Graduate
Management
Admission
Council®

Effect of Early Misfit In Computerized Adaptive Testing On the Recovery of Theta

Rick D. Guyer and David J. Weiss
University of Minnesota

Presented at the Realities of CAT Paper Session, June 2, 2009

Abstract

This study focused on how early misfit affected the recovery of θ for a computerized adaptive test (CAT). Number of misfitting items, generating θ, item selection method, and θ estimation method were independent variables in a monte-carlo simulation. It was found that CAT could recover from misfit-as-correct-responses for low ability simulees given a sufficient number of items. CAT could not recover from misfit-as-incorrect-responses for high ability simulees. Implications of the study and suggestions for future research are provided.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Copyright © 2009 by the Authors

Citation

Guyer, R. D. and Weiss, D. J. (2009). Effect of early misfit in computerized adaptive testing on the recovery of theta. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Rick Guyer, Assessment Systems Corporation, Suite 200, 2233 University Avenue,
St. Paul MN 55114, U.S.A. Email: guyerr@assess.com

Effect of Early Misfit in Computerized Adaptive Testing on the Recovery of Theta

A CAT is a test in which items are selected dynamically by computer based on the performance of the examinee. This is made possible by the estimation of examinee ability by the computer using item response theory (IRT) during the testing process. In a CAT, items are selected with difficulties similar to the ability of the examinee taking the test (Weiss, 1982).

Previous research in CAT has examined the recovery of the latent trait, θ, across θ estimation methods and/or item selection procedures. Previous studies (e.g., Bock and Mislevy, 1982; Chen and Ankenmann, 2004; Wang and Vispoel, 1998) have also compared the bias and standard error of Bayesian estimation methods to maximum likelihood estimation (MLE). Their results indicated that, although Bayesian methods reduced the standard errors, they biased θ toward the prior mean. The greater that θ deviated from the prior mean, the greater the bias was found to be.

Warm (1989) investigated the bias and SEs of his new weighted likelihood estimation (WLE) method. The results of his study were inconclusive due to small cell sizes. Recent research has found that WLE θ estimates provided lower SEs than MLE (Cheng & Liou, 2000; Yi, Wang, and Ban, 2001), but the difference between MLE and WLE became negligible after about 15 items were administered.

Use of Kullback-Leibler information (KLI) as an alternative to Fisher information (FI) for item selection was proposed by Chang and Ying (1996). Chang and Ying's study provided evidence that KLI selection reduced the bias and SE of θ until about 15 items were administered. The results from Chang and Ying must be tempered by their use of generating θ in the KLI equations, as generating θ is not known in applied testing.

The recovery of θ in CAT across item selection methods was investigated by a number of studies (Cheng & Liou, 2000; Cheng, Ankenmann, & Chang, 2000; Yi, Wang, and Ban, 2001; Chen & Ankenmann, 2004). These studies compared the performance of FI to KLI. The results of these studies indicated that the bias and standard errors for FI and KLI became similar after between 10-15 items were administered.

Item Selection in CAT

The equation for the probability of answering an item in the keyed direction (called the item response function, or IRF) for Birnbaum's three-parameter logistic (3PL) model (Lord and Novick, 1968, 2008) is

$$P_{ij}(u_i = 1 | \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{\exp[a_i \mathrm{D}(\theta_j - b_i)]}{1 + \exp[a_i \mathrm{D}(\theta_j - b_i)]}. \qquad (1)$$

The 3PL IRT model is a function of four parameters: a, b, c, and θ. The a parameter is proportional to the slope of the IRF at its maximum, while b represents the item difficulty parameter. The c parameter is defined as the probability of an examinee of infinitely low θ obtaining a correct response due to guessing. Thus, c is also the lower asymptote of the IRF. The latent trait θ is expressed on a standardized scale, so a one unit change equals a one standard deviation change. D equals 1.702 and is a constant used in the logistic model to approximate the normal ogive function.

The FI for an item (conditional on θ) is a transformation of the IRF and was defined by Lord (1977) as

$$I_i(\theta) = \frac{P_i'^2}{P_i Q_i}, \tag{2}$$

where $P_i'^2$ equals the squared first derivative of the IRF for item i, and Q_i equals $1 - P_i$. In CAT, the objective is to select items that provide maximum FI conditional on θ. Use of KLI as an alternative to FI was proposed by Chang and Ying (1996). The KLI function was defined as

$$K_i(\theta \| \hat{\theta}) = P_i(\theta) \log\left[\frac{P_i(\theta)}{P_i(\hat{\theta})}\right] + \left[1 - P_i(\theta)\right] \log\left[\frac{1 - P_i(\theta)}{1 - P_i(\hat{\theta})}\right], \tag{3}$$

where || denotes that $\hat{\theta}$ is separated from θ.

An index of KLI was obtained by Chang and Ying (1996) by integrating Equation 3

$$K_i(\hat{\theta}_n) = \int_{\hat{\theta}_n - \delta_n}^{\hat{\theta}_n + \delta_n} K_i(\hat{\theta} \| \hat{\theta}_n)\, d\hat{\theta}. \tag{4}$$

In this equation, δ_n equals the range over which the integral is to be calculated for the n^{th} item. The limits on the integral are with respect to $\hat{\theta}_n$, which is the θ estimate after n items have been administered. In Equation 4, $\hat{\theta}_n \pm \delta_n$ replaces $\hat{\theta}$ in the denominator of Equation 3 during the evaluation of the integral. The estimate $\hat{\theta}_n$ is substituted into Equation 3 for θ and is assumed to be fixed during calculation of the integral.

θ Estimation

MLE. The goal of MLE is to find an estimate of θ that maximizes the likelihood of observing the response pattern given the items administered. The log-likelihood (LL) function for the 3PL was defined by Baker and Kim (2000, p.66) as

$$LL(\mathbf{u}_j \mid \theta, a, b, c) = \sum_{i=1}^{n} u_{ij} \log[P_{ij}(\theta)] + (1 - u_{ij}) \log[Q_{ij}(\theta)]. \tag{5}$$

In order to locate the maximum of the log likelihood, an iterative procedure such as the Newton-Raphson method must be used (Embretson & Reise, 2000). Implementation of the Newton-Raphson method requires calculation of the first and second derivatives conditional on θ. The ratio of the first derivative to the second derivative (Hessian) is used to update the θ estimate ($\hat{\theta}$) until the ratio is smaller than a pre-determined criterion. This $\hat{\theta}$ is the MLE estimate of θ.

WLE. Warm (1989) proposed a WLE method that corrected for the first-order theoretical bias in MLE that was derived by Lord (1983). The weighted first derivative (WFD) of the LL is computed as

$$WFD = \frac{\partial(LL)}{\partial(\theta)} - BIAS_1(\hat{\theta}_{MLE}) I(\theta). \tag{6}$$

The WFD is obtained by subtracting the product of the test information function [$I(\theta)$] and the first-order bias function from the first derivative of the *LL* function. The WLE θ estimate can be obtained using a Newton-Raphson procedure modified to include the *WFD* and its derivative.

Bayesian methods. A Bayesian estimate of θ can be obtained by multiplying a prior distribution by the likelihood function. According to Bayes theorem, the posterior distribution is

$$f(\theta \mid \mathbf{u}_j, a, b, c) \propto P(\mathbf{u}_j \mid \theta_j, a, b, c) g(\theta) . \quad (7)$$

As discussed in Bock and Mislevy (1982), an expected a posteriori (EAP) θ estimate is obtained when the expectation (mean) of Equation 7 is computed. A modal a posteriori (MAP) θ estimate is obtained when a procedure such as Newton-Raphson iterations is used to locate the mode of the posterior distribution.

Misfit in CAT

The process of selecting items similar in difficulty to the examinee's θ level assumes that the θ estimate used is correct. One problem that can occur in IRT is person misfit. Misfit at the person level in IRT can be defined as item response(s) that are not likely given θ and the IRT model (Embretson and Reise, 2000). For low θ examinees, misfit would correspond to responding correctly to items above the person's θ, whereas for high θ examinees misfit would correspond to incorrect responses to items below the person's θ.

Rulison and Loken (2009) examined the effect of person misfit for the initial two items in a CAT on the recovery of θ. Recovery was assessed using the bias, standard error (SE), and root mean squared error (RMSE). Both MLE and EAP were used, in conjunction with FI as the item selection method. They examined the effect of misfit for both low and high θ examinees using the 3PL and Barton and Lord's (1981) four-parameter logistic (4PL) model. Support for the 4PL model derived from research by Reise and Waller (2003) and Waller and Reise (2009) who found that the 4PL model fit psychopathology data better than the 3PL model. The probability of a keyed response given the 4PL model is defined as

$$P_{ij}(u_i = 1 \mid \theta_j, a_i, b_i, c_i, d_i) = c_i + (d_i - c_i)\frac{\exp[a_i \mathrm{D}(\theta_j - b_i)]}{1 + \exp[a_i \mathrm{D}(\theta_j - b_i)]} . \quad (8)$$

The 4PL model introduces an upper-asymptote parameter (*d*), an upper bound for P_{ij}, into the model.

Rulison and Loken (2009) concluded that early misfit was more detrimental for high θ examinees when the 3PL model was used, as θ for low θ examinees recovered to near zero bias after 30 items. When the 4PL model was used with a fixed *d* of .98, they found that θ could be recovered with near zero bias after 45 items were administered.

Purpose

Rulison and Loken (2009) investigated the effect of early misfit on θ estimation using the 3PL and 4PL IRT models. Their study did not examine how θ estimates would recover from early misfit when WLE was used. In addition, their study used only FI as the item selection method. The present study introduced KLI item selection as an independent variable to investigate whether it affected recovery of θ. The generating θs in Rulison and Loken were not static, but rather the bottom and top 10% of the randomly generated sample. The present study controlled for variation in generating θ by using static values.

Method

Monte-Carlo Simulation

Based on previous research (e.g., Wang & Vispoel, 1998; van der Linden, 1998) it was observed that item banks in CAT typically have about 300 items. Item parameters for the 300 items used in this study were generated according to the following distributions: a ~ log-normal(–0.223, 0.2), b ~ U[–3.5, 3.5], c ~ N(.20, .02). The mean of a in the logistic metric was about 0.82 with a standard deviation of 0.15. A uniform distribution of b was used to avoid reduced precision for θ estimates above 2 in absolute value. The item response data for this study were obtained using a monte-carlo simulation. The 3PL IRT model defined by Equation 1 was used for item response generation.

Design

The following θ values were used in this study: –3, –2.5, –2, –1, 1, 2, 2.5, and 3. The initial θ estimate used to select the first item in the CAT was held constant at 0 for all generating θ conditions. The CAT was terminated after 50 items were administered. MLE, WLE and EAP were used for estimation of θ after each item in the CAT. A standard normal distribution was used as the prior for EAP. As MLE cannot obtain a finite estimate when the response pattern is non-mixed, $\hat{\theta}$ for MLE was incremented by –1 for each incorrect response, and +1 for each correct response, until $\hat{\theta}$ equaled 4 in absolute value.

Item selection. Both the FI and KLI item selection procedures were used in this study. This study set the limits of the confidence interval using $3/\sqrt{n}$ for δ. This δ was shown by Tang (1996) to provide the best recovery of θ (lowest bias and SE).

Introducing misfit. The number of responses that did not fit the 3PL model was varied from 0 to 4. Item responses for just the first k items (0 to 4) in the CAT were modified to introduce misfit. For examinees with θ above zero, misfit was operationalized as incorrect responses to the first k items. For examinees with θ below zero, misfit was operationalized as correct responses to the first k items.

Data generation. This study used a fully crossed 5 (misfitting items) × 3 (θ estimation)× 2 (item selection) × 8 (θ levels) design. To ensure stability in the results, 1,000 simulees were generated for each cell in the design. The program R (R Core Development Team, 2007) was used to simulate the CAT according to the specifications of this study.

Dependent Variables

Recovery of θ was assessed using the average unsigned bias, SE, and RMSE. The dependent variables were defined as

$$\text{Bias} = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta)}{N}, \qquad (9)$$

$$\mathrm{SE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \bar{\hat{\theta}})^2}{N}}, \tag{10}$$

and

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2}{N}}, \tag{11}$$

where *N* equals the number of simulees for the condition.

ANOVA

Although the average bias, SE, and RMSE indexed the recovery of *θ*, they did not provide information about any interactions among the independent variables. For this reason, an ANOVA approach was used for the data analysis. The empirical *θ* estimates were not appropriate for use as a dependent variable in an ANOVA for this study, because *θ* was an independent variable in and it would be possible to receive the same average *θ* estimate for two conditions – despite having different generating values of *θ.* Thus, the signed bias values for each simulee were used as the dependent variable for the ANOVA and provided 1,000 observations per cell.

The independent variables *θ* estimation, item selection, and *θ* were between-subjects factors in the ANOVA. As there was systematic redundancy in the misfitting item condition, that variable was a within-subjects factor in the ANOVA. The ANOVA was performed after 15 and 50 items were administered in the CAT.

Hypothesis testing for this study would not be as informative as an index of effect size due to the large sample size. An index of effect size was obtained for each effect in the ANOVA model. One advantage of the general η^2 is that it sums to 1.0. As shown, η^2 is a ratio of the sum of squares,

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}, \tag{12}$$

where SS_{effect} is the total variation attributable to a particular effect (e.g., *θ*) and SS_{total} is the total amount of variation in the study. For purposes of this study, an effect was defined as any non-error term in the ANOVA model.

Results

ANOVA

For the sake of brevity, only effects with η^2 larger than .01 are reported in Table 1. As seen in Table 1, the *θ* × misfit interaction had an η^2 of .286 after 15 items and .269 after 50 items. The main effects of *θ* and misfit accounted for a large portion of the total effect variance. For a more complete discussion of the ANOVA see Guyer (2008).

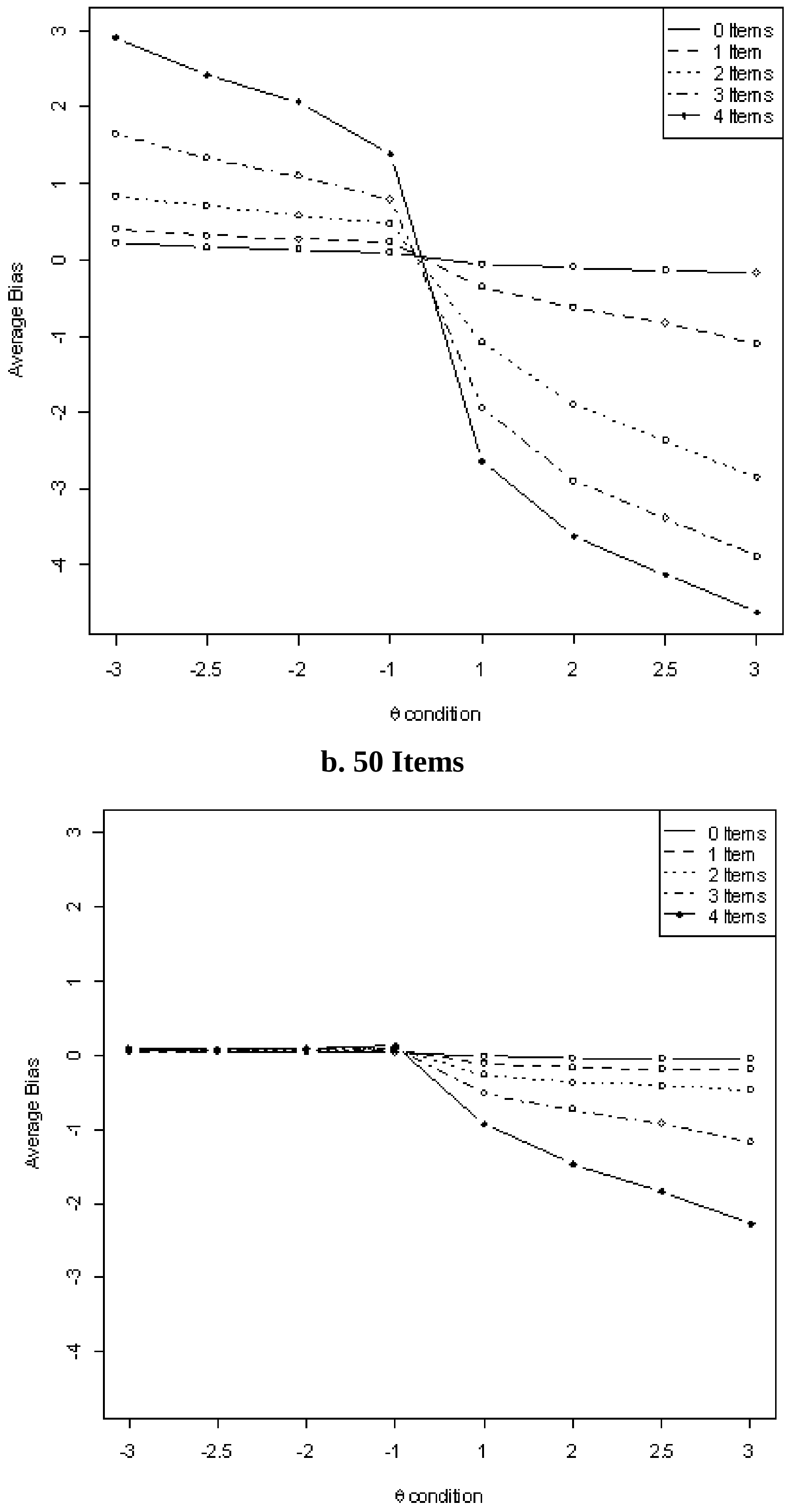
Table 1. Effects with η^2 Greater than .01 From the Mixed Design ANOVA For 15-Item and 50-Items CATs

Source of Variation	*SS*	*df*	*MS*	η^2
15-Item CAT				
Between Subjects				
θ	539684.186	9	59964.910	.538
Within Subjects				
Misfit	29839.703	4	7459.926	.030
θ × Misfit	286821.417	36	7967.262	.286
$\theta \times \hat{\theta}$ × Misfit	18185.274	72	252.573	.018
Total Effect				.888
Total Error				.112
50-Item CAT				
Between Subjects				
θ	34470.425	9	3830.047	.327
Within Subjects				
Misfit	17353.075	4	4338.269	.165
θ × Misfit	28341.566	36	787.266	.269
$\hat{\theta}$ × Misfit	1173.808	8	146.726	.011
$\theta \times \hat{\theta}$ × Misfit	1568.846	72	21.790	.015
Total Effect				.812
Total Error				.188

As shown by Figure 1a, there was substantial bias present in the θ estimates after 15 items were administered. It can be seen that the bias was greater for the misfit-as-incorrect-responses (MIR) conditions than the misfit-as-correct-responses (MCR) conditions. In addition, the bias increased as θ became larger in absolute value.

It is evident in Figure 1b that CAT could recover from MCR, given a sufficient test length. Negative bias still remained in the θ estimates after 50-items were administered for the 1-item misfit condition with θ = 1. Figure 1b shows that the bias increased both as θ increased and as the number of misfitting item responses increased.

Figure 1. Average Bias for the *θ* × Misfit Interaction



a. 15 Items

b. 50 Items

MCR

Bias. To better interpret the results of the ANOVA, the bias values after 6 to 50 items were administered are shown by Figure 2. For the sake of brevity, only the results for $\theta = -1$ with 2-item misfit are shown. Guyer (2008) found that it took longer for CAT to recover from MCR as both θ and number of misfitting items increased. It can be seen in Table 2 and Figure 2 that the bias of MLE and WLE were reduced to .129 and .127 after 35 items were administered. The results indicate that EAP had less bias than MLE or WLE for the first 20 items in the CAT. Figure 2 shows that KLI made the θ estimates more biased than FI, especially early in the CAT.

SE. Figure 3 displays the empirical SEs of the θ estimates across item selection methods. It is evident in Table 2 that EAP provided substantially lower SEs than MLE or WLE, especially when fewer than 25 items were administered. As with the bias, KLI resulted in larger SEs than FI across θ estimation methods. The effect of MCR on the SEs of MLE is evident in Figure 3, as the flattening of the likelihood due to early misfit caused the empirical SEs to spike above 1.0.

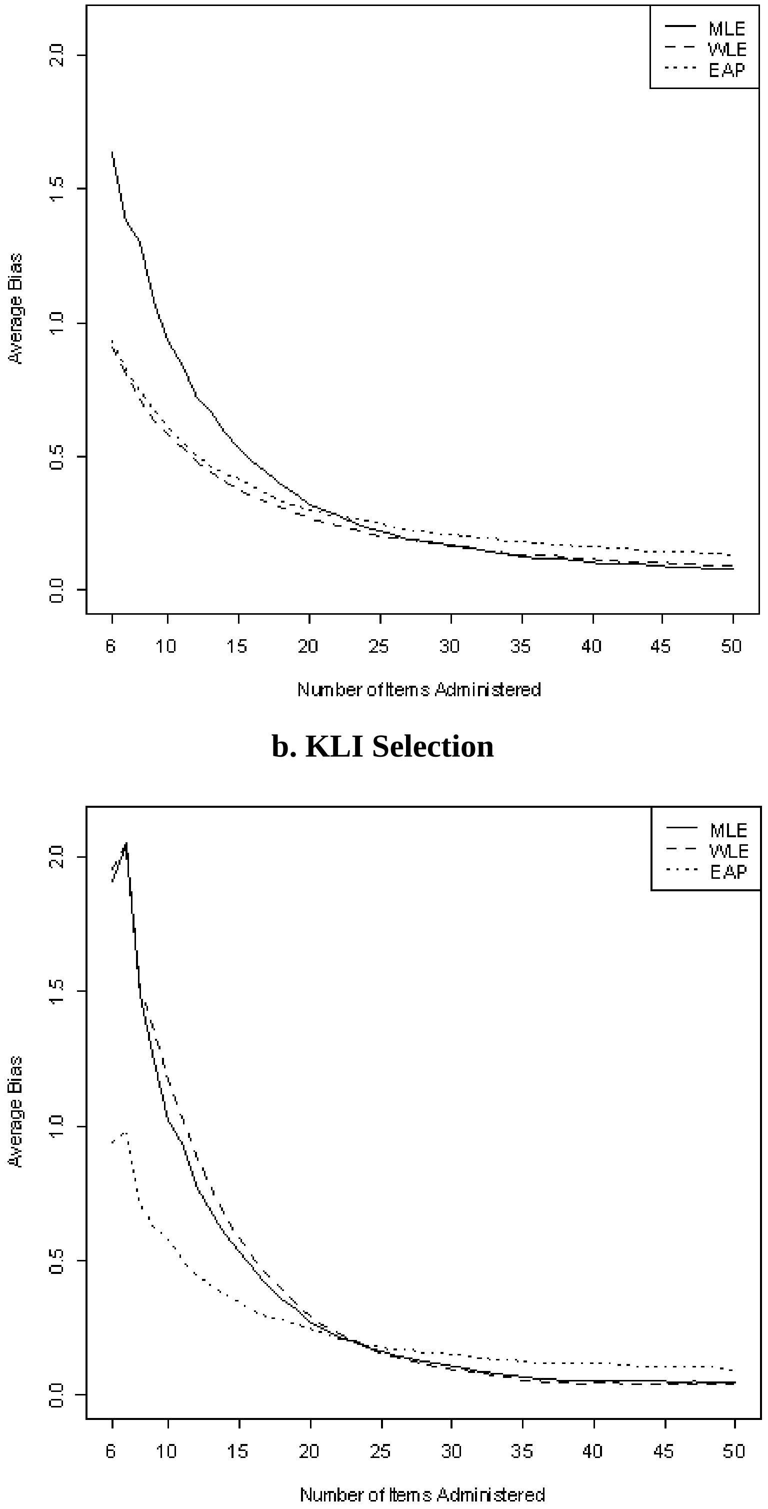
RMSE. As the RMSE is a linear combination of bias and SE, the result seen in Figure 4 that EAP provided the lowest RMSEs was expected. Guyer (2008) found that due to the increased bias, EAP estimation provided larger RMSEs than MLE or WLE when $\theta \leq -2$.

Table 2. Summary Statistics for the *θ* Estimates for the 2-Item Misfit Condition After Different Test Lengths for MCR at *θ* = −1 and MIR at *θ* = 1

θ Estimation	FI			KLI		
and Test Length	Bias	SE	RMSE	Bias	SE	RMSE
θ = −1 (MCR)						
MLE						
15	.534	.779	.944	.535	.900	1.047
25	.222	.487	.535	.167	.519	.545
35	.129	.349	.372	.069	.358	.364
50	.082	.268	.280	.050	.274	.279
WLE						
15	.376	.538	.656	.586	.852	1.034
25	.206	.397	.447	.159	.525	.549
35	.137	.313	.342	.057	.365	.369
50	.097	.257	.275	.040	.276	.279
EAP						
15	.416	.444	.609	.347	.486	.597
25	.249	.347	.427	.181	.340	.386
35	.184	.292	.345	.127	.286	.313
50	.137	.238	.274	.099	.244	.263
θ = 1 (MIR)						
MLE						
15	−1.392	.187	1.405	−1.087	.273	1.121
25	−.742	.316	.807	−.574	.329	.661
35	−.465	.308	.558	−.363	.311	.478
50	−.298	.265	.399	−.247	.263	.361
WLE						
15	−1.469	.183	1.481	−.841	.266	.883
25	−.811	.306	.867	−.448	.314	.547
35	−.502	.308	.589	−.295	.296	.418
50	−.314	.265	.411	−.208	.252	.327
EAP						
15	−.943	.264	.979	−.754	.277	.804
25	−.582	.285	.649	−.483	.280	.558
35	−.424	.268	.502	−.351	.264	.439
50	−.305	.238	.387	−.262	.235	.352

Figure 2. Average Bias Across CAT Lengths for the 2-Item Misfit Condition for $\theta = -1$ (MCR)

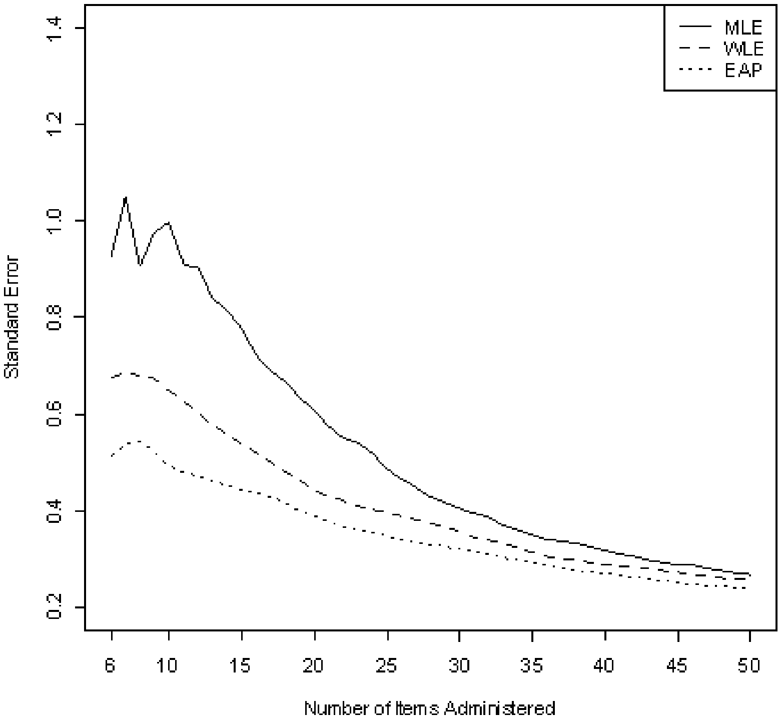
a. FI Selection



b. KLI Selection

Figure 3. Empirical SE Across CAT Lengths for the 2-Item Misfit Condition for $\theta = -1$ (MCR)

a. FI Selection



b. KLI Selection

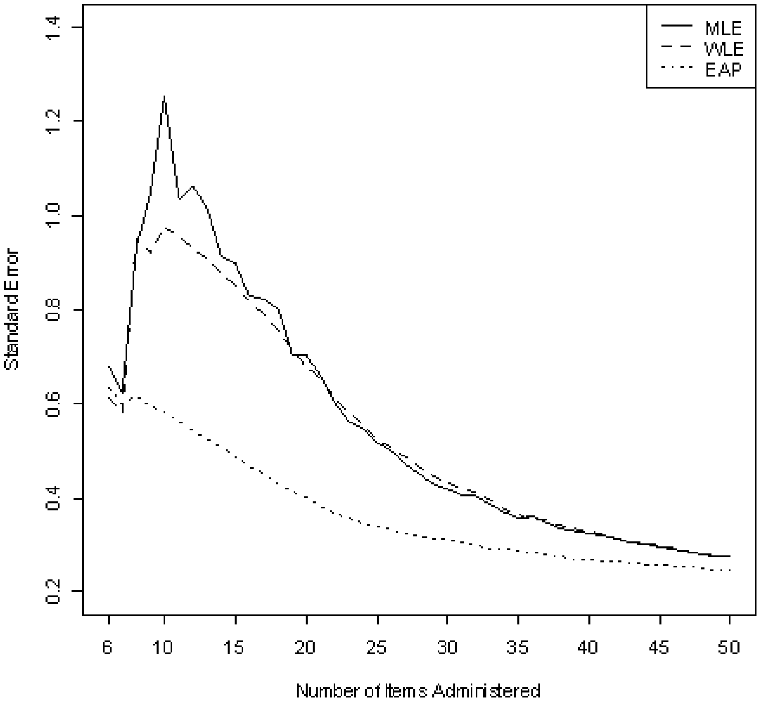
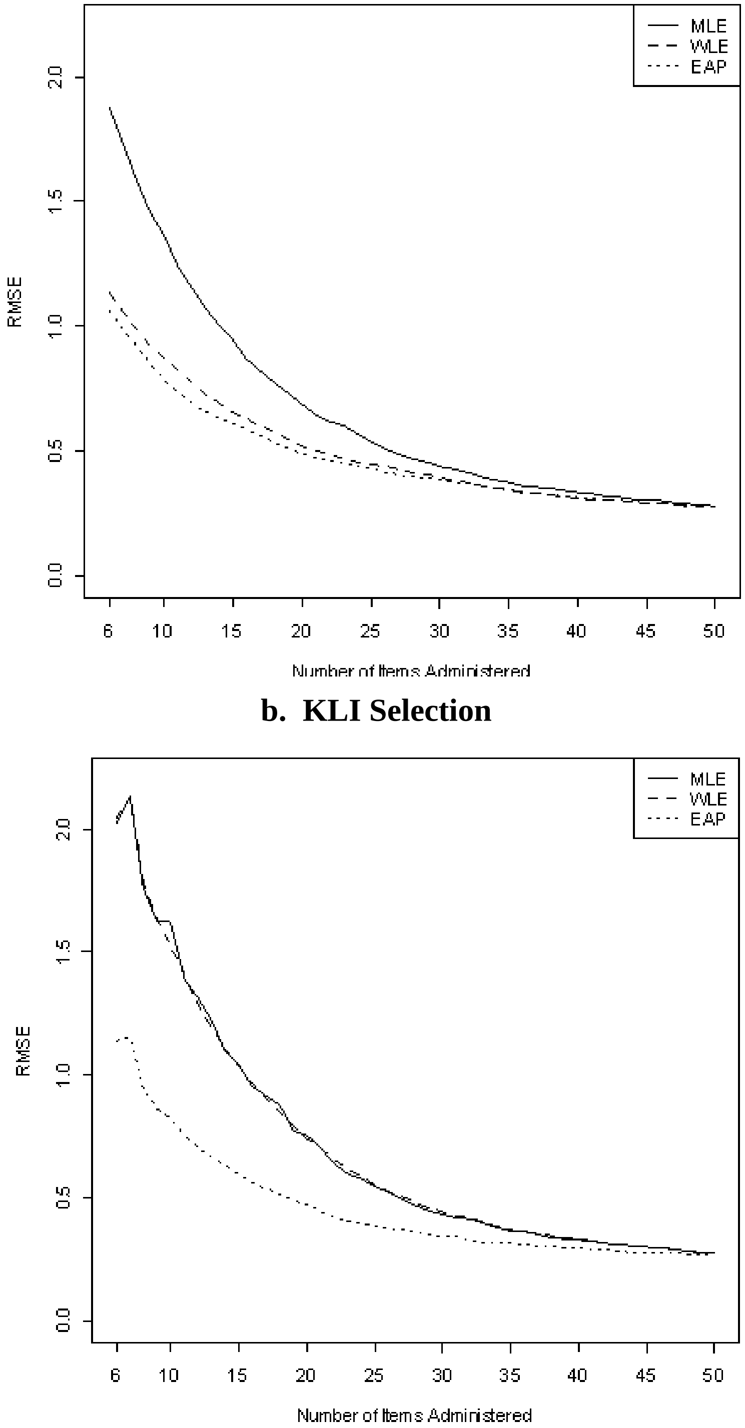


Figure 4. RMSE Across CAT Lengths for the 2-Item Misfit Condition for $\theta = -1$ (MCR)



a. FI Selection

b. KLI Selection

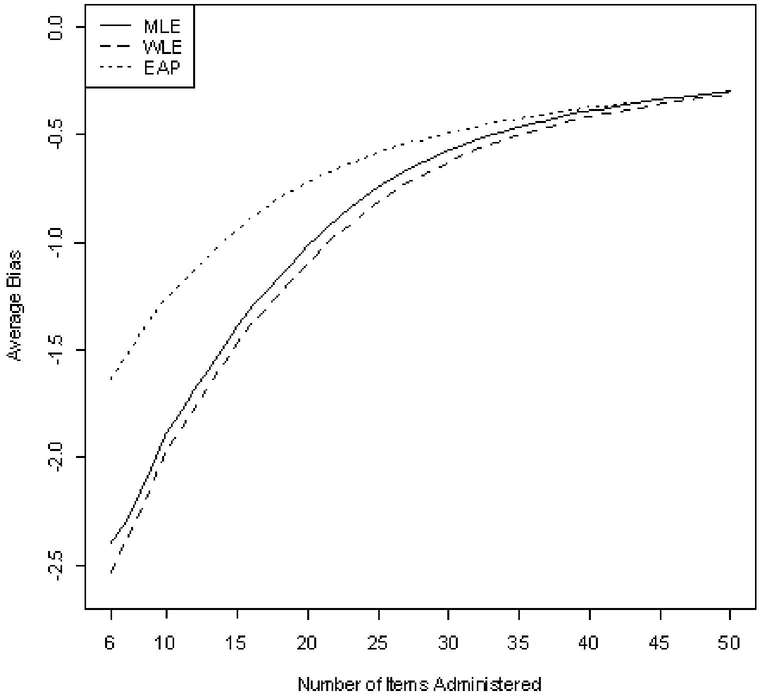
MIR

Bias. As shown in Table 2 and Figure 5, the θ estimates remained biased after 50 items for the 2-item misfit condition when $\theta = 1$. Guyer (2008) found that increasing both the number of misfitting items and θ resulted in increased bias. It can be seen that EAP estimation resulted in less biased θ estimates than MLE or WLE for at least the first 15 items in the CAT. This result was tempered by the effect of item selection method, as WLE θ estimates recovered better when KLI (Figure 5b) was used than FI (Figure 5a).

SE and RMSE. It can be seen in Table 3 and Figure 6 that the SEs of the θ estimates increased until 25 items were administered. Unlike with MCR, EAP had larger SEs than MLE for the first 20 items in the CAT. As the RMSEs were almost entirely comprised of bias, the same trends in the bias are evident in the RMSEs (Figure 7).

Figure 5. Average Bias Across CAT Lengths for the 2-Item Misfit Condition for $\theta = 1$ (MIR)

a. FI Selection



b. KLI Selection

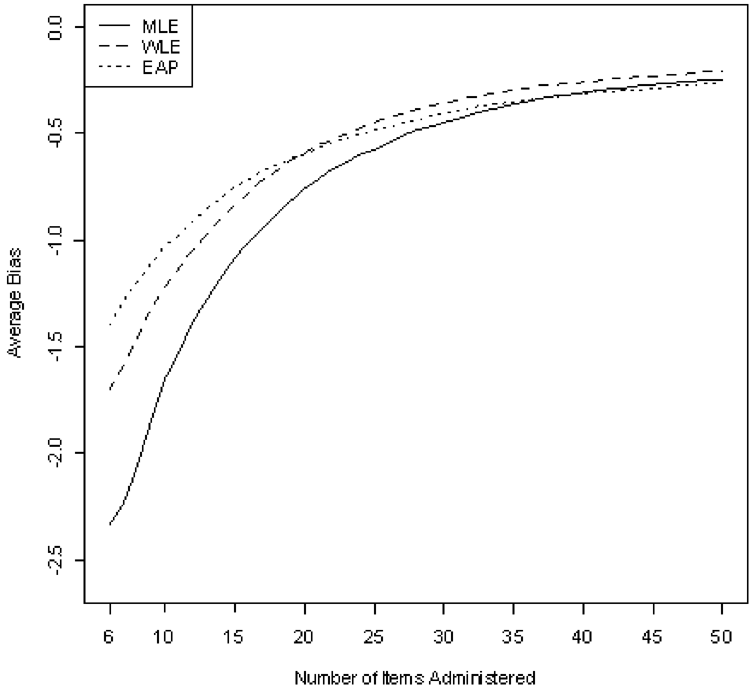
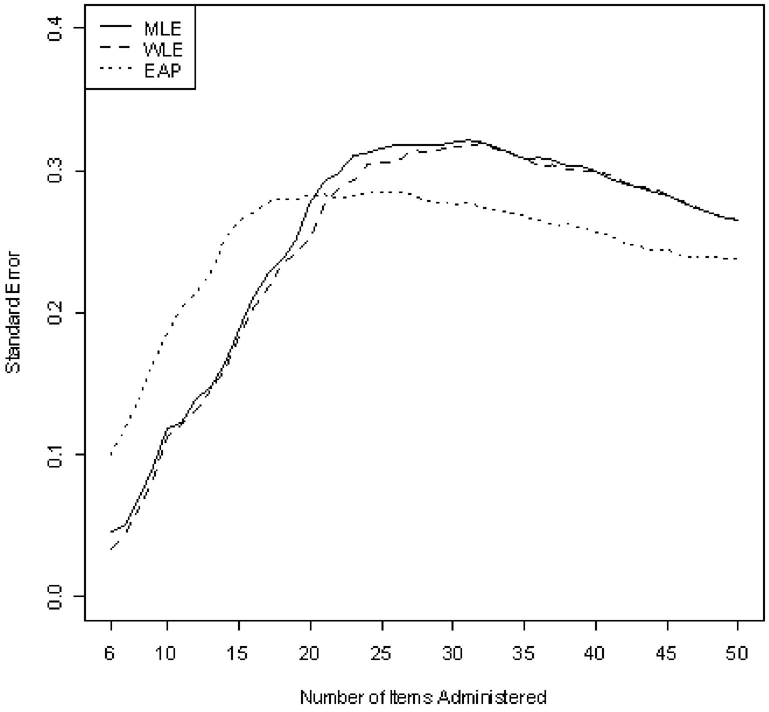


Figure 6. Empirical SE Across CAT Lengths for the 2-Item Misfit Condition for $\theta = 1$ (MIR)

a. FI Selection



b. KLI Selection

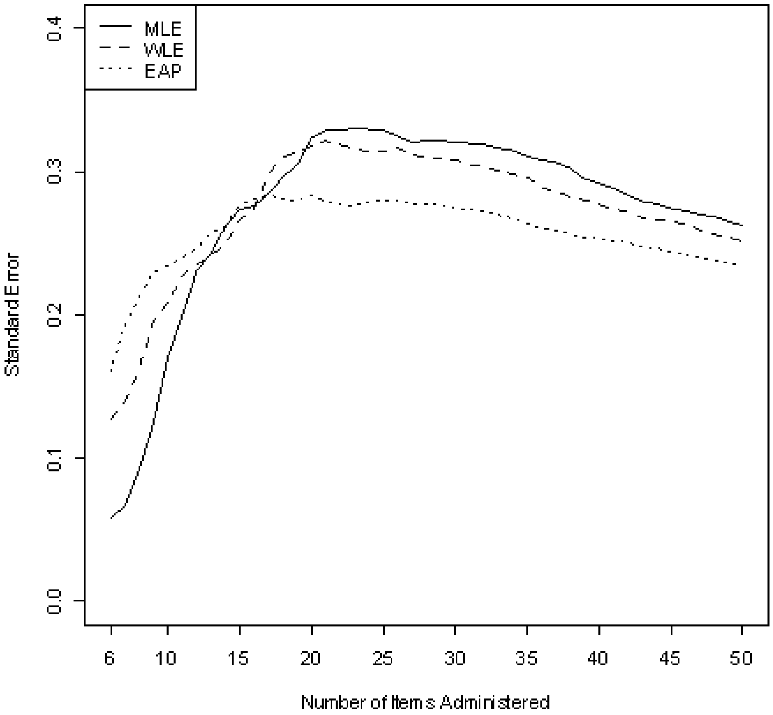
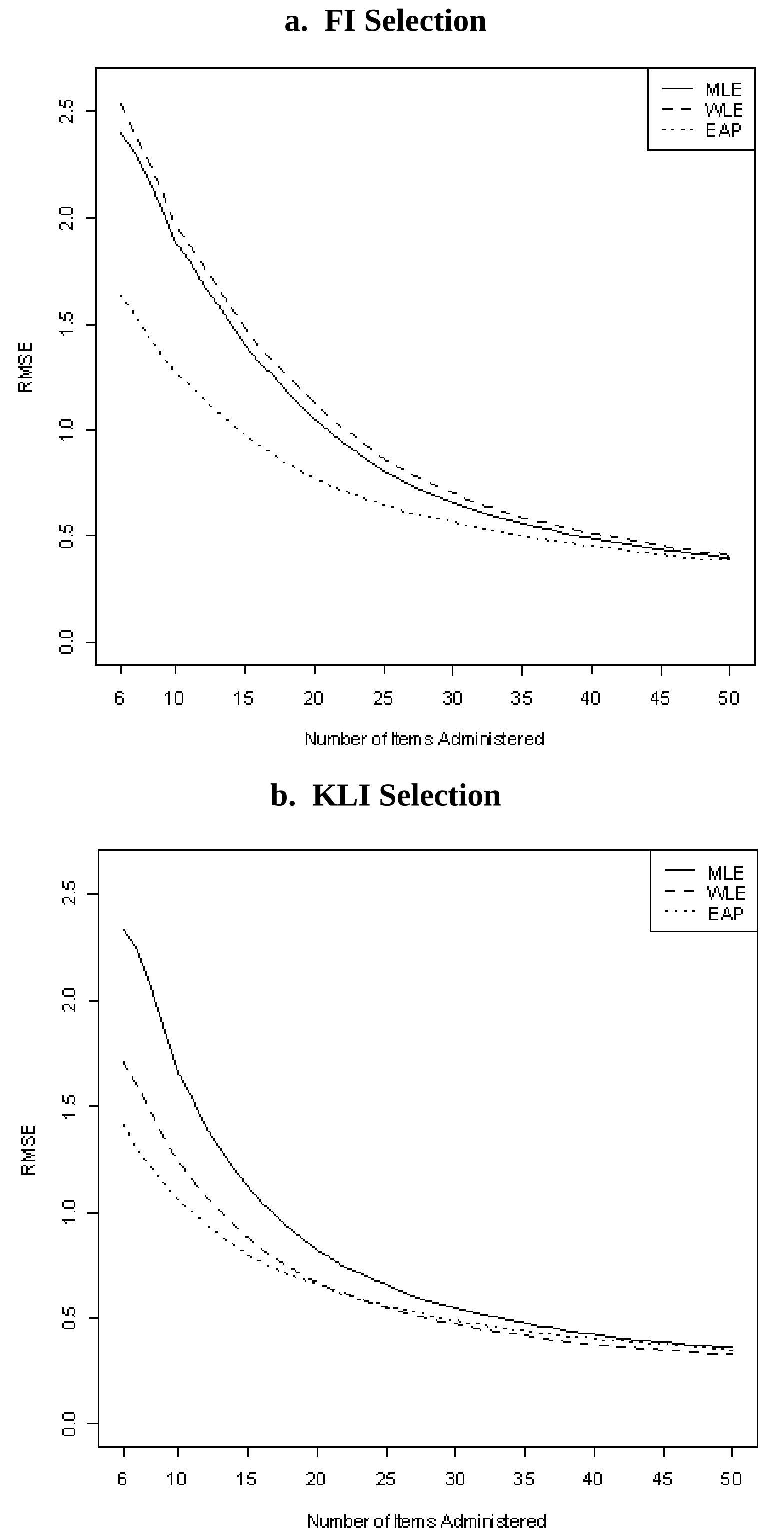


Figure 7. RMSE Across CAT Lengths
for the 2-Item Misfit Condition for θ = 1 (MIR)



a. FI Selection

b. KLI Selection

Discussion

Misfit

The results of this study indicated that CAT with the 3PL model could not recover from MIR within a test length of 50 items. These results suggest that an examinee with $\theta = 1$ who responds incorrectly to the first item in the CAT would not obtain an unbiased θ estimate even after 50 items were administered.

It was found that CAT could recover from MCR, given a sufficient test length. It took over 35 items for CAT to result in θ estimates with bias less than 0.1 when there was 2-item misfit and $\theta = -1$. These results suggest caution on the part of test developers who develop multiple-choice-based CATs with lengths less than 20 items. This study found that successful guessing early in the CAT resulted in biased θ estimates for such test lengths.

θ Estimation

There was evidence that EAP provided better recovery of θ than MLE or WLE, especially early in the CAT. This finding can, in part, be attributed to the regression to the prior mean of 0. The prior also prevented the θ estimates from destabilizing when there was MCR (higher SEs) due to a flattened likelihood, a problem that plagued MLE.

The increase in SEs for the first 20 items observed in the MIR condition can be explained by considering the conditional probabilities of a correct response. The items selected in the CAT initially have low difficulties relative to the generating θ when there was MIR. As a consequence, the probability of a correct response would be near 1.0. These extreme model-predicted probabilities reduced the variation possible in the monte-carlo simulation. As the bias in θ decreased, the difference between the items that were selected and θ also decreased. This resulted in increased variation in θ as test length was increased due to the model-based probabilities decreasing. This phenomenon was not observed for the MCR conditions due to guessing. In addition, EAP had larger SEs early in the CAT because the EAP estimates were less negative than their MLE or WLE counterparts.

It was evident that WLE was sensitive to item selection method. Guyer (2008) found this sensitivity resulted from a more difficult initial item being selected for KLI selection than FI. When KLI was used, WLE provided lower RMSEs than MLE when there was MIR, and larger RMSEs than MLE when there was MCR.

Item Selection Method

The results of this study did not provide evidence that KLI selection improved the recovery of θ when there was misfit. Apart from the results for WLE, KLI selection did not have much effect on the bias and SE of the θ estimates. There was evidence from the MCR conditions that KLI selection increased the bias and SEs of the θ estimates.

Considerations for Future Research

4PL Model

Rulison and Loken (2009) found that CAT could recover from MIR when a 4PL model with a *d* of .98 was used. Additional research is needed to determine the size of *d* parameter sufficient to reduce bias to near zero under various testing conditions. The effect of various *d* parameters on the bias and SE of θ for examinees who do fit the model also needs to be examined.

Base rates. One consideration for use of the 4PL model is the base rate at which MIR occurs in the population. This raises the question: How often does a high-ability examinee miss item(s) early in the CAT? A whole host of factors can be conceived (e.g., mood, testing environment, computer literacy) which would result in MIR. The practical question is how often such factors result in MIR. Information about base rates must be obtained to provide justification for the 4PL model as an acceptable alternative to the 3PL for achievement-type data, particularly if its use has as yet unknown effects for examinees whose responses do fit the 3PL model.

Use of Bayesian Estimation

It was found that EAP provided less biased θ estimates than MLE when there was a sufficient amount of misfit. This study used a fixed-increment method for handling non-mixed response patterns for MLE. It was found that EAP performed better than MLE, in part, because it regressed the initial θ estimates toward 0. As a result, future research should examine whether use of EAP for non-mixed response patterns early in a CAT would result in improved recovery of θ for MLE than a fixed-incremental approach, especially when there is the possibility of misfitting item responses to those items for high-ability examinees.

Limitations of the Present Study

The current study used an item bank that produced an essentially flat bank information function for much of the θ range. The rationale for this decision was to minimize the effect of the item bank on the results. However, applied CATs will not have the same ideal item bank that was used for this study. The findings for extreme θ (±3) might not generalize for real item banks, due to there being fewer items with extreme *b* parameters in real item banks (Chen & Ankenmann, 2004). Thus, future research needs to investigate the effect of misfit across θ for a real item bank.

This study limited the introduction of misfit to the first *k* items in the CAT. The purpose was to introduce a worst case scenario of misfit. Additional research is necessary to examine the effect of misfit at different stages of the CAT.

References

Baker, F. B., & Kim, S-H. (2000). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model.* Research Report 81-20. Office of Naval Research, Arlington, VA.

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

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

Chen, S-Y., Ankenmann, R. D., & Chang, H-H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24,* 241-255.

Chen, S-Y., & Ankenmann, R. D. (2004). Effects of practical constraints of item selection rules at the early stages of computerized adaptive testing. *Journal of Educational Measurement, 41*, 149-174.

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

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum Associates.

Guyer, R. D. (2008). *Effect of early misfit in computerized adaptive testing on the rcovery of theta.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, MN.

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

Lord, F. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48,* 233-246.

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

Lord, F. M. & Novick, M. R. (2008). *Statistical theories of mental test scores.* Charlotte, NC: IAP.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8,* 164-184.

Rulison, K. L. & Loken, E. (2009). I've fallen and I can't get up: Can high-abilitystudents recover from early mistakes in CAT? *Applied Psycholgical Measurement, 33,* 83-101.

Waller, N. G., & Reise, S. P. (2009). Measuring psychopathology with non-standard IRT models: Fitting the four parameter model to the MMPI. In S. E. Embretson (Ed.). *Measuring psychological constructs with model-based approaches.* Washington, D.C.: American Psychological Association Books.

Tang, K. L. (1996, April). *A comparison of the traditional FI method and the global information method in CAT item selection.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New York, N.Y.

The R Development Core Team. (2007). *R: A language and environment for statistical computing.* [Computer software and manual.]

van der Linden, W. J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika, 63,* 201-216.

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

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

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

Yi, Q., Wang, T., & Ban, J-C. (2001). Effects of scale transformation and test-termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement*, *38,* 267-292.