inaccurate" ability estimate. The value of any ability estimate in the early-stage CAT is, however, poorly estimated and is not as precise as that estimated at the final stage. It is apparent that an early-stage item's information estimate is not as accurate as one calculated during a later stage and should have correspondingly less effect in the process of seeking the "true" ability estimate.

In general, an item with a higher discrimination parameter together with a lower guessing parameter will produce a larger information value for a given difference between estimated examinee ability and item difficulty. This fact leads the MIF CAT to over-administer higher discrimination and lower guessing items in the pool to examinees at the beginning of CAT. Afterward, they might take items having lower discrimination but greater guessing values at the final stages of CAT if the item pool is not sufficiently large. As a result, we might inappropriately use the most valuable items with sound statistical characteristics at the beginning of CAT and not make the best use of those items at the final stages (Hau & Chang, 2001). Secondly, we are likely to overuse the higher discriminating items so that test security is threatened and item exposure rates are uneven. Finally, although the MIF CAT will result in the most accurate ability estimates, theoretically as well as empirically, this holds true only if item parameters in the item pool are assumed true values without contamination by measurement errors. If this condition is violated as usually happens in real testing situations, ability estimates may not be as accurate as those reported in literature. The reason is that the MIF CAT is more likely to administer highly discriminating items to examinees. As a result, such items may be contaminated with greater measurement errors than those items with low discrimination

parameters as demonstrated in Figure 1. This diagram is a plot of analytically based standard errors (for details, refer to Li & Lissitz, in press; Thissen & Wainer, 1982) of item discrimination as a function of both item difficulty and discrimination parameters given an item's guessing parameter of 0.25 and a sample size of 1000. Consequently, the MIF CAT ability estimates derived from most highly discriminating items in the pool might not be as accurate as we expect because their potentially larger measurement errors are not taken into account.

Other item selection methods are not as highly dependent on the discrimination parameters as the family of information methods are (e.g., MIF, KL information, weighted item information, Berger & Veerkamp, 1997). One is based on a classical adaptive testing rationale. This method selects the item with item difficulty that is closest to the test-taker's current ability estimate. We call this the matching item difficulty (MID) item selection method. Another item selection method, similar to the MID, is to select the item based on proximity of its optimal difficulty (OD, refer to equation 3) value as a criterion, in stead of using the item difficulty value. An item's OD value is defined as the location of the item's maximum information given its item parameter estimates (Lord, 1980, p.152). We call this item selection method as matching optimal difficulty (MOD). An example of MOD in CAT studies can be found in Cheng and Liou's study (2000), which showed that MOD CAT could result in comparable results with MIF CAT, although MIF CAT performed slightly better when item estimates were used in CAT.

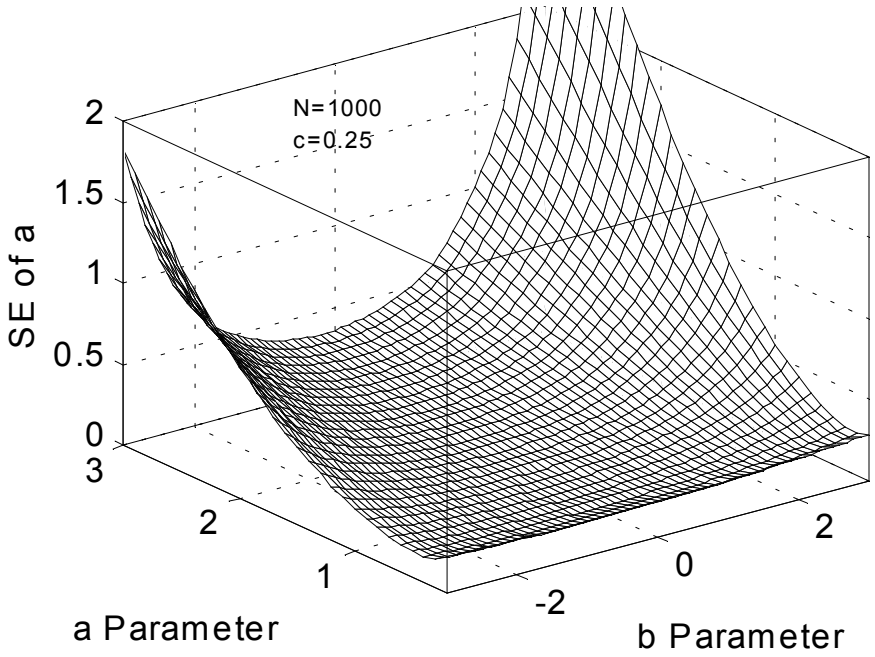

Figure 1. Standard Error of Item Discrimination Shown as a Function of Item Difficulty and Discrimination Parameters for the 3PL Model.

B. Research Purpose

MIF (or item-discrimination-oriented) item selection has dominated CAT's development since the beginning of CAT because it produces accurate ability estimates. But the MIF item selection method exhibits several potential problems as discussed earlier. Further, if error in item

parameter estimates is factored into the process of CAT's ability estimation, the degree to which accurate ability estimation can be maintained is of interest. On the other hand, an empirical study conducted by Li and Schafer (2003) indicated that both non-item-discrimination-oriented methods (e.g., MID and MOD) were capable of mitigating some potential problems (e.g., test security) caused by MIF. This finding encouraged us to further compare the robustness of these two methods with the MIF method when they were all employed using item parameters that were contaminated with some degree of error. If one of non-item-discrimination-oriented methods performs well on this issue, it might become a viable choice. These questions provided the motivation for the present Monte-Carlo study. The ultimate goal of this research was to explore how accurate these three item selection methods are, in terms of their corresponding estimated abilities, under a variety of simulation conditions, and especially when item parameters used to estimate CAT abilities are contaminated with measurement errors.

II. Overview of CAT Techniques

A. 3PL Model

The commonly-used three-parameter (3PL) logistic IRT model was used to model the dichotomous scored items in this study. Under the 3PL model, the probability, P_{ji}, of the correct response on an item i for an examinee with ability θ_j is given by the following function (Lord, 1980).

$$P_{ji} = c_i + (1 - c_i)\frac{\exp(Da_i(\theta_j - b_i))}{1 + \exp(Da_i(\theta_j - b_i))} \quad (1)$$

where
the symbol of "exp" stands for the mathematical function of the natural logarithm exponential,
a_i is the item discrimination,
b_i is the item difficulty,
c_i is the lower asymptote parameter (also known as the guessing parameter), and
D is a scaling factor (usually equal to 1.702).

The scaling factor D is included in the model to make the logistic function as close as possible to the normal ogive function (Baker, 1992).

B. Item Selections Used in this Study

In this study, the item selection methods of MIF, MID, and MOD are operationally defined as:

(a) MIF: select an item with a maximum Fisher information value from the item bank,
(b) MID: select an item with a minimum absolute difference value between the test-taker's current ability estimate and this item's difficulty value, and
(c) MOD: select an item with a minimum absolute difference value between the test-taker's current ability estimate and this item's OD value.

The Fisher information is computed at the current ability estimate, that is:

$$I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)} \quad (2)$$

where, $P_i'(\theta)$ is the first partial derivative of $P_i(\theta)$ with respect to θ. An item's OD value is defined at where this item's maximum information is located, based on the item-pool difficulty scale, (Lord, 1980, p.152). For the 3PL model, it is expressed as:

$$OD = b_i + \frac{1}{Da_i} \ln\left(\frac{1+\sqrt{1+8c_i}}{2}\right) \tag{3}$$

For the case of 3PL data modeling, both FI and OD values are derived from the 3PL item's parameters, a, b, and c. The difference is that the FI value depends on the ability (θ) scale; in contrast, the OD does not. A FI value may be misleading, as indicated previously, if it is computed from an ability estimate that is far away from its true value.

C. CAT Ability Estimates

1. Maximum Likelihood Estimator (MLE)

Assuming that the local independence assumption holds, given an examinee with an ability, θ, who responds to a set of n items with the response pattern $\underline{u}$, then the probability (or likelihood function) of obtaining this response pattern $\underline{u}$ can be modeled by:

$$L = L(\underline{u} \mid \theta) = \prod_{i=1}^{n} P_i(\theta)^{u_i} Q_i(\theta)^{1-u_i} \quad , i \in S_n \tag{4}$$

where $Q_i(\theta) = 1 - P_i(\theta)$ and S_n connote the n items that have been administered (or selected) to the examinee during the CAT testing process. The log of this likelihood function is given by:

$$\ln(L) = \sum_{i=1}^{n} \left[u_i \log(P_i(\theta)) + (1-u_i)\log(Q_i(\theta))\right] \tag{5}$$

Several methods of CAT ability estimation exist. For the maximum likelihood estimate of θ, MLE (θ), the log likelihood function (Equation 5) should be partially differentiated with respect to θ, then set to equal zero and finally used to solve this equation 6 for θ using the Newton-Raphson method or some other suitable numerical strategy. This equation is given below (Lord, 1980) :

$$\frac{\partial \ln(L)}{\partial \theta} = \sum_{i=1}^{n} \frac{P_i(\theta)'(u_i - P_i(\theta))}{P_i(\theta)Q_i(\theta)} \tag{6}$$

where ∂ denotes partial differentiation. A problem for the MLE is that it is unable to estimate examinees' abilities when they get all items right or wrong. This has become an issue especially at the early stage of CAT. Thus, this estimator was not considered in this research.

2. Bayesian Modal or Mean Estimator

θ is a parameter that needs to be estimated. If the prior information f(θ) for the distribution (or probability density) of θ together with the observed response pattern $\underline{u}$, are available, we are then able to approximate the posterior distribution of θ according to Bayes' rule. The posterior density of θ is:

$$f(\theta \mid \underline{u}) = \frac{L(\underline{u} \mid \theta) f(\theta)}{f(\underline{u})} \tag{7}$$

Where $f(\underline{u})$ is the marginal probability of $\underline{u}$ given by Bock and Lieberman (1970) and Bock and Aiken (1981):

$$f(\underline{u}) = \int_{-\infty}^{\infty} L(\underline{u} \mid \theta) f(\theta) d\theta \tag{8}$$

$f(\underline{u})$ is irrelevant while finding the solution of the θ parameter. Hence, the posterior function can simply be proportional to a prior function times a likelihood function. That is:

$$f(\theta \mid \underline{u}) \propto L(\underline{u} \mid \theta) f(\theta) \tag{9}$$

The relative influence of observed data (the input for the likelihood function) and prior information on the posterior function (related to the updated belief) depends on test-lengths, item-pool characteristics, and the magnitude of prior dispersion. As the prior becomes vague or diffuse, the posterior function is closely approximated by the likelihood function and consequently the Bayesian approach will result in the same solution as the likelihood approach. In contrast, if the prior is very informative or specific, then it would have a relatively greater influence on the posterior function.

For the maximum a posteriori (MAP) estimator , the estimate is the value that maximizes the posterior density function of $f(\theta \mid \underline{u})$. The $\hat{\theta}$ can be derived by partially differentiating the log-posterior density function with respect to θ , setting this equation equal to zero (Equation 10), and solving this non-linear equation:

$$\frac{\partial}{\partial \theta} \ln f(\theta \mid \underline{u}) = 0 \tag{10}$$

MAP is the mode of the posterior distribution. Another method to solve equation 10 is to find the mean of the posterior distribution of θ. This method is called the expected a posteriori (EAP) estimator. The mathematical expression for this estimator can be found in Bock and Aitkin (1981) and its features in CAT has been well documented by DeAyala, Schafer, and

Sava-Bolesta (1995). Wang, Hanson & Lau (1999) compared EAP with MAP estimation; EAP had slightly lower standard errors, but was slightly more biased.

III. Methodology

A. Adding Errors in the Estimation of Item parameters

This study used the procedure illustrated below to add a measurement error in each of true (or population) item parameters.

1. A Simulated Item Bank with True Item Parameters

In order to cover a variety of possible combinations of the 3PL item parameters, a, b and c, we included all possible combinations of 12 item discriminations (ranging from 0.60 to 1.70 in increments of 0.1), 21 item difficulties (ranging from –2.5 to 2.5 in increments of 0.25) and 3 guessing parameters (0.2, 0.3, and 0.4) into the item bank. Accordingly, a simulated 756-item bank with true item parameters was developed.

2. An Item Bank with Estimated Item Parameters

The item parameters in the simulated item bank were assumed to be true without measurement errors. The following steps were implemented to create an item bank whose parameters were derived with errors from the simulated bank;

(a) Create Simulated Tests: The 756 items were divided into 12 simulated tests, where each test had 63 items.

(b) Generate a Group of Simulees' Abilities: The abilities of the 1890 simulees were randomly generated with the assumption that they follow a standard normal distribution, N(0,1). The sample size was set at 1890 to meet the requirement that the sample size to parameter ratio be 10:1. The ratio was defined by the sample size to the number of item parameters (De Ayala & Sava-Bolesta, 1999) and the sample size of 1890 was derived from: 3 (3PL) x 63 (Number of test items) x 10.

(c) Generate Simulees's Responses to Test Items: The probability given by the true item parameters and an ability parameter was computed using Equation 1 and then compared to a random uniform number with the range [0,1]. When the probability is larger than the value of the random number, the corresponding item response was correct one, otherwise incorrect. This procedure was repeatedly carried out for all 1890 simulees taking each of the 12 tests. Finally, 12 test datasets with an 1890 x 63 data matrix were obtained.

(d) Estimate Item Parameters: We fit the 3PL model to each of the 12 test datasets to obtain item parameter estimates using the computer program BILOG (Mislevy & Bock, 1990), in which the MMLE (marginal maximum likelihood estimation) /Bayesian estimation method was chosen, the program default prior distributions for the slope, threshold and guessing parameters were used, and the convergence criterion was set at .0001. Since all 12 test datasets were generated from the same group of 1890 simulees, the item estimates independently calibrated from the 12 tests have been placed on the same scale.

(e) Transform Item Parameters: We transformed the metric of the estimated parameters to the one defined by the population (or true) parameters using matching test characteristics curves (Stocking & Lord, 1983). Thus, a 756-item item bank was established.

3. 100 Item Banks with Estimated Item Parameters

Up to this point, we have created two item banks, one with true item parameter values; the other with estimated item parameter values. The difference in item parameters between two item banks is the result of measurement error. However, the measurement error generated by the above steps is not a constant; it will vary if we repeat the steps again. Accordingly, we repeated the above steps 100 times to generate 100 item banks with estimated item parameters.

B. True Ability Levels and Test Starting Points

We included 31 points on the true ability or θ scale, ranging from –3.0 to 3.0 in increments of 0.2. The initial ability of all simulated subjects at the beginning of the test was taken as a randomly drawn value from the uniform distribution with range [0,1]. The EAP ability estimator was used to estimate abilities, assuming the prior distribution of examinee's abilities to be distributed as a standard normal distribution, N(0,1).

C. Item Selections and Test Length

There were three item selection methods implemented in this study (see the introduction) and four types of test length (TL). The lengths were 10, 20, 30 and 40 items.

D. Algorithms in Simulating CAT

For each simulated examinee these procedures were followed:

(a) Set the an initial ability estimate for the examinee. This was done by randomly drawing a uniform number with the range [0,1].
(b) Select the item from the first item bank (with estimated item parameters) using one of item selection methods.
(c) Re-estimate abilities based on the examinee's response(s) to the items that have been administered.
(d) Repeat the steps b-c until the fixed n items have been administered.

Using the above steps, an examinee's ability estimate was obtained. This estimate was the result of using item parameter estimates, instead of using the true item parameters. Up to this point, for each simulee, we know what items this simulee has taken and what true parameters values these administered items have. This set of information would make it possible to re-estimate this examinee's ability estimate using the true item parameters. This ability estimate using the true item parameters was obtained at the end of the adaptive test and did not influence item selection. Afterwards, two ability estimates existed for this simulee; one obtained from the item parameter estimates; one produced by the true parameters.

So far, the above steps have been implemented in the first item bank. They were then repeatedly implemented to the second, third … 100^{th} item banks. Finally, each examinee had 100 ability estimates based on item parameter estimates for each of the item selection methods and 100 based on true item parameters.

E. Data Analyses and Evaluation

One hundred replications for each condition were conducted. Afterward, the BIAS and RMSE (root mean squared error) for each of the ability estimates were calculated by the formulas shown below.

$$BIAS(\theta_j) = \frac{\sum_{j=1}^{r}(\hat{\theta}_j - \theta_j)}{r} \quad \text{and} \quad (11)$$

$$RMSE(\theta_j) = \sqrt{\frac{\sum_{j=1}^{r}(\hat{\theta}_j - \theta_j)^2}{r}} \quad (12)$$

where θ_j is the true ability parameter, $\hat{\theta}_j$ is the corresponding estimated ability parameter, and r is the number of replications, which was 100 in this study.

RMSE is a measure of total error of estimation that consists of the systematic error (BIAS) and random error (SE). These three indexes relate to each other as follows (Rao, 2000):

$$RMSE(\theta_j)^2 \cong SE(\theta_j)^2 + BIAS(\theta_j)^2 \quad (13)$$

As can be seen from Equation 13 either a large variance (SE^2) or a large BIAS will produce a large RMSE. It is apparent that an estimator will have much practical utility only if it must not only be highly precise (or small SE^2), but also has small BIAS (Rao, 2000). The accuracy of an estimator is inversely proportional to its RMSE so that this RMSE index is the criterion of accuracy for an estimator (Rao, 2000). Accordingly, this index was primarily used to compare the accuracy of ability estimates when they were estimated under various simulation conditions. We also provided the BIAS results for reference if needed.

IV. Results

A. The Effect of Measurement Errors in Item Parameter Estimates on the Ability Estimate

1. Maximum Fisher Information Method

Regarding the MIF method, when test length was 10, the average RMSE (computed across 31 simulees) were .481 for both conditions of use of item parameters with and without measurement errors. When test length was 20, the average RMSE were .269 and .274 for use of item parameters with and without measurement errors, respectively. When test length was 30, the average RMSE were .212 and .217 for use of item parameters with and without measurement errors. When test length was 40, the average RMSE were .180 and .185 for use of item parameters with and without measurement errors. These summary statistics seem to differ very little in the accuracy of MIF-based ability estimates when they were estimated either from the item estimates or from the true parameters. However, when we take a closer look of Figure 2, which shows RMSE as a function of true θ for MIF method for test length (TL) = 10, 20, 30 and 40, we find that RMSEs results from the MIF/Item Estimates was less than those resulted from the MIF/True Parameters when ability parameter was larger than about -.5. This implies that when the MIF is implemented in a real CAT testing program, the accuracy of its estimated ability reported in the literature might be overestimated for this range of ability estimates. This result is consistent with our initial speculation.

On the other hand, for an ability parameter that is less than about –0.5, we found that RMSEs resulted from the MIF/Item Estimates were larger than those resulted from the MIF/True Parameters. This implies that MIF tends to inappropriately inflate the amount of error for this range of low ability estimates.

Table 1
Descriptive Statistics of BIAS and RMSE of CAT's Ability Estimates for the Three Item Selection Methods (MIF, MID, and MOD) under Four Types of Test Length and Two Types of Item Parameters

TL	Method	Error	BIAS				RMSE			
			Mean	SD	Min	Max	Mean	SD	Min	Max
10	MIF	Yes	.057	.373	-.611	1.065	.481	.227	.288	1.236
		No	.025	.366	-.623	1.028	.481	.215	.297	1.205
	MID	Yes	.076	.609	-1.058	1.503	.685	.328	.347	1.665
		No	.056	.600	-1.061	1.473	.683	.319	.359	1.637
	MOD	Yes	.035	.561	-1.061	1.178	.651	.261	.366	1.314
		No	.018	.554	-1.064	1.159	.651	.254	.378	1.296
20	MIF	Yes	.010	.134	-.289	.324	.269	.070	.181	.476
		No	-.028	.139	-.370	.295	.274	.070	.190	.461
	MID	Yes	.031	.296	-.505	.708	.420	.159	.276	.873
		No	.005	.292	-.541	.696	.418	.149	.279	.848
	MOD	Yes	.011	.259	-.530	.534	.391	.127	.261	.690
		No	-.007	.257	-.535	.559	.391	.120	.271	.692
30	MIF	Yes	.008	.092	-.208	.209	.212	.048	.150	.324
		No	-.032	.102	-.311	.172	.217	.054	.161	.367
	MID	Yes	.023	.172	-.302	.410	.315	.101	.210	.610
		No	-.006	.184	-.413	.428	.316	.097	.218	.589
	MOD	Yes	.000	.160	-.371	.256	.297	.069	.214	.471
		No	-.019	.165	-.408	.290	.297	.067	.220	.471
40	MIF	Yes	.005	.063	-.144	.125	.180	.030	.137	.256
		No	-.034	.074	-.260	.085	.185	.037	.139	.307
	MID	Yes	.012	.117	-.240	.288	.246	.065	.168	.446
		No	-.020	.135	-.372	.319	.249	.064	.171	.440
	MOD	Yes	.001	.104	-.239	.186	.234	.052	.160	.345
		No	-.021	.115	-.324	.222	.235	.054	.165	.383

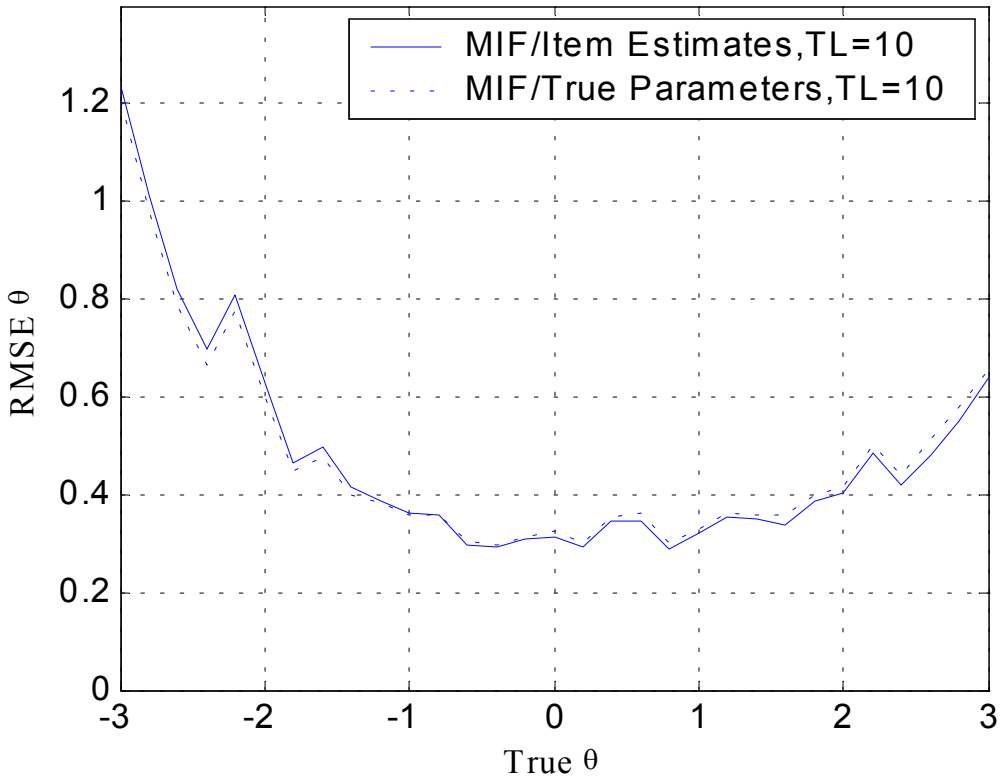

Figure 2a. RMSE as a Function of True θ for MIF Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 10.

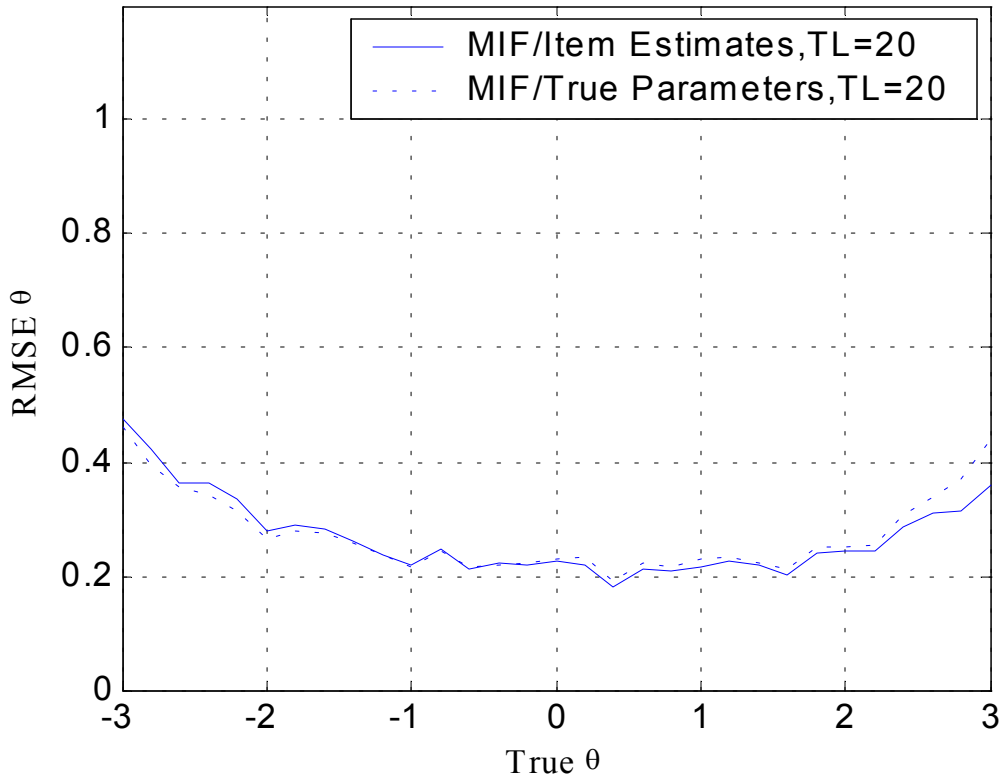

Figure 2b. RMSE as a Function of True θ for MIF Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 20.

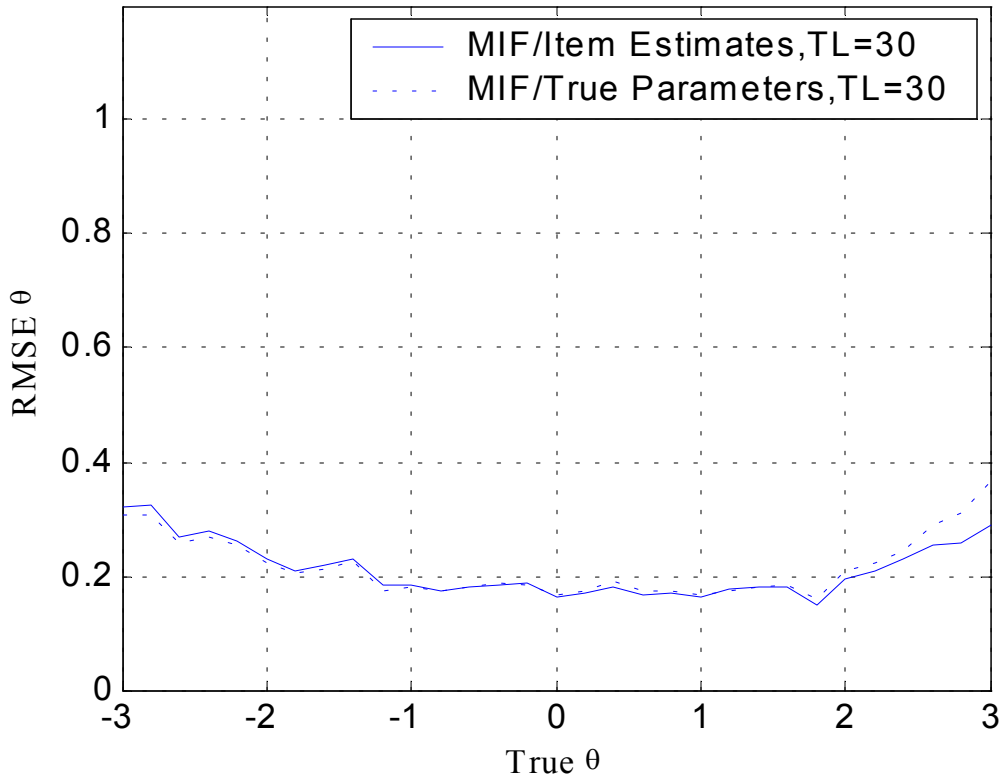

Figure 2c. RMSE as a Function of True θ for MIF Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 30.

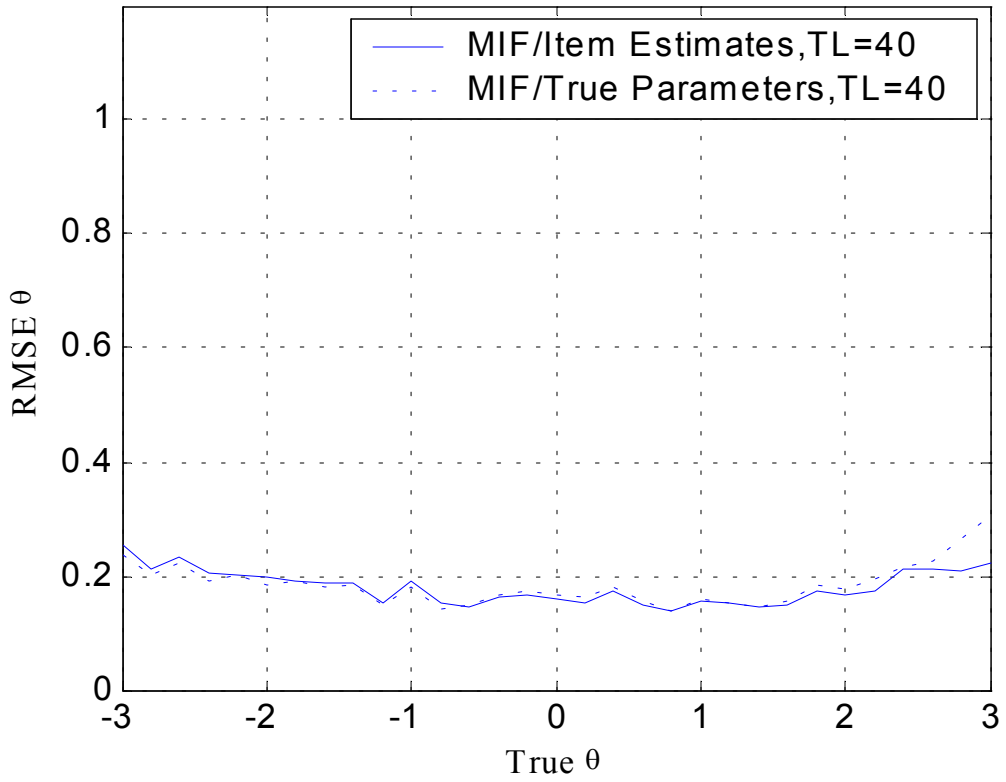

Figure 2d. RMSE as a Function of True θ for MIF Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 40.

2. MID Method

The results of the MID method are summarized in Table 1. As noted earlier, there is little difference in estimated abilities produced by item estimates or true parameters. As seen in Figures 3a to 3d, a similar result as that found for MIF was found. The accuracy of the estimated ability produced by MDO was slightly overestimated for those abilities that are larger than about -.5. For the abilities which are less than -.5, we might report more measurement errors than they should have if MID was implemented in CAT.

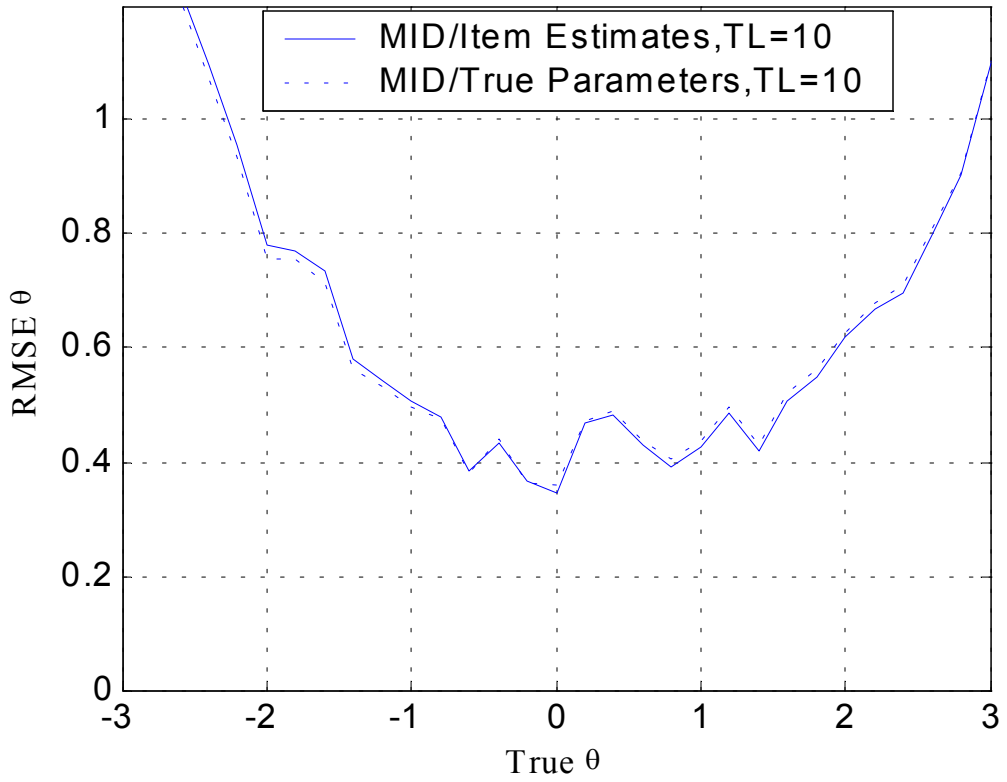

Figure 3a. RMSE as a Function of True θ for MID Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 10.

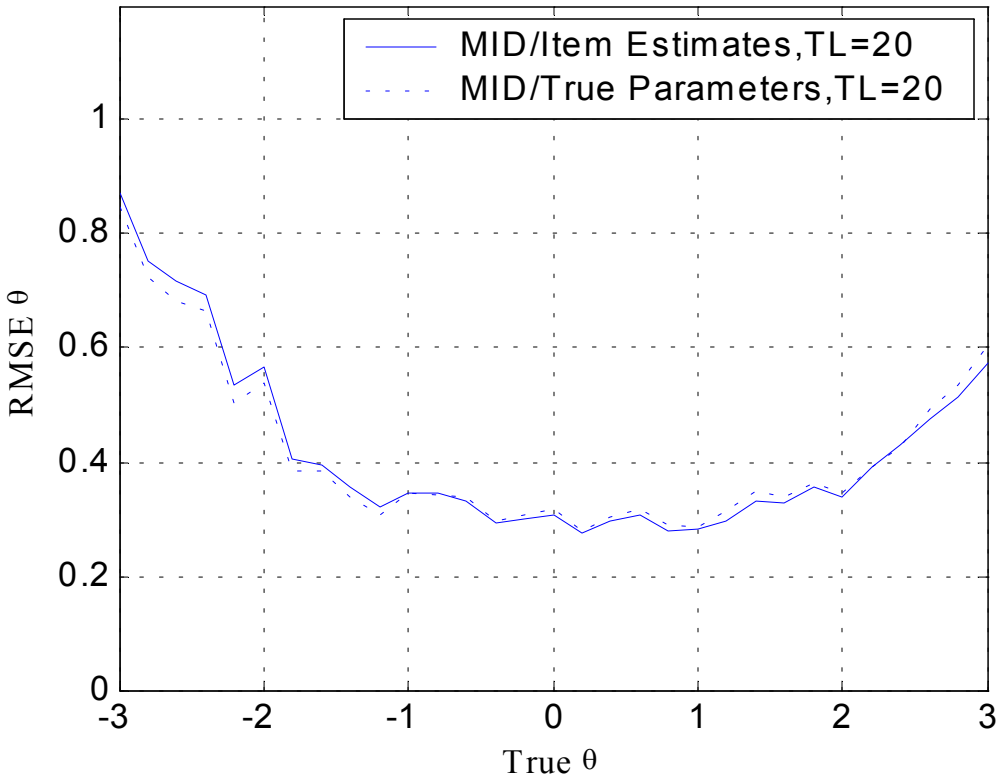

Figure 3b. RMSE as a Function of True θ for MID Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 20.

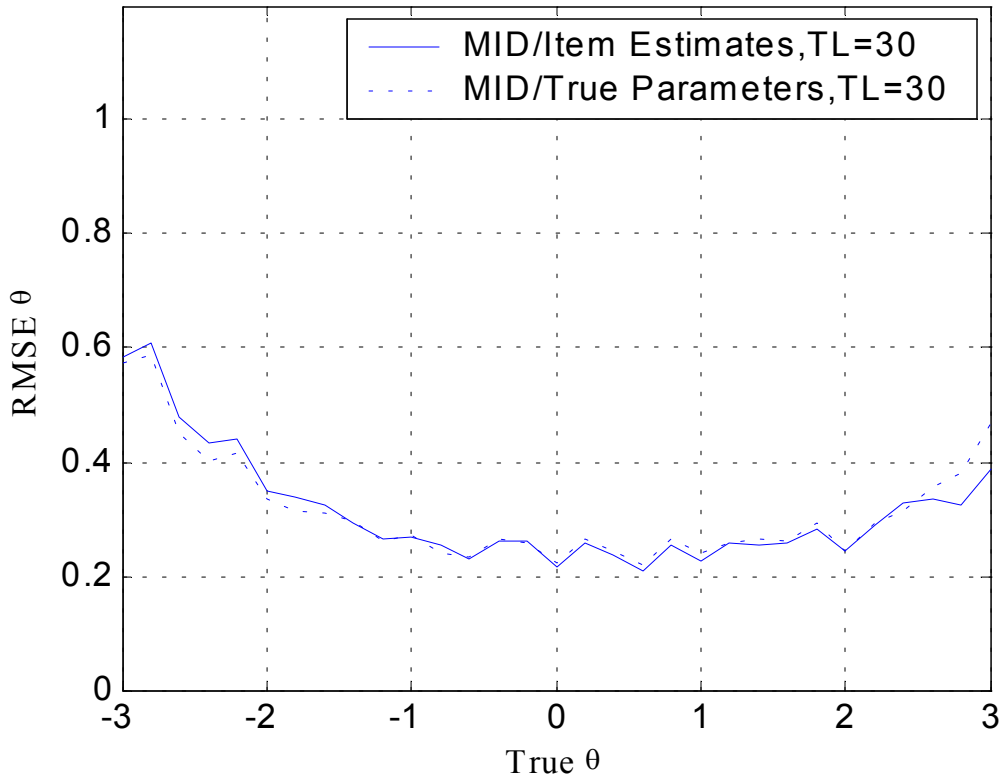

Figure 3c. RMSE as a Function of True θ for MID Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 30.

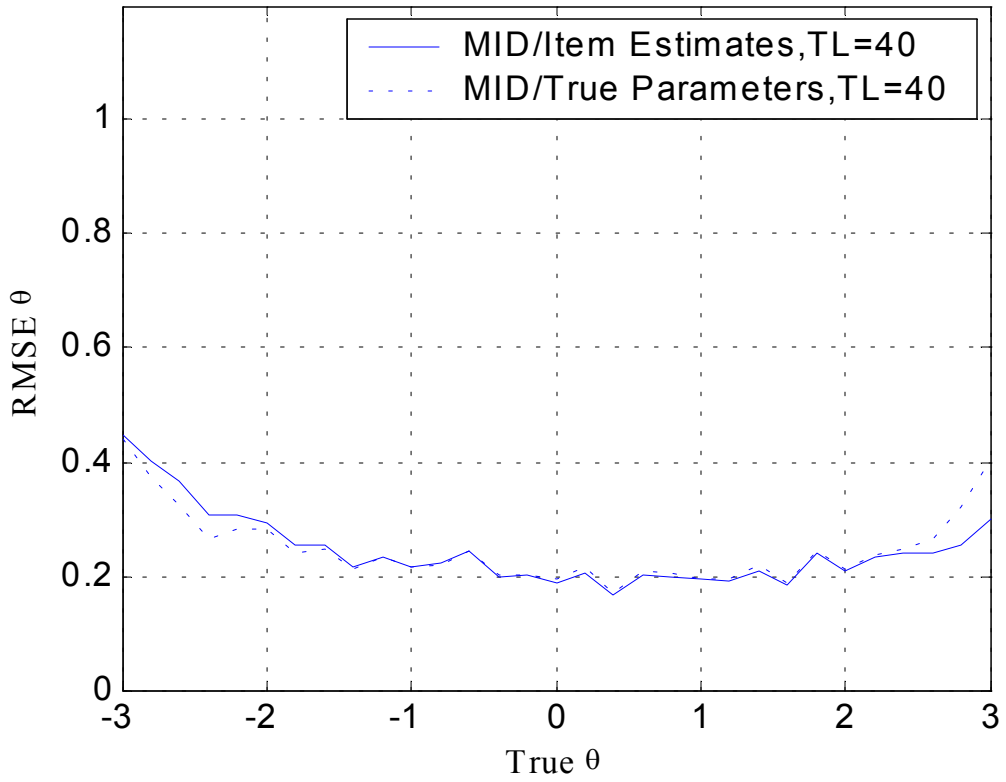

Figure 3d. RMSE as a Function of True θ for MID Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 40.

3. MOD Method

The results for the MOD method are summarized in Table 1. Again, there is little difference in estimated abilities produced by item estimates vs. true parameters. Further, the plot of RMSE against the true θ as shown in Figure 4 shows the same pattern as for the other two methods. As seen in Figure Figures 4a to 4d, the RMSE of the estimated ability produced by MOD was underestimated for abilities larger than about -.5. For abilities less than -.5, we might overestimate the RMSE of the estimated ability if MOD was implemented in CAT.

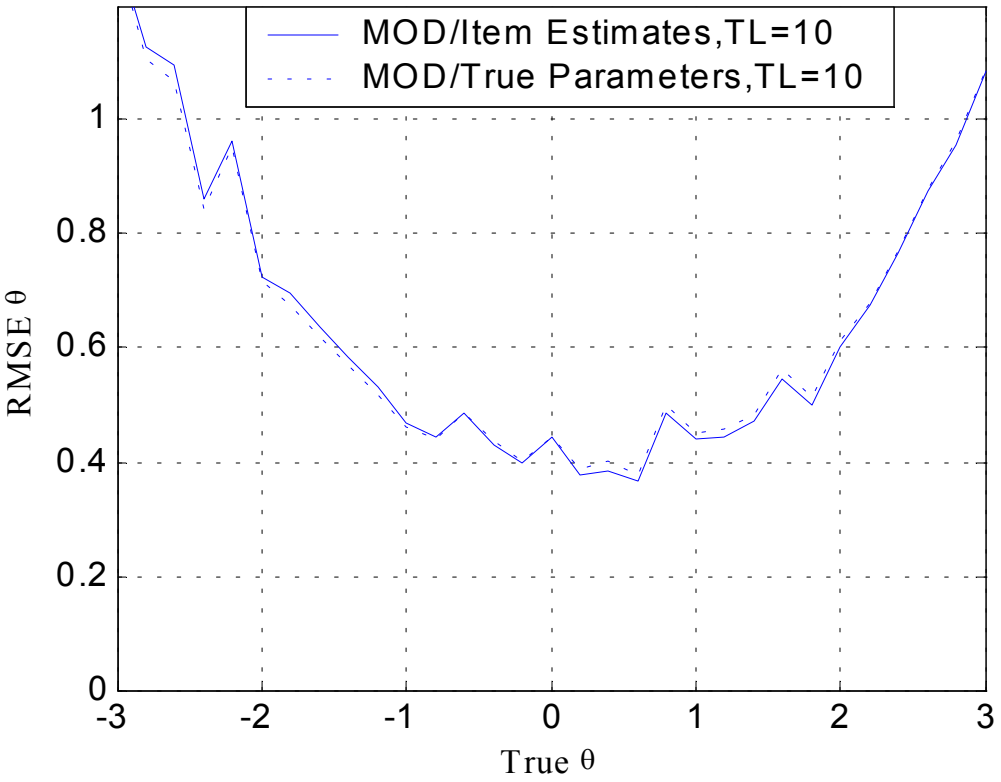

Figure 4a. RMSE as a Function of True θ for OID Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 10.

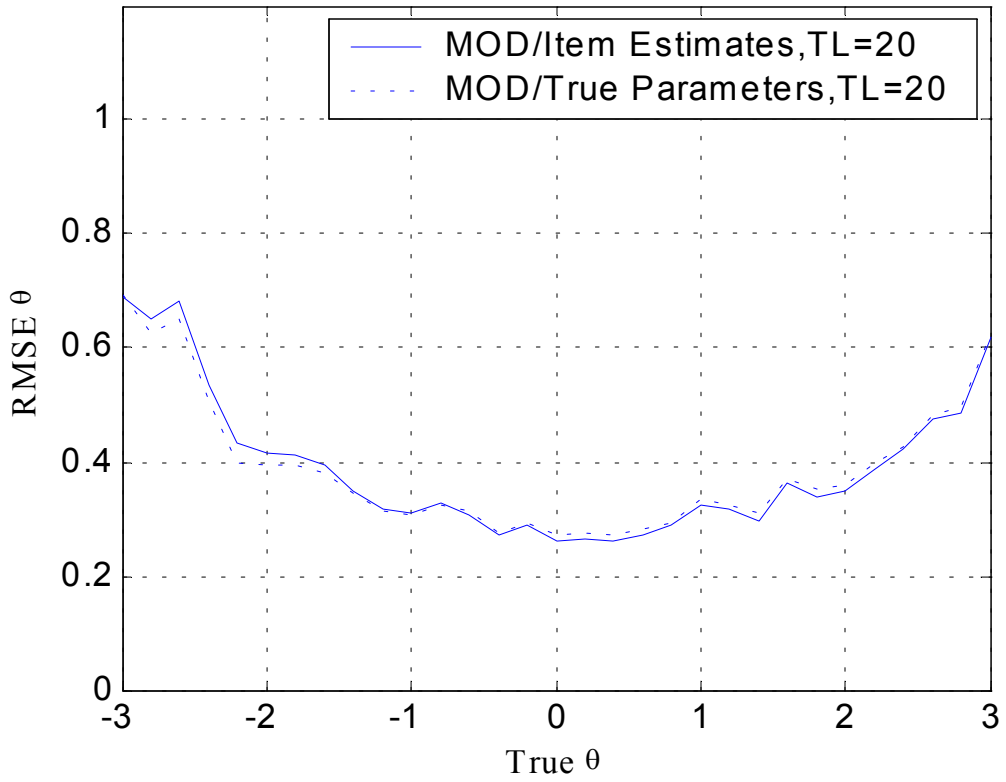

Figure 4b. RMSE as a Function of True θ for MOD Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 20.

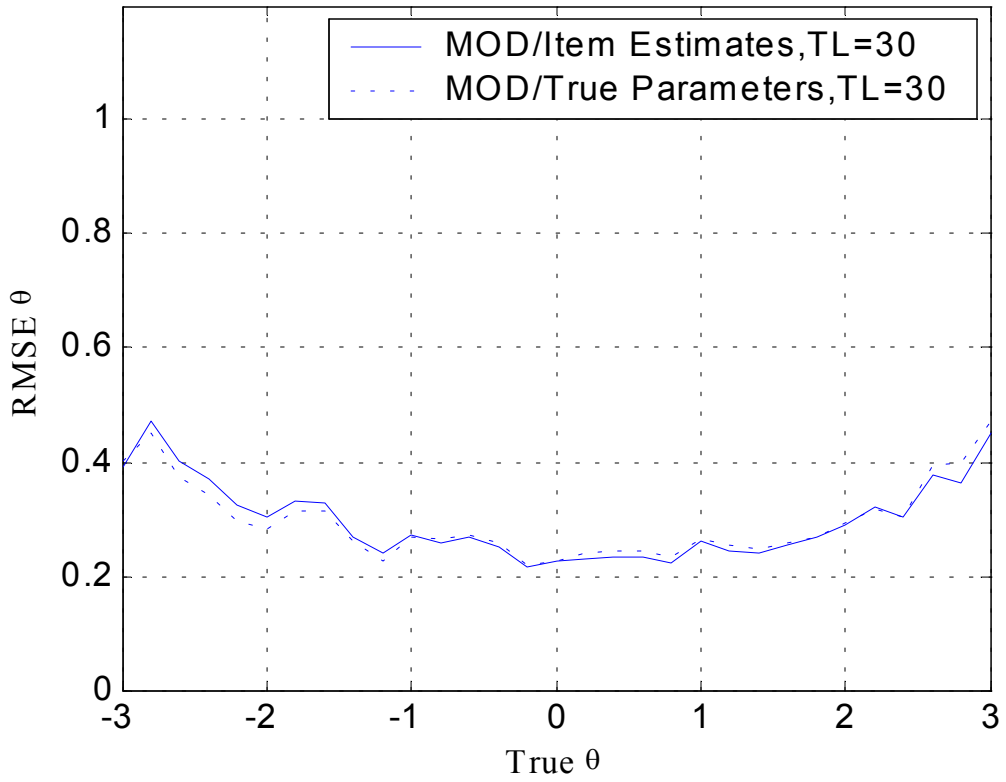

Figure 4c. RMSE as a Function of True θ for MOD Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 30.

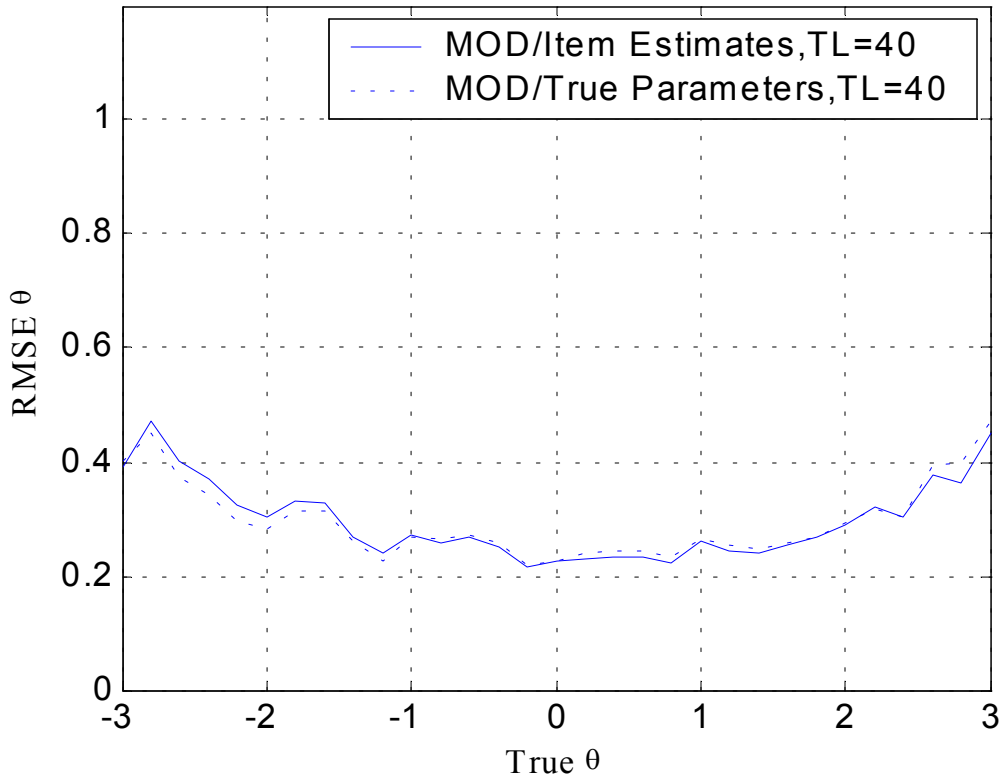

Figure 4d. RMSE as a Function of True θ for MOD Method for the Conditions of Use of Item Parameters with and without Measurement Errors When Test Length = 40.

B. Comparison Among MIF, MID and MOD Item-selection Methods

Comparisons among the MIF, MID, and MOD item selection methods in terms of accuracy of CAT ability estimates were explored by plotting their RMSE at different levels of ability. As seen as Figures 5a to 5d, MIF produced the lowest RMSE. The performance for MID and MOD was mixed. MOD performed slightly better than MID at some locations of ability, but did slightly worst than MID at others.

At the early-stage of CAT (e.g., TL=10), large RMSEs were found for high and low abilities when even the best MIF method was used. As the test length increased to 40, which is very close to a typical test length (e.g., 40 or 50) in real testing conditions, these three methods did not make much difference in assessing their corresponding RMSE, especially for almost 99% examinees whose ability values range between –2 and 2. [Note that we simulated examinees ability as a normal distribution, N(0,1).] While MIF made a sizeable difference in estimating extreme high or low abilities, examinees with such extreme abilities are rare in the population.

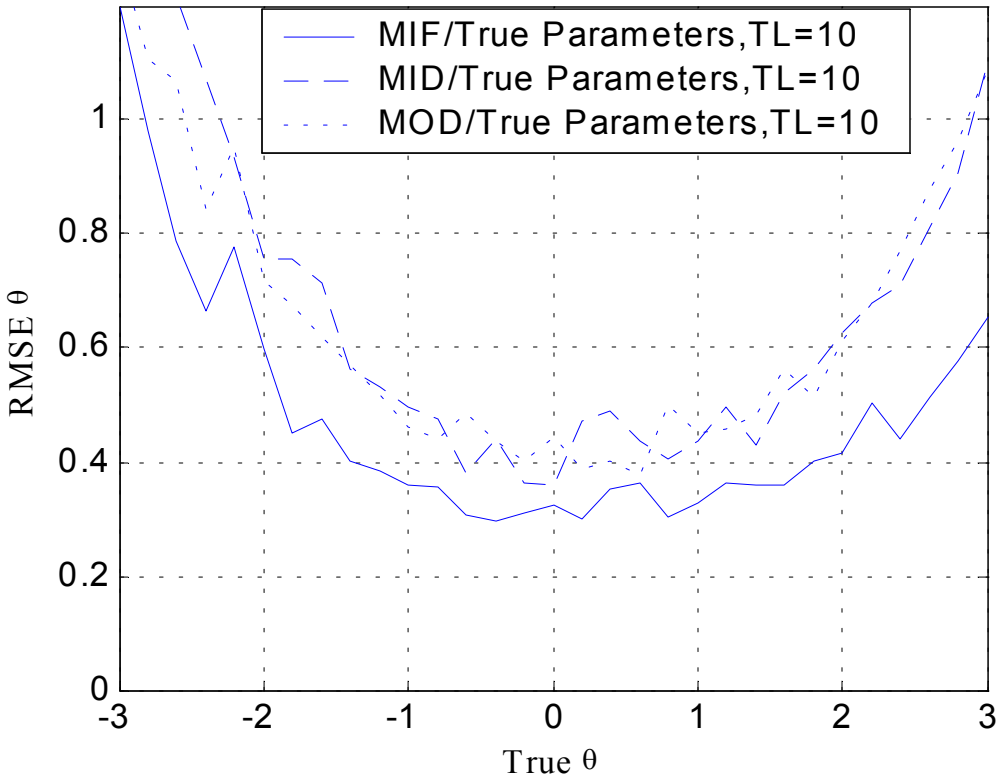

Figure 5a. RMSE as a Function of True θ for the MIF, MID and MOD Methods with True Item Estimates When Test Length = 10

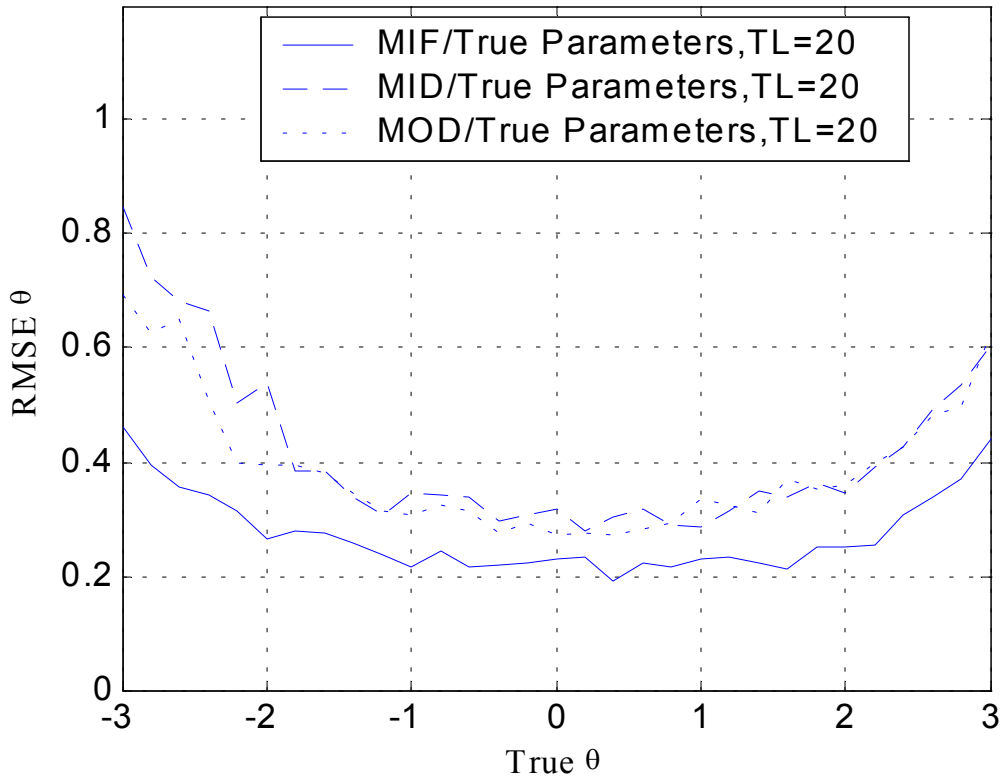

Figure 5b. RMSE as a Function of True θ for the MIF, MID and MOD Methods with True Item Estimates When Test Length = 20

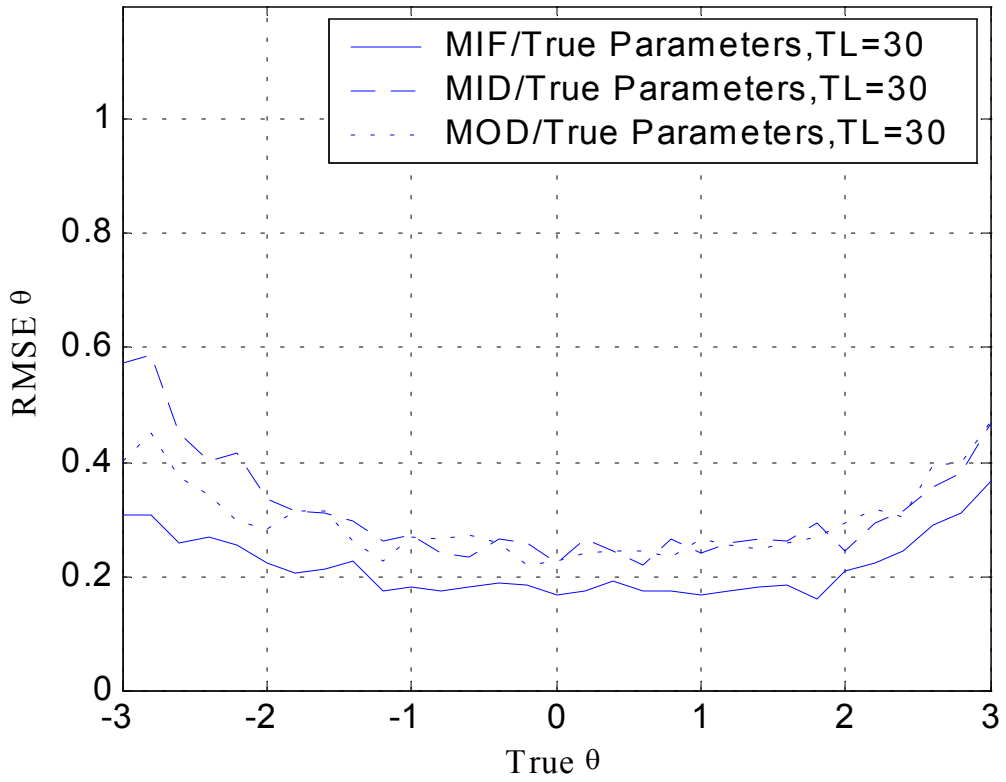

Figure 5c. RMSE as a Function of True θ for the MIF, MID and MOD Methods with True Item Estimates When Test Length = 30

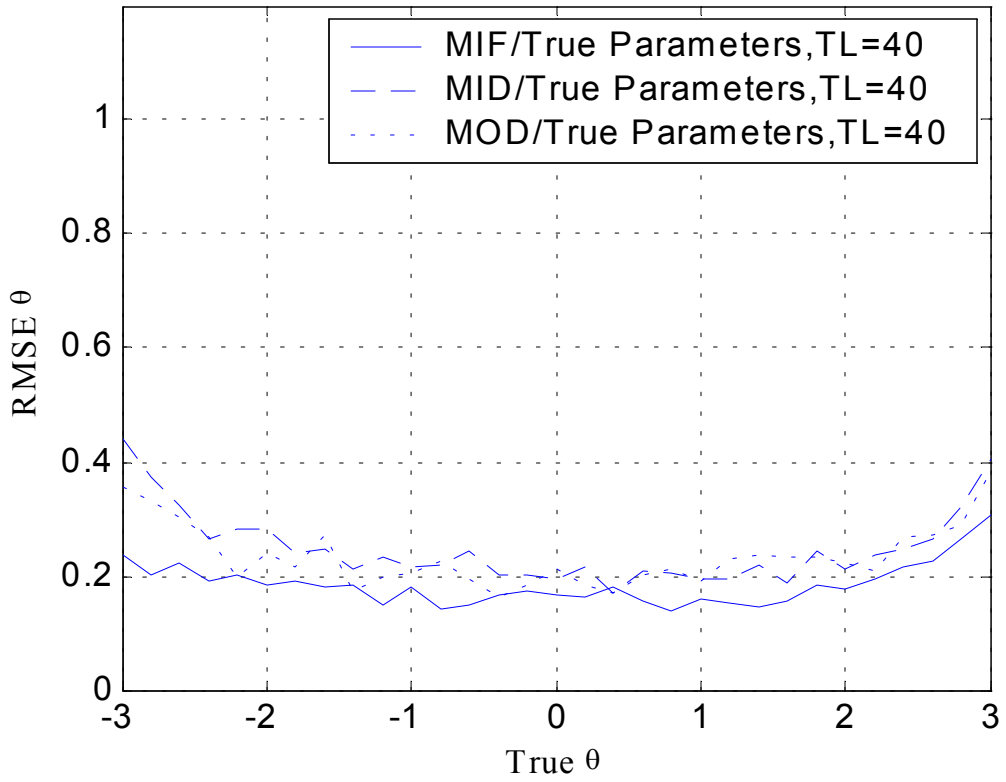

Figure 5d. RMSE as a Function of True θ for the MIF, MID and MOD Methods with True Item Estimates When Test Length = 40

V. Conclusions

As described in the introduction, MIF CAT tends to administer the most highly discriminating items in the pool to examinees. Its ability estimates appear accurate and stable, in part because of the assumption that the comparatively highly discriminating parameters of the items it selects are true, without the contamination by measurement errors. However, item parameters are contaminated with measurement errors in real settings and more highly discriminating parameters are likely to be contaminated with more measurement errors. Accordingly, the accuracy of ability estimates reported by the popular MIF CAT may be exaggerated . This study was initially motivated by this speculation

When the presence of item parameters that were contaminated with measurement errors was incorporated into the MIF CAT algorithm, we found that the above speculation was not as serious as we originally thought possible. The accuracy of ability estimates was slightly exaggerated in a range of abilities when they were larger than about –0.5. But we also found that in the opposite range of abilities (smaller than about –0.5), the accuracy of ability estimates was slightly understated. This mixed result was found not only in the MIF CAT, but also in other two methods, MID and MOD. When the data-model fit condition was met, as implemented in this study, measurement-error seemed to not have a strong effect on the quality of CAT in terms of accuracy of ability estimates.

This finding of no systematic measurement errors occurred when items were calibrated under data-model fit condition. Further, the measurement errors manipulated into this study

design had little impact on ability estimates as the test length increases. In the real testing situation, the model may not fit the data well and consequently quantifying the magnitudes of measurement errors in the item parameters can be much more complicated than the method used in this study. Thus, in real CAT implementations, we should be cautious in relying on the above conclusion based on our data.

Nevertheless, the above findings are encouraging and mitigate our initial speculation that when item estimates are used in CAT, CAT might not result in as accurate ability estimates as those having been reported in literature. We also found that the MOD CAT or MID CAT was almost as capable of recovering ability parameters as the MIF CAT when the number of test items is large (e.g., TL=40).

The apparent value of MIF CAT over MOD and MID in this study should be considered as well on the basis of item-exposure rate (see Revuelta & Ponsoda, 1998). That was not addressed in this study but has been investigated by Li and Schafer (2003). That study helps us clarify how these three methods compare and adds to the sufficiency of this study.

According to Li and Schafer's study (2003), when an additional content-balance control (van der Linden & Reese, 1998) was imposed into MIF CAT, the correlation of item-exposure rate and the a_i parameter was .60 and about 55% of items from the pool were not administered. In contrast, if the MOD was used under the same conditions as MIF, no correlation between item-exposure rate and the a_i parameter existed and only about 1 % of items from the pool were never administered. The mechanism of MOD item selection does not depend on item parameter values and thus make use of more items in the pool without significantly reducing the precision of ability estimates. These desirable features should make this method more appealing in real CAT testing applications in the near future, especially if more studies into this promising item selection method are explored.

References

Baker, F. B. (1992). Item Response Theory: Parameter Estimation Techniques. New York: Marcel Dekker, Inc.

Berger, M. P. F. & Veekamp, W. J. J. (1997). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics, 22, 203-226.

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

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Chang, H. H., Qian, J. & Ying, Z. (2001). a-Stratified multistage computerized adaptive testing with b blocking. Applied Psychological Measurement, 25, 333-341.

Chang, H. H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20, 213-229.

Chen, S., Ankenmann, R. D. & Chang, H. H. (2000). A comparison of item selection rules at early stages of computerized adaptive testing.

Cheng, P. & Liou, M. (2000). Estimation of trait level in computerized adaptive testing. Applied Psychological Measurement, 24, 257-265.

De Ayala, R. J. & Sava-Bolesta, M. (1999). Item parameter recovery for the nominal response model. Applied Psychological Measurement, 23, 3-19.

De Ayla, R. J., Schafer, W. & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. British Journal of Mathematical and Statistical Psychology, 47, 385-405.

Hau, K. & Chang, H. H. (2001). Item selection in computerized adaptive testing: Should more discriminating items be used first? Journal of Educational Measurement, 38, 249-266.

Leung, C., Chang, H & Hau, K. (2002). Item selection in computerized adaptive testing: Improving the a-stratified design with the sympson-hetter algorithm. Applied Psychological Measurement, 26, 376-392.

Li, Y. H. & Lissitz, R. W. (in press). Applications of the analytically derived asymptotic standard errors of IRT item parameter estimates. Journal of Educational Measurement.

Li, Y. H. & Schafer, W. D. (2003). Increasing the homogeneity of CAT's Item-exposure rates by minimizing or maximizing varied target functions while assembling shadow tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

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

Meijer, R. R. & Nering, M L. (1999). Computerized adaptive testing: Overviews and introduction. Applied Psychological Measurement, 23, 187-194.

Mislevy, R. J. & Bock, R. D. (1990). BILOG-3 (2nd ed.): Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software.

Rao, P. S. (2000). Sampling methodologies with applications. New York: Chapman & Hall/CRC.

Revuelta, J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measurement, 35, 311-327.

Stocking, M. L. Lord, F. M. (1983). Developing a common metric in items response theory. Applied Psychological Measurement, 7, 201-210.

Thissen, D. & Wainer, H. (1982). Some standard errors in item response theory. Psychometrika, 47, 397-412.
van der Linden, W. J. & Reese, L. M. (1998). A model for optimal constrained adaptive testing. Applied Psychological Measurement, 22, 259-270.
Veerkamp, W. & Berger, M. (1999). Optimal item discrimination and maximum information for logistic IRT models. Applied Psychological Measurement, 23, 31-40.
Wang, T., Hanson, B. A., & Lau, C. A. (1999). Reducing bias in CAT trait estimation: A comparison of Approaches. Applied Psychological Measurement, 23, 263-278.
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 4, 427-450.