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

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL

OF THE UNIVERSITY OF MINNESOTA

BY

Rick D. Guyer

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

December 2008

© Rick D. Guyer, December 2008

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 this study. 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 were provided.

Table of Contents

List of Tables

List of Figures

Chapter 1:

REVIEW OF PAST RESEARCH

The quality of our measurement is fundamentally important to the quality of our inferences. If our measurements are not consistent (i.e., reliable) over time we can attribute any inferences to the time the individual was measured and not to the variables we are measuring. For at least this reason, the precision of measurement is fundamentally important to the appropriateness of our inferences. Over the years, many methods have been proposed to improve the precision of measurements in psychology. The focus of this review is dedicated to one such method – computerized adaptive testing (CAT).

There are several general properties of CAT that are important to consider (Weiss, 1982):

1) The starting point can be varied by the test administrator.

2) Items are scored during the test administration process.

3) Examinee performance is assessed during the testing process.

4) The item selection procedure is based on the performance of the examinee.

5) A pre-specified criterion is used to terminate the test.

As seen above, 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 during the testing process. In a CAT, items are selected with difficulties similar to the ability of the examinee taking the test. This is in contrast to a conventional test in which the test items are determined before the test is administered and each examinee receives the same set of items in the same order, independent of

performance.

Item Response Theory Models

In 1968, Birnbaum contributed four chapters to Lord and Novick's (1968, 2008) book on psychological measurement. In these chapters, Birnbaum introduced the psychometric community to the three-parameter logistic item response theory (IRT) model. The three-parameter IRT model is commonly used in adaptive testing, so it is important to consider its properties.

The equation for the probability of answering an item in the keyed direction (called the item response function or IRF) for the three-parameter logistic model (3PL) is

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

The probability of correctly endorsing item i ($u_i = 1$) conditional on the latent trait θ for person j (P_{ij}) is a function of both θ and the a, b, c parameters. 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. The a parameter is proportional to the slope of the IRF at the location on θ where P_{ij} equals .5 + (c / 2), and is where the slope of the IRF is at its maximum. D equals 1.702 and is a constant used in the logistic model to approximate the normal ogive function. The b parameter is the item difficulty parameter and is the location on the θ continuum where the probability of a correct response equals .5 + (c / 2).

The two-parameter logistic model can be obtained from Equation 1 by fixing c to 0. This ensures that the IRF asymptotes to 0 and not c. As a result of setting c to 0, the b parameter can be found where the probability of a correct response equals .5. The Rasch

(one-parameter) model results when the item discrimination parameters for all items are constrained to be equal. Commonly an *a* parameter of 1.0 in the logistic metric (D = 1.0) is used in the Rasch model (Embretson and Reise, 2000). Items in the Rasch model only differ only in terms of their *b* or difficulty parameters.

Local independence. The assumption in IRT of local independence states that items are independent when the item parameters and person parameters are taken into account (Embretson & Reise, 2000). This means that the item and person parameters can account for all of the intercorrelations among a set of items. Thus, each item is not correlated with any other item in the test conditional on the item and person parameters. The assumption of local independence is necessary in order to estimate θ, as the IRFs are multiplied together to obtain a likelihood. If local independence does not hold, then the multiplication of probabilities would not be justified.

IRT-Based CAT

In order to implement a CAT using IRT, the following components must be specified. First, an item bank appropriate for the measurement objective must be developed. Next, an initial θ estimate must be specified for the adaptive test. Then, a θ estimation method must be selected. Once the θ estimation method is selected, an item selection method must be specified. Finally, the test developer must decide on a set of criteria for terminating the CAT.

Information in adaptive testing. In order to define the psychometric properties of a good item bank, it is necessary to first consider how IRT defines precision at the item level. Because the IRF for the 3PL is a function of four parameters, the IRF must be considered as a whole in order to index the precision of a given item. Fisher information

(FI), or expected information, 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

$P_i'^2$ = the squared first derivative of the IRF for item *i*, and

Q_i = $1 - P_i$.

The item information functions can be summed across items to obtain the test information function (TIF). Test information indexes the total amount of measurement precision for a test conditional on θ. The test information function was defined by Lord (1977) as

$$I(\theta) = \sum_i^n I_i(\theta), \tag{3}$$

where *n* equals the number of items in the test.

Equiprecise Measurement

Before the development of an item bank, the goal of the measurement process must first be considered. If the goal of adaptive testing is to provide each examinee with θ estimates of equal precision, then the test information function should be high and constant across θ. The property of equal measurement precision across θ is known as equiprecise measurement (Weiss, 2004). To obtain equiprecise measurements, the test developer needs a number of items with *b* parameters that span the θ continuum, yet provide an acceptable amount of psychometric information.

Methods for Implementing a CAT

The Item Bank

The bank information function (BIF) is the sum of the item information functions for each of the items in the test bank – it is analogous to the TIF in a conventional test. Once the target BIF has been determined based on the goals of the CAT (e.g., equiprecise measurement), the next step is to develop an item bank that can support the CAT . The target BIF is determined by both the amount of measurement precision desired and the area on the θ continuum that is to be measured. To develop an item bank, the test developer first must write a series of items and then administer them to a calibration sample. Item parameters are estimated for the calibration sample using the selected IRT model. If items have undesirable psychometric properties (e.g., low discrimination) they can be discarded and new items are then written. The goal for an equiprecise CAT is a bank with a large number of items with high item discrimination parameters and a rectangular distribution of *b* parameters (Flaugher, 2000), which can result in a BIF that is high and flat.

Maximum Likelihood-Based Estimation of θ

The fundamental unit in θ estimation is the IRF. The IRF is a function of all of the item parameters and expresses the probability of a correct response given θ. When an examinee encounters an item, they either provide a keyed response (scored 1) or a non-keyed response (scored 0). Information from the item responses and the IRFs is used in the estimation of θ. In order to estimate θ in CAT, the item parameters are assumed to be known.

Maximum Likelihood

The goal of maximum likelihood θ estimation is to find an estimate of θ that maximizes the likelihood of observing the response pattern given the items administered. Therefore, the likelihood of a response pattern given θ is a function of the response pattern and the IRFs. The item parameters are estimated for the item bank before administration of the CAT. The log-likelihood (*LL*) function was defined by Baker and Kim (2000, p.66) as

$$LL(\mathbf{u}_j \mid \theta, \xi) = \sum_{i=1}^{n} u_{ij} \log[P_{ij}(\theta)] + (1 - u_{ij}) \log[Q_{ij}(\theta)], \quad (4)$$

where

$\mathbf{u}_j$ = response pattern for person *j*,

ξ = the item parameters for the administered item(s),

n = the number of items that have been administered,

P_{ij} = probability of a keyed response,

$Q_{ij} = 1 - P_{ij}$, or the probability of a non-keyed response, and

u_{ij} = item response.

In order to locate the maximum of the log likelihood, an iterative procedure such as Newton-Raphson must be used (Embretson & Reise, 2000). The Newton-Raphson procedure is used to locate where on the θ continuum the first derivative of the log-likelihood is zero.

Implementation of the Newton-Raphson method requires calculation of the first and second derivatives conditional on θ. An initial θ estimate is needed to complete the first iteration. Commonly, a starting value of 0 is used during the first iteration. The ratio of the first derivative to the second derivative (Hessian) is used to update the θ estimate ($\hat{\theta}$) for the i^{th} iteration as shown by

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{\partial(LL)/\partial(\hat{\theta})}{\partial^2(LL)/\partial(\hat{\theta}^2)}. \quad (5)$$

The procedure continues the iterative process until the ratio of the derivatives is less than a pre-defined criterion (Baker & Kim, 2000). When the criterion is met, the Newton-Raphson procedure is said to have converged to the maximum of the function. This $\hat{\theta}$ is the maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$.

The standard error for MLE. The theoretical standard error of $\hat{\theta}_{\text{MLE}}$ is defined as

$$SE(\hat{\theta}_{MLE}) = \frac{1}{\sqrt{-\partial^2(LL)/\partial(\hat{\theta}^2)}}. \quad (6)$$

As the standard error is an inverse function of the second derivative, larger second derivatives indicate greater precision for $\hat{\theta}_{\text{MLE}}$.

Non-mixed response patterns. In the early stages of a CAT, the examinee will not always have both a 0 and a 1 in his/her response pattern. This is known as a non-mixed response pattern. When the response pattern is non-mixed, the likelihood function will still be a monotonically increasing (or decreasing) function like the IRF. This poses a problem for the estimation of θ as the maximum of this likelihood function will be located at either $-\infty$ or $+\infty$ depending on whether the response(s) are 0s or 1s.

The test developer must decide how to handle non-mixed response patterns. One method is to assign an arbitrary value for θ until a mixed response pattern is obtained. An alternative is to use Bayesian estimation until there is a mixed response pattern (van der Linden & Pashley, 2000). As will be discussed shortly, Bayesian methods can obtain a finite $\hat{\theta}$ even for non-mixed response patterns.

Weighted Maximum Likelihood

Bias of maximum likelihood. The estimator $\hat{\theta}_{\text{MLE}}$ is a biased estimator of θ when the expected value of $\hat{\theta}_{\text{MLE}}$ does not equal θ (Lord, 1983). Thus, bias can be defined as

$$BIAS(\hat{\theta}_{MLE}) = E(\hat{\theta}_{MLE} - \theta). \quad (7)$$

Lord (1983) indicated that MLE was unbiased in the limit – as the number of items approaches infinity. In applied testing circumstances, *n* (the number of items) is often quite small. In addition, the estimation of θ assumes the item parameters are known, which is not the case for any applied testing circumstance when item parameters are estimated. Any bias in the item parameters will cause θ to become biased also. Thus, the asymptotic property of non-zero bias will not hold. The first order bias for θ using the 3PL was derived by Lord using a Taylor series approximation and is defined by

$$BIAS_1(\hat{\theta}_{MLE}) = \frac{1}{I(\theta)^2}\sum_{i=1}^{n}[a_i \mathrm{D}]I_i(\theta)(\phi_i - .5), \quad (8)$$

where

$$\phi_i = \frac{P_i - c_i}{1 - c_i}. \quad (9)$$

As shown by Equation 8, the first-order bias is an inverse function of the square of the TIF. Notice that the $\phi_i - .5$ term determines the sign of the first order bias for an item. If $\hat{\theta}_{\text{MLE}}$ is less than *b* there will be negative bias, and if $\hat{\theta}_{\text{MLE}}$ is greater than *b* there will be positive bias in the estimate. The amount of bias would be quite large if there was little test information at the location on θ where the trait was being estimated (Samejima, 1993). This outward bias (estimates being pulled away from θ) is one limitation of MLE estimation, particularly with small numbers of items.

Estimation of θ. To correct for the first-order bias of the MLE estimator, Warm (1989) proposed an adjustment to the first derivative of the log likelihood. The weighted first derivative (*WFD*) is defined as

$$WFD = \frac{\partial(LL)}{\partial(\theta)} - BIAS_1(\hat{\theta}_{MLE})I(\theta). \quad (10)$$

In Equation 10 the product of the first-order bias function and the TIF is subtracted from the derivative of the log likelihood function. Warm (1989) defined a θ estimate from the modified likelihood function as a weighted likelihood estimate (WLE) of θ.

The same approach to θ estimation is used in WLE as is in MLE. Equation 10 is set equal to zero, and the Newton-Raphson procedure is used to locate $\hat{\theta}_{\text{WLE}}$. The Newton-Raphson procedure for WLE is defined as

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{WFD}{\partial(WFD)/\partial(\hat{\theta})}. \quad (11)$$

Notice that *WFD* is substituted into Equation 11 and the derivative of *WFD* replaces the second derivative of the log likelihood in Equation 5. Since the bias of MLE is subtracted from the derivative of the likelihood, the WLE estimator ($\hat{\theta}_{\text{WLE}}$) is said to be unbiased to the order n^{-1} (Warm, 1989), where n^{-1} is the first order bias.

The standard error. The standard error of $\hat{\theta}_{\text{WLE}}$ is defined as

$$SE(\hat{\theta}_{WLE}) = \frac{1}{\sqrt{-\partial(WFD)/\partial(\hat{\theta})}}. \quad (12)$$

The standard error of $\hat{\theta}_{\text{WLE}}$ equals the reciprocal square root of the negative of the derivative of *WFD*. Because additional information is included in the likelihood equation, the standard error of $\hat{\theta}_{\text{WLE}}$ is less than the standard error for $\hat{\theta}_{\text{MLE}}$. However, the standard

error of $\hat{\theta}_{\text{WLE}}$ will approach the standard error of $\hat{\theta}_{\text{MLE}}$ because the bias function approaches zero as the amount of test information increases.

Non-mixed response patterns. It is possible to obtain a finite θ estimate for a non-mixed response pattern in weighted likelihood estimation. Because the bias function is subtracted from the derivative of the log likelihood, the resulting *WFD* function will cross zero (have a maximum) before θ reaches infinity.

Bayesian θ Estimation

In Bayesian estimation, prior information about the population of θ is introduced into the likelihood equation. As shown by Baker and Kim (2000, p. 192), Bayes' theorem can be used to obtain the posterior distribution,

$$P(\theta \mid \mathbf{u}_j, \boldsymbol{\xi}) = \frac{P(\mathbf{u}_j \mid \theta_j, \boldsymbol{\xi}) g(\theta)}{P(\mathbf{u}_j)} \tag{13}$$

where:

$P(\mathbf{u}_j \mid \theta_j, \boldsymbol{\xi})$ = the log likelihood as defined by Equation 4

$g(\theta)$ = the prior distribution for θ

$P(\mathbf{u}_j)$ = the probability of the response pattern.

The prior is a pre-specified distribution of θ, and is typically a standard normal distribution with a mean of 0 and a standard deviation of 1 (although any prior can be used). The probability of the response pattern is a constant and is ignored during θ estimation.

Non-Mixed Response Patterns

It is possible to obtain a finite θ estimate from the posterior distribution for a response pattern that is not mixed, provided a uniform prior is not used. When a uniform prior is

used, the posterior distribution will have the same shape as the likelihood.

Bayesian Modal Estimation

Bayesian modal (known as Modal a Posteriori or MAP) θ estimation locates the $\hat{\theta}$ that maximizes the likelihood of observing the response pattern given the prior and the item parameters. An iterative procedure, such as Newton–Raphson, is commonly used to locate the maximum of the posterior. When the starting θ value is identified, the first and second derivatives of the posterior are calculated conditional on the starting value. Then, the Newton-Raphson procedure uses the following equation to update $\hat{\theta}$ for the i^{th} iteration:

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \frac{\partial[P(\hat{\theta} \mid \mathbf{u}_j, \boldsymbol{\xi})] \big/ \partial(\hat{\theta})}{\partial^2[P(\hat{\theta} \mid \mathbf{u}_j, \boldsymbol{\xi}_i)] \big/ \partial(\hat{\theta}^2)}. \tag{14}$$

The Newton-Raphson procedure continues until the ratio of the derivatives is smaller than a pre-specified criterion. This estimate is defined as the MAP estimate of θ ($\hat{\theta}_{\text{MAP}}$).

The standard error. The model-based standard error for MAP estimation is a function of the second derivative and is defined as

$$SE(\hat{\theta}_{MAP}) = \frac{1}{\sqrt{-\partial^2(P(\hat{\theta} \mid \mathbf{u}_j, \boldsymbol{\xi})) \big/ \partial(\hat{\theta}^2)}}. \tag{15}$$

As with the other θ estimation methods, the square root of the negative second derivative is the observed standard error.

Expected a Posteriori Estimation

In 1975, Owen proposed a method of estimating θ that used an approximation to the posterior. This approximation of the posterior was necessary due to the limited computing power available during the 1970s. Improved computing capabilities in the 1980s made it

possible to evaluate the full posterior (Bock and Mislevy, 1982). The expected value of the full posterior distribution is computed and equals the EAP estimate of θ ($\hat{\theta}_{\text{EAP}}$). As defined by Bock and Mislevy, the expected value equals

$$\hat{\theta}_{EAP} = \frac{\sum_{k=1}^{q} X_k \ell(X_k \mid \mathbf{u}_j, \boldsymbol{\xi}) W_k(X_k)}{\sum_{k=1}^{q} \ell(X_k \mid \mathbf{u}_j, \boldsymbol{\xi}) W_k(X_k)}, \tag{16}$$

where

k = a given quadrature node,

q = the total number of quadrature nodes,

X_k = one of q quadrature nodes,

$\ell(X_k \mid \mathbf{u}_j, \boldsymbol{\xi})$ = the likelihood evaluated at X_k, and

$W_k(X_k)$ = the quadrature weight for that quadrature node.

In EAP estimation, the quadrature weights equal the probabilities taken from the corresponding location on the prior distribution. If the normal distribution were used, then the weights would equal the area under the normal curve contained between quadrature points (Bock & Mislevy, 1982). To maintain interpretability, the quadrature weights are scaled to have a sum of 1.0, which is the total area under a probability density function. Bock and Mislevy recommended using 80 quadrature weights that span a range of θ from −4 to +4.

The standard error. The standard error for EAP estimation is calculated directly using the posterior distribution. Bock and Mislevy (1982) defined the standard error for EAP as

$$SE(\hat{\theta}_{EAP}) = \left[\frac{\sum_{k=1}^{q} (X_k - \hat{\theta}_{EAP})^2 \ell(X_k \mid \mathbf{u}_j, \boldsymbol{\xi}) W_k(X_k)}{\sum_{k=1}^{q} \ell(X_k \mid \mathbf{u}_j, \boldsymbol{\xi}) W_k(X_k)} \right]^{\frac{1}{2}}. \quad (17)$$

As with $\hat{\theta}_{\text{EAP}}$, no iterative procedures are required to obtain the standard error. Because the posterior distribution is evaluated for both EAP and MAP, the standard error in EAP estimation shares the same properties as the MAP standard error.

Other Bayesian Methods

In an effort to reduce the bias of EAP θ estimates, Wang (1997) proposed an essentially-unbiased procedure for EAP estimation (EU-EAP). Rather than using the standard normal distribution for the prior, Wang proposed use of a beta distribution. Wang argued that the shape of a beta distribution is flexible, and could be used to reduce the bias of the EAP θ estimates. The shape of the normal distribution is determined by its mean and standard deviation, while the shape of the beta distribution can be directly modified. The beta distribution is defined as

$$g(\theta \mid \alpha, \beta, l, u) = \frac{(\theta - l)^{\alpha - 1} (u - \theta)^{\beta - 1}}{B(\alpha, \beta)(u - l)^{\alpha + \beta - 1}}, \quad (18)$$

where

α and β are parameters that control the shape of the beta distribution,

l and u are parameters that control the lower and upper bounds of the distribution, and

B is the beta function.

When α and β are equal, the beta distribution will be symmetric. If α is greater than β the distribution will be negatively skewed, otherwise the distribution will be positively skewed.

The properties of the item bank will affect the beta distribution parameters needed to reduce bias. For this reason, Wang (1997) proposed a post-hoc modification of the parameters from the beta distribution. The set of parameters that provided the greatest reduction of bias would be obtained using a trial-and-error process. This required a number of simulations in order to determine which parameters provided the largest reduction in the bias.

Once an optimal set of beta distribution parameters are found, then the beta distribution is substituted into Equation 13 for the prior to obtain the posterior. Equations 16 and 17 are used to obtain the essentially-unbiased EAP (EU-EAP) θ estimate and its standard error, respectively. The motivation for the EU-EAP estimator was to provide a θ estimation method that had the desirable properties of the EAP method (low standard error) yet did not suffer from inward bias, as is the case for EAP using a standard normal prior (Wang, 1997).

Wang, Hanson, and Lau (1999) generalized Wang's (1997) essentially-unbiased method to MAP estimation (EU-MAP). The EU-EAP and EU-MAP estimation procedures are limited by the post-hoc nature of the modification to θ. For this reason, they will not be considered further .

Item Selection Procedures

Starting Value for $\hat{\theta}$

In order to implement a CAT, an item selection method must be chosen. The test administrator must first specify a starting value for θ in order to select the first item and begin the CAT. The choice of a starting value depends on the testing situation, and whether previous information is known about the examinee.

Maximum Information Item Selection

One method for item selection in CAT is to select the item that provides the maximum Fisher information at $\hat{\theta}$ (Lord, 1977; Samejima, 1977; Weiss, 1982). FI, defined by Equation 2, indexes the amount of measurement precision at a given $\hat{\theta}$. The item that provides maximum FI at $\hat{\theta}$ will provide the greatest increase in test information and the greatest reduction in the standard error (SE) when administered.

Kullback-Leibler Information

Global information. FI item selection selects an item that provides the most Fisher information at $\hat{\theta}$. This process takes into account information at only a fixed point on the θ continuum, as information is a function of the second derivative at $\hat{\theta}$. Chang and Ying (1996) described one limitation to this method: Early in an adaptive test the precision of $\hat{\theta}$ is low and information at just $\hat{\theta}$ does not take into account this imprecision.

One way to take into account this uncertainty is to incorporate information about both generating θ (θ_0) and $\hat{\theta}$ in the item selection process. Generating θ is the θ defined by the researcher in a monte carlo study. In order to do this, we must evaluate the likelihood function using both θ_0 and $\hat{\theta}$. The likelihood ratio test was advocated by Chang and Ying (1996) to test how different θ_0 was from $\hat{\theta}$ given the IRT model. The likelihood-ratio test is used in statistics to determine how disparate two functions are. Thus, it can be used to index differences between two sets of parameters over an interval.

The Kullback-Leibler (K-L) information function is defined as

$$K_i(\hat{\theta}\,||\,\theta_0) = P_i(\theta_0)\log\left[\frac{P_i(\theta_0)}{P_i(\hat{\theta})}\right] + \left[1-P_i(\theta_0)\right]\log\left[\frac{1-P_i(\theta_0)}{1-P_i(\hat{\theta})}\right]. \qquad (19)$$

where

|| denotes that $\hat{\theta}$ is separated from θ_0.

Properties. Kullback-Leibler (K-L) information provides what is defined as global information (Chang & Ying, 1996). What makes K-L information global is the fact that it takes into account uncertainty about θ. Equation 19 is a function, not an index of global information. A global information index was proposed by Chang and Ying and equals

$$K_i(\hat{\theta}_n) = \int_{\hat{\theta}_n - \delta_n}^{\hat{\theta}_n + \delta_n} K_i(\hat{\theta} || \hat{\theta}_n)\, d\hat{\theta}. \quad (20)$$

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 20, $\hat{\theta}_n \pm \delta_n$ replaces $\hat{\theta}$ in the denominator of Equation 19 during the evaluation of the integral. The estimate $\hat{\theta}_n$ is substituted into Equation 19 for θ_0 and is assumed to be fixed during calculation of the integral. The specification of δ_n must take into consideration the fact that $\hat{\theta}$ will stabilize as the number of items increases. Chang and Ying recommended a confidence interval based approach, where

$$\delta_n = \frac{\mathrm{d}}{\sqrt{n}} \quad (21)$$

and d is a constant selected according to a pre-specified coverage probability based on the standard normal distribution. Note that as the number of items increases, the denominator will progressively increase; therefore, the range of the integral will approach zero as n increases. It is important to note that there are an unlimited number of possible values for d. The specification of d depends on the rate of convergence for the integral that is

desired by the researcher.

Bayesian index. It is possible to generalize the K-L index to a Bayesian approach. Chang and Ying (1996) defined K-L information with a posterior (K-L-P) as

$$K_i^B(\hat{\theta}_n) = \int K_i(\hat{\theta}||\hat{\theta}_n)\, p(\hat{\theta} \mid \mathbf{u}_n)\, d\hat{\theta}\,. \qquad (22)$$

This integral is computed over the entire range of θ. In Equation 22, $p(\hat{\theta} \mid \mathbf{u}_n)$ equals the posterior density of the random variable $\hat{\theta}$ after the full set of item responses (**u**). The K-L-P index weights the K-L function by the posterior and takes into account the uncertainty in $\hat{\theta}$

Interval Information Methods

Rather than select an item based on maximum information at a single point, Veerkamp and Berger (1997) suggested integrating over a range of θ. They proposed the Fisher interval information (FII) and likelihood weighted information methods as alternatives to maximum information selection. These procedures are limited to MLE and will not be discussed further here.

Other Bayesian Criteria

Four alternative Bayesian item selection procedures were proposed by van der Linden (1998). These methods take into account the possibility of an examinee responding correctly or incorrectly to candidate items. A Bayesian θ estimation method is required for use of these methods, and as this study focused on methods applicable to both MLE and Bayesian methods, they will not be further considered.

Termination Criteria

Standard Error of θ

The selection of a termination criterion depends on the goals of the test administrator. It

also is in part dependent on the practical constraints on the adaptive testing process. The CAT can be terminated when the standard error of θ falls below a pre-specified criterion. To obtain equiprecise measurement, it is necessary to continue the adaptive test until the error of measurement is reduced to a common criterion, if it is possible given the information structure of the item bank.

Fixed Length

In many applied testing circumstances, an adaptive test is terminated after a fixed number of items have been administered (Weiss, 2004). This is typically done due to the practical concern in high-stakes testing that examinees would take legal action if they felt mistreated by the testing process. This could occur if different examinees received different numbers of items and one examinee felt their low score resulted from receiving fewer items.

One operational problem with fixed-length tests is they do not take into account the precision of the θ estimate before termination of the CAT (Weiss, 2004). If an individual is at a location on θ where there is less bank information, then a fixed-length CAT could provide poorer measurement precision than if they were elsewhere on θ. This compromises the goal of equiprecise measurement.

Research on CAT Methods

Properties of Simulation Studies

Monte Carlo Design

In a monte carlo design, item responses are generated according to the IRT model that is used. Simulees are typically created conditional on θ (i.e., a specified number of simulees are generated at each of a discrete number of points on θ). Item parameters then

are generated according to the researcher's specifications, or an item bank with predetermined item parameters can be used. Once the person and item parameters are generated, the researcher generates a matrix of random numbers drawn from a uniform distribution with a minimum of 0 and maximum of 1 (U[0, 1]). This random number matrix has as many rows as persons and as many columns as items. A probability matrix is generated based on θ and the IRT parameters according to the IRT model. For each cell in the matrix, the random number is compared to the probability to create the dichotomous item response. If the random number is greater than the probability, the response is a 0, while a probability greater than the random number will result in a response of 1.

The focus of the current review is on monte carlo studies. Monte carlo studies enable the researcher to control generating θ and allow for the assessment of θ recovery. This enables researchers to take into account sampling variability in θ estimates obtained at a fixed value of θ.

Recovery of $\boldsymbol{\theta}$

One goal of a monte carlo simulation is to assess how well θ is recovered by the CAT. A number of statistics have been proposed in the CAT literature to index θ recovery. One statistic commonly used in the CAT literature is bias, defined as

$$\text{Bias} = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta)}{N}, \tag{23}$$

where

N = number of simulees in the study.

Bias is averaged across simulees in a simulation study by computing the mean of bias

across those simulees. It is also possible to compute the absolute value of bias and compute the mean of those values across simulees. This is known as mean absolute bias (MAB) in the literature.

A commonly employed alternative to MAB is to compute the squared difference between $\hat{\theta}$ and θ. One such index is the mean squared error (MSE) and is defined by

$$\text{MSE} = \frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2}{N}. \tag{24}$$

The root mean square error (RMSE) is the square root of Equation 24, and has the advantage of being in the same metric as θ. It is defined by

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2}{N}}. \tag{25}$$

The final index is the standard error. It is defined as

$$\text{SE} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \bar{\hat{\theta}})^2}{N}}. \tag{26}$$

The SE equals the standard deviation of $\hat{\theta}$ over the N simulees in the study, and indexes instability in $\hat{\theta}$.

Research on θ Estimation

Bayesian *θ* Estimation

In order to demonstrate the computational benefits of their EAP estimator, Bock and Mislevy (1982) performed a simulation study in which the *a* parameters were in the logistic metric (D = 1.0). Bock and Mislevy used a constant *a* parameter of 1.0 with *c* = .20. The *b* parameters were generated to provide maximum FI at $\hat{\theta}$. Thus, FI item

selection was used in their study. Bock and Mislevy varied the termination criteria for the CAT – a Bayesian standard error (as defined by Equation 17) of either 0.2, 0.3, or 0.4 was used in their study. A total of 100 simulees were generated in increments of 0.2 from $\theta = -3$ to 3. This provided a means for assessing the bias of the estimates conditional on θ.

The results indicated that the bias of the Bayesian estimates was largest at extreme θ values (Bock & Mislevy, 1982). This was an effect of the use of a standard normal prior, and resulted in the regression of θ toward 0. It was also found that the regression toward 0 was largest for the SE = 0.4 condition and lowest for the SE = 0.2 condition. This likely resulted from the increased test length of the SE = 0.2 condition.

A simulation study was performed by Wang and Vispoel (1998) to compare the properties of the Bayesian θ estimation methods and MLE. In their study, Wang and Vispoel used MLE, MAP, and EAP for θ estimation. θ was generated in 0.4 increments from −3.2 to 3.2 resulting in 17 different θs. A total of 100 simulees were generated for each of the θ values. They also varied the starting value of θ. In one condition, the starting value was fixed to 0. In the other condition, for θ between −3.2 and −1.2 a prior mean of −2 was used, for θ between −0.8 and 0.8 a starting θ of 0 was used, and for θ between 1.2 and 3.2 a starting θ of 2 was used.

In addition, the properties of the item bank were varied by use of two different "ideal" item banks and one "realistic" item bank, which each were comprised of 300 items. The first ideal item bank used $a \sim N(1.9, 0.1)$; the second used $a \sim N(1.1, 0.1)$. The b parameters for both banks were equally spaced in the interval −3.2 to 3.2 (Wang & Vispoel, 1998). The c parameter was set to .15 for both banks. The realistic item bank was based on pre-calibrated items used in the Iowa Tests of Educational Development.

The mean (minimum and maximum) equaled: a = 1.149 (0.385, 2.0), b = 0.213 (−2.13, 3.781), and c = .15 (.09, .15) for the realistic item bank.

Two different termination criteria were used by Wang and Vispoel (1998). They used either a fixed number of items (10, 20, 30, 40, or 50 items) or a SE of 0.32 or 0.45 to terminate the CAT. The dependent variables in this study were the average bias, the SE, and the RMSE.

Due to the complexity of their research design, only the main trends from Wang and Vispoel's (1998) study are discussed. They provided graphs of the SE, bias, and RMSE for the four θ estimation methods across θ, using the 30-item fixed-length CAT. Separate graphs were provided for the combinations of different discrimination and test entry conditions (a total of nine graphs). They found that MLE had consistently higher SEs than the Bayesian methods for the realistic item bank, as it did not have enough items at the extremes of θ to provide satisfactory precision.

The graphs for the bias revealed that the Bayesian methods were more biased than MLE (Wang & Vispoel, 1998). Of the four θ methods studied, MLE had the lowest bias for each of the nine conditions. For the realistic item banks MLE performed the poorest of the θ estimation methods. Although the results for MAP and EAP were quite similar, EAP provided slightly lower RMSE across θ than MAP. Because $RMSE^2 = SE^2 + bias^2$ (Wang & Vispoel, 1998) the large bias in EAP and MAP contributed to them having larger RMSEs than MLE for θ greater than 1.2.

Efficacy of the WLE Method

In order to assess the bias of his WLE θ estimation method, Warm (1989) compared MLE, WLE, and MAP θ estimation methods. Warm varied the discrimination of the items

in the item bank. In one condition, *a* was constant at 2.0 for all items, and in the other condition *a* descended from 2.0 for the *i* items using the formula $a_i = (71 - i)/35$. All of the *c* parameters were set equal to .20. Warm generated items that provided maximum information at $\hat{\theta}$.

There were 17 sets of 100 simulees generated in increments of 0.5 from $\theta = -4$ to 4. The CAT was terminated when the TIF equaled 20 for the simulee or when 50 items had been administered. The dependent variables in this study were average bias, the SE, and the MSE.

The results of the declining *a* condition indicated that the average bias of MAP was quite large in absolute value at the extremes of θ. In addition, WLE θ estimates were slightly less biased than MLE estimates, particularly from $\theta = -1.5$ to 4. WLE and MLE had about equal bias from $\theta = -4$ to -1.5 (Warm, 1989).

The results for the SE were inconsistent in the declining *a* condition. This was particularly the case for θ values below -1.5, where the SE became quite large for each of the three methods. The SE of the methods also was unstable between the θ increments of 0.5 used by Warm (1989). More stable results were found for the constant *a* condition. Across θ, MLE resulted in consistently higher SEs than either WLE or MAP. The SEs of WLE and MAP were virtually indistinguishable from -2 to 4 on θ.

For the declining *a* condition, no discernible trend emerged for the MSE across θ. In the constant *a* condition, MLE had a higher MSE than WLE or MAP, particularly in the middle of the θ continuum. This result was a function of the instability of the MSE for MLE across the 0.5 increments used by Warm. For both conditions, the MSE of MAP rapidly increased from -3 to -4 on θ.

Evaluation. The results of Warm's (1989) study revealed that WLE provided slightly less biased estimates than MLE or MAP. However, the results for the other statistics were prone to fluctuation across θ. Perhaps this resulted from use of only 100 simulees for each value of θ. It is possible that sampling error caused the observed fluctuations seen in the plots for the SE and MSE.

Research on Item Selection Procedures

Kullback-Leibler Item Selection

Tang (1996) performed a study to investigate the properties of the K-L information item selection method. Her study used a real and a generated item bank, both with 500 items. Items from the real item bank were taken from the Test of English as a Foreign Language. Items were generated using the 3PL model with the following parameter distributions: a ~ lognormal(0, 0.5), which resulted in a mean of 1.09 and a standard deviation of 0.49, and $b \sim N(0, 2)$. The average of the simulated b parameters was –0.02 with a standard deviation of 1.55. The c parameter was generated using a beta distribution with $\alpha = 4$ and $\beta = 13$, which resulted in an average c of .24. The dependent variables in her study were average bias and the MSE.

The means and standard deviations of the item parameters for the real bank were reported by Tang (1996) to be: a (1.29, 0.44), b (0.17, 0.66), and c (.21, .13). The following procedures were used for the real item bank: θ values were generated at values of −3, −2, −1, 1, 2, and 3 on the θ continuum, with 100 simulees generated for each θ value. The K-L item selection method was used with the following values for δ: $3/\sqrt{n}$, $1/\sqrt{n}$, $1/\exp(0.1n)$, and $3/\exp(0.1n)$. Tang was investigating the effect of varying δ on the performance of the K-L item selection procedure. In addition, FI was used during

item selection. The CAT was terminated after 20 items had been administered. Tang used generating θ rather than estimated θ in the numerator of Equation 19.

The results for the real item bank provided evidence that the K-L method resulted in equal or less bias than the FI method (Tang, 1996). In addition, the δ of $3/\sqrt{n}$ resulted in the least bias of any of the δ values. For θ values of −3, −2, −1 and 2, the $3/\sqrt{n}$ δ value had a lower MSE than the other δs. Tang (1996) indicated that the desirable properties of the $3/\sqrt{n}$ δ value might be due to the fact that it was the slowest to converge to FI.

For the generated item bank, the same θs as the real item bank were used. Tang (1996) indicated that the $3/\sqrt{n}$ δ value worked best for the real data, so it was the only δ value used for the generated items. FI was compared to the K-L index. The test length was fixed to 30 items.

The results for the generated item bank differed conditional on θ. It was observed that K-L had average bias 0.1 units lower than FI until about 10–15 items were administered. The difference in bias dissipated as test length increased. The results for the MSE were similar to the bias, as the difference in MSE between the K-L methods and FI was about 0.2 units after 10 items and essentially 0 after 20 items.

The K-L item selection procedure was evaluated in a number of other simulation studies. Chang and Ying (1996) performed two simulation studies to assess the performance of K-L selection compared to FI item selection. Both studies used MLE for estimation of θ. The dependent variables in their study were average bias and the MSE. In both studies a δ equal to $3/\sqrt{n}$ was used to set the interval for the integral, as Tang (1996) found that it provided the lowest MSE and bias.

In Study 1, 800 items were generated from the 3PL with parameters drawn from a

uniform distribution with $a \sim U[0.5, 2.5]$, $b \sim U[-3.6, 3.6]$, and $c \sim U[.0, .25]$. The first item administered had $a = 1.0$, $b = -6$, and $c = .2$. In order to force a mixed response pattern, the next item(s) were of either increasing or decreasing difficulty depending on whether the response pattern was all 1s or 0s.

There were 1,000 simulees for each of the 1-unit increments from −3 to 3 on θ. The CAT was terminated after 14 items had been administered, as Chang and Ying (1996) were interested in the early stages of the adaptive testing process. However, the results were saved after 5–14 items were administered. The results of Study 1 revealed a consistent improvement over maximum information selection in terms of bias and MSE when the K-L method was used. This result generalized across θ and was found after 5–14 items had been administered.

Study 2 used the same conditions as Study 1, except the θ range for generation was limited to −2 to 2. The starting value for b was also set to 0, while the test length was fixed to 40 items. In addition, 254 items from the Reading Assessment of the 1992 NAEP sample were used as the item bank. The use of θ from –2 to 2 was due to the item bank not containing any items with b parameters greater than 2.5 in absolute value. Chang and Ying used generating θ rather than estimated θ in the numerator of Equation 19.

The results for Study 2 revealed that the reduced bias and SE of the K-L selection procedure diminished as the number of items increased. An examination of the graphs provided by Chang and Ying (1996) indicated that the bias of θ was equal across K-L selection and FI selection after 30 items had been administered. This result generalized across θ.

This finding was expected by Chang and Ying, as K-L information was found by

Tang (1996) to converge to FI after about 30 items had been administered. Given that enough items are administered in the CAT, the informative items near $\hat{\theta}$ will be exhausted, meaning that K-L and FI will yield equivalent results.

Evaluation. The results of the studies by Chang and Ying (1996) and Tang (1996) provided evidence that K-L information can improve recovery of θ compared to FI. The strength of this result is modified by the number of items administered. As the length of the CAT increased, Chang and Ying found that K-L information and Fisher information provided increasingly similar results. It was evident that after about 15 items the two methods yielded similar recovery of θ in terms of bias and MSE.

These results are limited by the fact that generating θ was used in the numerator of the K-L selection equation. As the numerator determines the location over which the integral is to be computed, it follows that use of θ makes the K-L procedure more likely to select items near θ. This provided K-L selection an advantage (θ as known) over FI selection in these studies that would not exist in a real CAT environment (θ is unknown).

Comparisons Between K-L and Other Methods

The studies discussed in this section used estimated θ rather than generating θ in the numerator of Equation 19. The properties of the K-L information item selection method were investigated by Cheng and Liou (2000). In their simulation study, both MLE and WLE were used to estimate θ. FI, optimal b, and K-L item selection procedures were used during their simulation and resulted in a 3 × 2 design.

The items in Cheng and Liou's (2000) study were taken from the NAEP. They used item parameters from 204 NAEP items, which had the following ranges: a: 0.452 to 2.502, b: −2.325 to 3.061, and c: .0 to .373. An initial θ of 0 was used to begin the CAT. A

total of 1,000 simulees were generated for each condition at θ values of −2, −1, 1, and 2. The dependent variables in their study were average bias and the MSE. Test length was fixed to 30 items, although plots were provided that showed bias and MSE after 1 to 30 items were administered.

The results revealed a few trends. First, WLE resulted in θ estimates that were less biased than MLE, particularly in the early stages of the CAT (Cheng & Liou, 2000). The optimal *b* selection procedure resulted in estimates with greater bias and MSEs than FI. The results also indicated that the difference between K-L information and Fisher information became trivial after 10 items had been administered.

Chen, Ankenmann, and Chang (2000) compared five different item selection procedures using a simulation study. They used FI, posterior-weighted information, FII, K-L, and K-L-P item selection procedures. Items were generated from the following distributions: $a \sim N(1, 0.25)$, $b \sim U[-3.6, 3.6]$, and $c \sim U[.0, .30]$. θ was estimated using EAP with a standard normal prior.

The CAT was terminated after 20 items, yet information was retained to compare the different selection methods after each item was administered. θ was generated from −3 to 3 in 1-unit increments for the conditions in this study. A total of 1,000 simulees were generated for each condition in this study. The initial θ estimate was 0.0. The dependent variables were bias, RMSE, and SE.

The results for bias, RMSE, and SE indicated that FII and FI performed more poorly than the other three methods at θs of −3 and −2 (Chen et al., 2000). The selection methods were quite similar at other locations on θ. For θs other than –3 and –2 the differences between these methods were negligible after 10 items were administered. For

θs of –3 and –2, it took 20 items for the different methods to be equivalent. These results were consistent with those found by Chang and Ying (1996).

A simulation study was performed by Chen and Ankenmann (2004) to assess the performance of four different item selection procedures. The researchers used FI, Fisher information with a posterior distribution, K-L-P, and random item selection procedures in their study.

Item parameters were based on 360 items from the ACT Math section. The authors did not provide descriptive statistics for the items; however, most of the *a* parameters fell between .5 and 1.5, most of the *b* parameters fell between −1.2 and 2.5, and most *c* parameters fell between .10 and .35. The initial θ estimate for item selection was 0.0. EAP with a standard normal prior was used in this study to estimate θ. The CAT was terminated after 20 items had been administered.

θ was generated from −3 to 3 in 1-unit increments for each of the 1,000 simulees. The recovery of θ was assessed using average bias and RMSE. These statistics were recorded after each of the 20 items had been administered.

As the number of items increased, the recovery of θ improved. The FI method had slightly higher RMSE values than both the posterior weighted and K-L-P selection methods. However, this difference disappeared after about 10 items had been administered (Chen & Ankenmann, 2004).

The recovery of θ differed across θ values. For the RMSE, the recovery was consistently poor for θ values of −3, −2 and 3. An explanation for this result comes from the distribution of the *b* parameters, as there were very few items with difficulties below −1.2 or above 2.5. The results for bias were very similar to the RMSE, and no graphs of

bias were reported by Chen and Ankenmann (2004).

Evaluation. The results of these three studies revealed a few trends. First, the benefits of interval-based item selection methods are reduced as the number of items are increased. This result generalizes across K-L, K-L-P and the FII item selection procedures. As the recovery of θ was similar after about 10 items have administered, the benefit of interval item selection for real CATs is limited. The results of Chen and Ankenmann (2004) must be also tempered by their use of EAP θ estimation which was biased for non-zero θ.

Interaction Between θ Estimation Method and Termination Criteria

The interaction between θ estimation method and termination method was explored by Yi, Wang, and Ban (2001). Yi et al. used MLE, WLE, EAP, and MAP θ estimation in their study. In addition, two different item banks were used. The first item bank was comprised of the 420 item ACT bank used by Wang (1997). The second item bank was comprised of 420 items with *a* parameters generated from a log-normal distribution with a mean of 1.2 and a standard deviation of 0.4 in non-logarithmic units. The *b* parameters came from a uniform distribution (U[−5,5]), while the *c* parameters came from a normal distribution [N(.15, .05)].

Yi et al. (2001) used FI as the method for item selection. The first termination criterion used by Yi et al. was a fixed-length CAT with 30 items. A fixed standard error criterion of .32 also was used in their study, and the CAT was terminated if the standard error was not reached after 60 items. The third criterion used target information to terminate the CAT. For the real item bank the BIF from the ACT item bank was used to obtain the target information function. For the generated item bank, Yi et al., (2001) obtained the average

of information conditional on θ, and then multiplied the average by 60 to obtain the target information function. This resulted in a target information function that was shaped like a normal distribution for the real item bank, and a uniform function for the generated item bank. When the total information for the CAT exceeded the target (conditional on $\hat{\theta}$), the CAT was terminated. As before, the CAT was terminated if the criterion was not reached after 60 items.

Simulees were generated at one of 21 points on θ using increments of 0.4 from −4 to 4. There were 1,000 simulees for each value of θ. The dependent variables in the study were average bias, SE, and RMSE.

Overall, the bias of EAP and MAP was larger than it was for MLE and WLE (Yi et al., 2001). The amount of bias differed across termination criteria for the real item bank. For the fixed SE termination criterion, MLE actually had smaller bias than WLE. However, WLE was less biased than MLE for the target information criterion. Test length was found to vary substantially across the fixed SE and target information termination conditions.

For the generated item bank, the bias of MLE and WLE were consistent across termination criteria. Yi et al. (2001) also found that the bias of MAP and EAP was somewhat greater than MLE or WLE. Overall, the bias in the θ estimates was greatest under the target information termination rule.

The results for the generated item bank indicated that the Bayesian methods had the lowest SE across θ. MLE had the largest SEs with WLE slightly lower than MLE (Yi et al., 2001). The SEs for the real item bank followed a similar pattern except for the target information condition. In the target information condition, the SEs were much higher

than the other two conditions. In addition, the SE for MLE became large at the low and high ends of the θ continuum.

For the RMSE, Yi et al. (2001) found that the MAP and EAP θ estimators in the real item bank condition showed high RMSE at the extremes of θ. In addition, MLE showed high values for the RMSE at the extremes of θ for the real item bank with a target information termination criterion.

Evaluation. The study by Yi et al. (2001) provided more consistent results than previous studies. This can be attributed to the larger number of simulees generated per condition. Their finding that a SE-based termination criterion resulted in greater bias for WLE than MLE was somewhat surprising. Additional research is needed to see if this result is replicable.

One limitation of their study was the use of termination criteria that systematically altered the length of the CAT for the real item bank condition. Because the target information function at values of θ below −2.5 and above 3.5 was below 1, the CAT would terminate very quickly for those θs. There were not many discriminating items in those intervals of θ, so the CAT required a large number of items to reach the fixed SE termination criterion. For this reason, it is difficult to interpret the differences they found, as the two termination criteria differed greatly in the number of items administered.

Implications from Past Research

Item Banks

The results of previous studies have revealed some considerations for selection of an item bank for an equiprecise CAT. First, it is important to have sufficient numbers of items with *b* parameters at extremes of the θ continuum, or else the estimates will have

large bias and RMSEs (Chen & Ankenmann, 2004; Wang, 1997).

There was an apparent discrepancy between the item banks generated for the simulation studies and the real item banks. The generated item banks were typically designed to have a uniform distribution of *b* parameters that ranged from about −3 to 3. In contrast, many of the real item banks (e.g., Chen & Ankenmann, 2004; Wang, 1997) had few items with *b* parameters larger in absolute value than 2. This suggests that the *b* parameter generation procedure should be modified to better approximate real item banks if the two are compared directly in a simulation study.

Theta Estimation Method

Bias. The results of the simulation studies provided evidence pertaining to the bias of different θ estimation methods. It became evident that Bayesian θ estimation methods will result in inward bias at the extremes of the θ continuum. In addition, MLE has outward bias which is most prevalent when few items have been administered.

Several new methods were proposed to reduce the bias of MLE, EAP, and MAP. The properties of the WLE estimator have been documented in only two studies (Warm, 1989; Yi et al., 2001). In general, Warm and Yi et al. found that WLE provided less severely biased estimates than MLE. The advantage of WLE over MLE in terms of bias decreases as test length increases. Given an informative item bank, WLE will converge to MLE after about 10 items have been administered. This is likely attributable to the adaptive item selection procedure, in which items that provide large amounts of test information are selected, thereby reducing the value of the bias function.

New methods were proposed by Wang (1997) and Wang et al. (1999) to reduce bias in EAP and MAP, respectively. The so-called EU θ estimation methods were successful in

reducing bias. However, this result is attributable to the use of a beta distribution to minimize bias – given the item bank. No research has been done to compare the generality of the beta distribution across item banks.

SE. There has been strong evidence that Bayesian θ estimation methods reduce the SE of the θ estimate compared to MLE. This can be attributed to the additional information (the prior) that is introduced into the likelihood. As the number of items increases, the SE for Bayesian θ estimation methods will approach that of MLE.

The SEs for WLE were quite similar to MLE after about 10 items had been administered. As mentioned earlier, the bias function is close to zero at $\hat{\theta}$ after about 10 items had been administered during the CAT. Thus, the subtraction of the bias function will have a decreasing effect on the likelihood as the number of items increases.

One practical question that has not been well addressed is the validity of the standard normal prior commonly used during Bayesian estimation. If the prior distribution is based on empirical evidence, then any reduction in the SE can be attributed to prior knowledge about the population. It seems somewhat inappropriate to use a standard normal prior without first considering the distribution of θ in the population. If θ were skewed or kurtotic, then the use of a normal prior would further bias the estimates.

Item Selection Procedures

A number of new methods have been proposed to take into account uncertainty in $\hat{\theta}$ during item selection. These include Kullback-Leibler information-based item selection procedures. Simulation studies found that these methods provided improvements in terms of bias and RMSE in the early stages of the CAT. However, the benefits disappeared as the test length increased to 10 items and θ became more accurately estimated.

For this reason, FI provides measurements that are equal in precision to these new methods – provided the adaptive test is at least 10–15 items long. Thus, the benefits of the alternative item selection procedures are limited to very short tests. For longer tests the different methods provide similar results.

Purpose of the Current Study

Misfit in IRT can be defined as item responses that are not likely given the IRT model. Given the convergence of $\hat{\theta}$ to θ as the number of administered items is increased in a CAT, it holds that misfit for the first item(s) in a CAT (early misfit) would result in responses that are less likely given the 3PL than misfit at other stages of the CAT. Contrast this with misfit at the end of the CAT, when $\hat{\theta}$ is more precisely estimated, and where the effect of the misfit on the item selection procedure would be much more inconsequential.

Early misfit would occur if a high ability examinee ($\theta = 3$) responded incorrectly to easy items ($b = -3$). In CAT, early misfit for high ability examinees could occur due to: unfamiliarity with the computer terminal, psychological factors such as nervousness, or environmental factors such as background noise. In the case of the 3PL, early misfit would also occur if low ability examinees guessed correctly on the initial item(s) in the CAT or had prior knowledge of correct answers to items that might appear early in a CAT.

In a study that was published about a year after the present study was begun, Chang and Ying (2008), reported a very limited examination of the effect of early misfit on the recovery of θ. They noted that early in a CAT an imprecise θ estimate is being used to select the next item for administration because few items have been administered. In their

study, they varied the initial θ used to select the first item, as well as the discrimination parameters used in the CAT. They found that as initial θ deviated from θ, the θ estimates became more biased. Chang and Ying concluded that early misfit would likely cause the item selection procedure to select items that are not as appropriate for the examinee. Chang and Ying did not examine how introducing early misfit would affect the θ estimates.

The goals of this study were multi-faceted. First, the effect of early misfit on the recovery of θ was investigated. As shown by the review of previous research, the new methods for item selection and θ estimation did not make a practical difference in the recovery of θ after about 10–15 items had been administered. However, these studies assumed that the examinee responses fit the IRT model. In the present study, the recovery of θ was examined across θ estimation method, item selection procedures, and levels of θ. This study also examined whether WLE, MLE, and EAP θ estimation methods affected the recovery of θ when there was early misfit. In addition, both K-L and FI item selection were used during the item selection procedure to determine if they were differentially affected by early misfit.

Chapter 2:

METHOD

Data Generation

Item Banks

The a parameter distributions were selected to be similar to those obtained for real CAT item banks (e.g., Wang, 1999; Chen & Ankenmann, 2004). Item parameters were generated according to the following distributions: $a \sim$ log-normal(–0.223, 0.2), $b \sim$ 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, and to ensure that the goal of equiprecise measurement was not compromised during the data generation process.

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. For this reason, an item bank with 300 items was used in this study.

Monte Carlo Simulation

The item response data for this study were obtained using monte carlo simulation. The 3PL IRT model defined by Equation 1 was used for item response generation. Then the monte carlo procedure described previously was used for generation of the item response data.

Design

θ

To examine the effect of value of θ on θ recovery, θ values of –3, –2.5, –2, –1, 0, 1, 2, 2.5, and 3 were used in this study. To provide a comprehensive look at the effects of misfit conditional on θ, θ was crossed with the other variables in this study. There were 1,000 simulees generated at each θ for each condition in this study.

Misfitting Items

Starting value for θ. As no prior information about θ was assumed in this study, the first item in the CAT was selected using an initial θ estimate of 0. Specification of a constant initial θ was necessary to obtain consistent operationalization of misfit across the θ continuum.

Number of misfitting items. The number of responses that did not fit the 3PL model was varied from 0 to 4. The zero misfitting items condition served as a null condition to assess the recovery of θ across different θ estimation and item selection methods. For the 1 to 4 item misfit conditions, the direction of misfit differed conditional on θ. Misfit was introduced during the item selection stage of the CAT. This ensured that the item selection process was affected by misfit.

Introducing misfit. As examinees with a θ below 0 are expected by the model to have a probability less than .5 + (c / 2) correctly answering an item with a b of 0, it follows that misfit would result when they answer such an item correctly. Likewise, for an examinee with θ above zero, the model indicates a probability greater than .5 + (c / 2) of them correctly answering an item with a b of 0.

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. Item responses for the first k items in the CAT were designated as incorrect in order to introduce misfit. For examinees with θ below zero, misfit was operationalized as correct responses to the first k items. For the θ = 0.0 condition, misfit was operationalized as either incorrect or correct responses to the first k items. For the misfit-as-correct-responses (MCR) condition, the first k responses were changed to correct. For the misfit-as-incorrect-responses (MIR) condition, the first k responses were changed to incorrect.

θ Estimation

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. In CAT, the item selection

procedure required an estimate of θ after 1 to n items were administered. For this reason it was necessary to specify an alternative to an MLE θ estimate for response patterns that produce a likelihood that had no maximum. As WLE and EAP can obtain finite θ estimates for non-mixed response patterns, no additional specifications were required for these methods.

The Newton-Raphson procedure was used to estimate θ for MLE and WLE. As shown by Equation 5, the Newton-Raphson procedure locates the maximum of the likelihood using an iterative procedure. When the incremental change in $\hat{\theta}$ became less than the criterion of .001, the Newton-Raphson procedure was considered to have converged. As EAP can be estimated in closed form, no iterative procedure was necessary. A total of 80 quadrature points from –4 to 4 on the standard normal distribution were used to estimate θ using EAP.

Alternative to MLE for problematic response patterns. In this study, $\hat{\theta}$ was incremented by -1 for each incorrect response, and $+1$ for each correct response, until $\hat{\theta}$ equaled 4 in absolute value. This procedure was employed until the response pattern became mixed. Thus, an incorrect response to the first item would yield a $\hat{\theta}$ of -1. One additional problem with MLE for the 3PL was that θ cannot be estimated when the proportion of correct responses was less than the lower asymptote of the TRF. When this occurred, the sum of the log of the IRFs does not yield a function with a maximum, but rather an IRF-shaped function. In these situations, $\hat{\theta}$ was fixed to –4 to obtain a finite estimate for use during item selection.

Modifications to the Newton-Raphson procedure. When the response pattern was not mixed, the R program set the θ estimate to a common value as defined above. If the

procedure could not converge for a mixed response pattern due to the likelihood function not having a maximum, the θ estimate was fixed to –4. This only occurred at the low end of θ due to the effect of the c parameter.

In order to locate the maximum of a function, the second derivative must be negative. If the second derivative was positive, the procedure would be iterating toward $-\infty$ due to the inflection point (local minimum) found when there is a lower asymptote (c) parameter in the model. If the second derivative was found to be positive, then the R program reversed the sign of the second derivative to ensure that the procedure moved toward a maximum rather than a minimum. In addition, it would be possible for the increment to become quite large when the second derivative approached zero. For this reason, the incremental change in θ for an iteration was constrained to be no larger than 1.0 in absolute value.

Item Selection

This study investigated whether an interval item selection procedure provided improved recovery of θ compared to a FI fixed-point procedure . The K-L item selection procedure proposed by Chang and Ying (1996) was used as the interval item selection method. 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).

Termination Criterion

A fixed-length CAT that terminated after 50 items were administered was implemented. The item responses, current θ estimate, and the model-based standard errors were saved after each item was administered, for additional follow-up analyses.

Conditions

This study used a 5 (misfitting items)× 3 (θ estimation)× 2 (item selection) × 10 (θ

levels) design. Each independent variable was fully crossed with each other. In total there were 300 cells in this study. To ensure stability in the results, 1,000 simulees were generated for each cell of this design. Item response data were generated independently for each of the 300 cells, to ensure that there was no capitalization on chance for any condition.

Analysis

CAT Simulation Program

As no commercially available software was developed to simulate a CAT using WLE crossed with K-L information item selection, the author wrote a program in R (R Core Development Team, 2007) to implement the CAT. The CAT program was developed to modify item responses in real time to introduce misfit. Given the number of misfitting responses (0 to 4), the program modified item responses directly to introduce misfit. For example, if a high ability simulee were to not-fit-the-model by getting the first two items incorrect, the item responses for the first two items would be set to 0 in real time during the item selection process. This was an additional advantage of the program, as no commercial program has been developed that can introduce such a specialized case of misfit.

To ensure the integrity of the CAT program, results from the author's CAT program were compared to software that estimates θ using EAP, WLE, and MLE for the two-parameter logistic model (Choi, 2007). The results provided evidence that the EAP, MLE, and WLE θ estimates from the author's program were the same as the program developed by Choi.

Dependent Variables

The θ estimates obtained after 15, 25, 35, and 50 items were administered were used for these analyses. The use of θ estimates after different test lengths provided information about how θ recovery changed as more items were administered. Average bias indexes the average deviation of θ from $\hat{\theta}$ and was defined by Equation 23. The empirical SE equals the standard deviation of the distribution of $\hat{\theta}$ and was defined by Equation 26. The RMSE equals the standard deviation of $\hat{\theta}$ about θ and was defined by Equation 25. These analyses provided descriptive information about the recovery of θ, and were useful for comparisons across the different cells in the research design.

ANOVA

Although the average bias, SE, and RMSE index the recovery of θ, they do 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. This was because θ was an independent variable in this study 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-subject factors in the ANOVA. As there was systematic redundancy in the misfitting item condition, that variable was a within-subjects factor in the ANOVA. This redundancy occurred as, for example, the three misfitting item condition shared three items in common with the four misfitting items condition – given that the other factors were held

constant.

Hypothesis testing for this study would not be as informative as an index of effect size due to the number of effects tested. In addition, the large sample size used in this study ensured that every effect would likely be statistically significant. 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,

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}. \quad (27)$$

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.

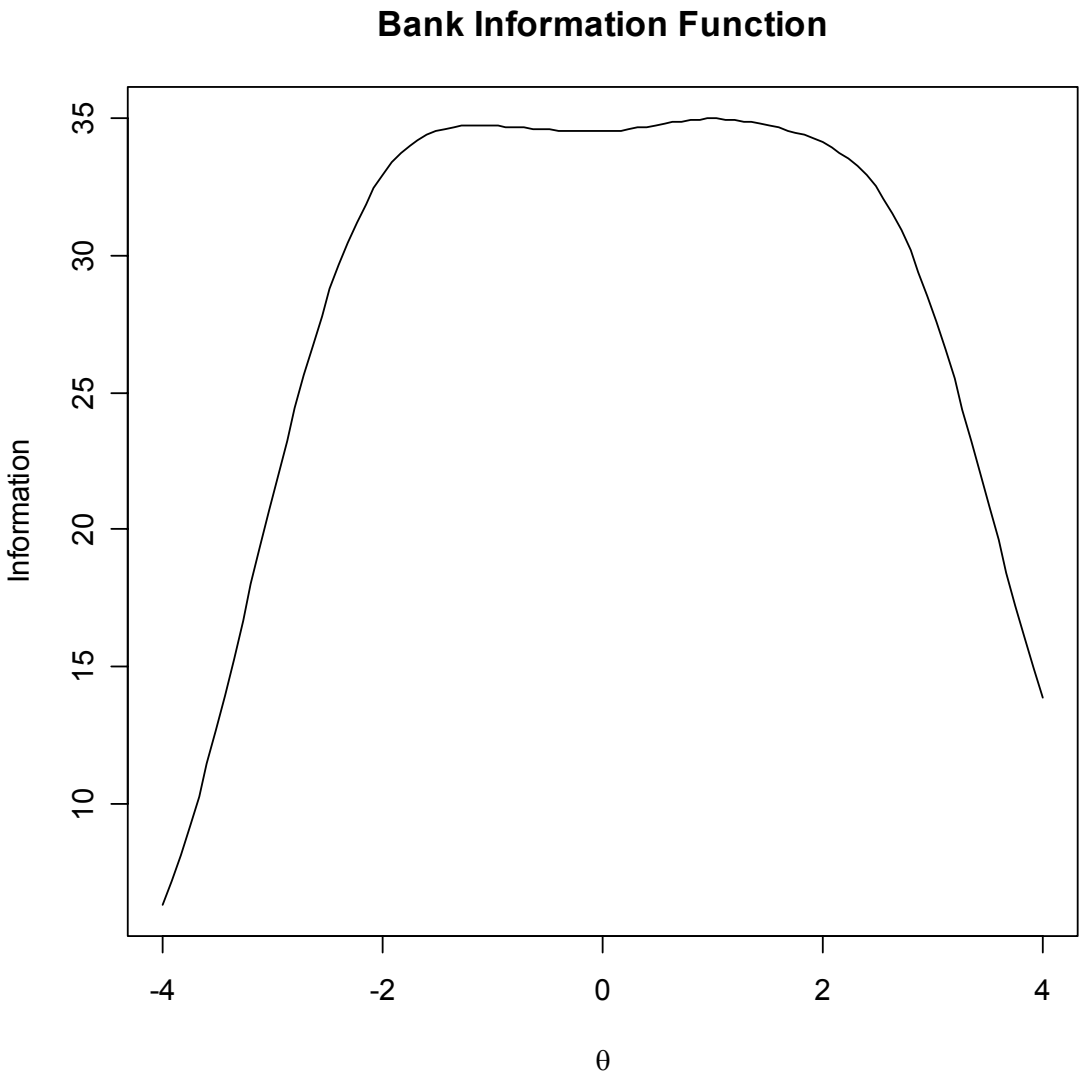
Chapter 3: RESULTS

Item Bank

As b was generated using a uniform distribution, the BIF for this simulation study met the goal of equiprecise measurement, in the range of θ between approximately −2.0 and +2.5, as shown by Figure 1. The c parameter reduced the psychometric information for lower θ, and caused the BIF to be asymmetric. The goal was to obtain a flat BIF, to ensure that differences in information conditional on θ did not lead to large differences in recovery for different θ.

Figure 1

BIF for the Simulation Study



Convergence Failures

It was found that MLE θ estimation failed to converge for certain conditions in this study. To ensure that there were 1,000 simulees within each condition, additional simulees were generated until 1,000 MLE θ estimates converged. If the θ estimate did not converge for a simulee for MLE, then the θ estimation program would exclude the EAP,

MLE, and WLE estimates for that simulee from any additional analyses. This ensured that the same simulees were used for analyses based on EAP, MLE, and WLE θ estimates.

Information about the number of convergence failures is provided by Appendix Table A1. As shown by Table A1, almost all of the convergence failures occurred for low generating values of θ, especially for shorter CATs. The introduction of early misfit increased the number of convergence failures.

Convergence failures occurred most frequently for the 3 misfitting item response MCR conditions. To document this finding, graphs of the *LL* and first and second derivatives for the first 12 items in the CAT for the 3-items-of-misfit MCR (with $\theta = -3$) condition are provided in Figure A1. It can be seen in Figure A1 that the first derivative of the *LL* flattened as the simulee continued to respond incorrectly to items. The *LL* function did not have a maximum after 11 items (Figure A1k), and resulted in a convergence failure. The θ estimate was fixed to –4 in that circumstance.

It was also noteworthy that K-L item selection resulted in a higher number of convergence failures. As shown by Table A1, this result was most striking after 15 items had been administered. For example, for a θ of –2.5 and three misfitting items, K-L selection failed to converge 104 times while FI selection only failed 27 times.

The ANOVA

A 5 (misfitting items)× 3 (θ estimation)× 2 (item selection) × 10 (θ) design ANOVA was performed after 15, 25, 35, and 50 items had been administered in the CAT. The observed unsigned bias values were the dependent variable in the ANOVA. See Appendix Formulas A1–A47 for the formulas used to compute the sums of squares and degrees of

freedom in the ANOVA. The sums of squares, mean squares, degrees of freedom, and η^2 are reported in Tables 1−4 for CATs of 15 to 50 items.

Table 1
Results from the Mixed Design ANOVA After 15 Items were Administered

Type of Effect Source	*SS*	*df*	*MS*	η^2
Between Subjects				
θ	539684.186	9	59964.910	.538
θ est.	354.252	2	177.126	<.001
Selection	3237.887	1	3237.887	.003
$\theta \times \theta$ est.	6227.084	18	345.949	.006
$\theta \times$ Selection	275.441	9	30.605	.003
θ est. × Selection	2258.283	2	1129.142	.002
$\theta \times \theta$ est. × Selection	188.138	18	10.452	<.001
Error	22569.215	59940	.377	
Within Subjects				
Misfit	29839.703	4	7459.926	.030
$\theta \times$ Misfit	286821.417	36	7967.262	.286
θ est. × Misfit	1097.041	8	137.130	.001
Selection × Misfit	991.69	4	247.923	<.001
$\theta \times \theta$ est. × Misfit	18185.274	72	252.573	.018
$\theta \times$ Selection × Misfit	690.358	36	19.177	<.001
θ est. × Selection × Misfit	965.31	8	120.664	<.001
$\theta \times \theta$ est. × Selection × Misfit	262.538	72	2.646	<.001
Misfit × Individuals (Error)	89887.953	239760	.090	
Total Effect				.888
Total Error				.112

Table 2
Results from the Mixed Design ANOVA After 25 Items were Administered

Type of Effect and Source	*SS*	*df*	*MS*	η^2
Between Subjects				
θ	193606.372	9	21511.819	.440
θ est.	1763.021	2	881.511	.004
Selection	1014.368	1	1014.368	.002
$\theta \times \theta$ est.	957.667	18	53.204	.002
$\theta \times$ Selection	866.589	9	96.288	.002
θ est. $\times$ Selection	645.906	2	322.953	.001
$\theta \times \theta$ est. $\times$ Selection	345.812	18	19.212	<.001
Error	12291.796	59940	.205	
Within Subjects				
Misfit	47833.450	4	11958.363	.109
$\theta \times$ Misfit	122574.667	36	3404.852	.279
θ est. $\times$ Misfit	1994.703	8	249.338	.005
Selection $\times$ Misfit	459.101	4	114.775	.001
$\theta \times \theta$ est. $\times$ Misfit	5367.011	72	74.542	.012
$\theta \times$ Selection $\times$ Misfit	479.613	36	13.323	.001
θ est. $\times$ Selection $\times$ Misfit	400.442	8	50.055	<.001
$\theta \times \theta$ est. $\times$ Selection $\times$ Misfit	271.191	72	3.767	<.001
Misfit $\times$ Individuals (Error)	48863.056	239760	.204	
Total Effect				.861
Total Error				.139

Table 3
Results from the Mixed Design ANOVA After 35 Items were Administered

Type of Effect and Source	*SS*	*df*	*MS*	η^2
Between Subjects				
θ	90329.447	9	10036.605	.385
θ est.	1341.847	2	670.924	.006
Selection	540.632	1	540.632	.002
$\theta \times \theta$ est.	364.043	18	20.225	.002
$\theta \times$ Selection	725.520	9	80.613	.003
θ est. $\times$ Selection	302.828	2	151.414	.001
$\theta \times \theta$ est. $\times$ Selection	359.615	18	19.979	.002
Error	6043.778	59940	.101	
Within Subjects				
Misfit	37072.589	4	9268.147	.158
$\theta \times$ Misfit	66992.819	36	1858.967	.285
θ est. $\times$ Misfit	2016.068	8	252.009	.009
Selection $\times$ Misfit	348.379	4	87.095	.001
$\theta \times \theta$ est. $\times$ Misfit	2839.537	72	39.438	.012
$\theta \times$ Selection $\times$ Misfit	479.697	36	13.325	.002
θ est. $\times$ Selection $\times$ Misfit	255.526	8	31.941	.001
$\theta \times \theta$ est. $\times$ Selection $\times$ Misfit	340.561	72	4.730	.002
Misfit $\times$ Individuals (Error)	24203.945	239760	.101	
Total Effect				.871
Total Error				.129

Table 4
Results from the Mixed Design ANOVA After 50 Items were Administered

Type of Effect and Source	*SS*	*df*	*MS*	η^2
Between Subjects				
θ	34470.425	9	3830.047	.327
θ est.	531.006	2	265.503	.005
Selection	229.462	1	229.462	.002
$\theta \times \theta$ est.	153.031	18	8.502	.001
$\theta \times$ Selection	367.263	9	40.807	.003
θ est. $\times$ Selection	136.373	2	68.187	.001
$\theta \times \theta$ est. $\times$ Selection	218.96	18	12.164	.002
Error	3903.131	59940	.065	
Within Subjects				
Misfit	17353.075	4	4338.269	.165
$\theta \times$ Misfit	28341.566	36	787.266	.269
θ est. $\times$ Misfit	1173.808	8	146.726	.011
Selection $\times$ Misfit	226.735	4	56.684	.002
$\theta \times \theta$ est. $\times$ Misfit	1568.846	72	21.790	.015
$\theta \times$ Selection $\times$ Misfit	351.697	36	9.769	.003
θ est. $\times$ Selection $\times$ Misfit	171.085	8	21.386	.002
$\theta \times \theta$ est. $\times$ Selection $\times$ Misfit	289.313	72	4.018	.003
Misfit $\times$ Individuals (Error)	15863.555	239760	.066	
Total Effect				.812
Total Error				.188

Effect Sizes Greater Than .10

As shown by Tables 1–4, the θ, and $\theta \times$ misfit factors accounted for the most variation in the ANOVA model as defined by η^2. It was observed that as the number of items increased from 15 to 50, the variation accounted for by the θ factor decreased from .538 to .327. The η^2 values for the $\theta \times$ misfit interaction remained largely stable as test length increased. Interestingly, the η^2 for the misfit factor increased from .03 to .165 as test length increased from 15 to 50 items. It appeared that the misfit factor began to absorb some of the variation accounted for by the θ condition when test length increased. The rest of the factors in the mixed-design ANOVA accounted for a negligible amount of variation across test lengths.

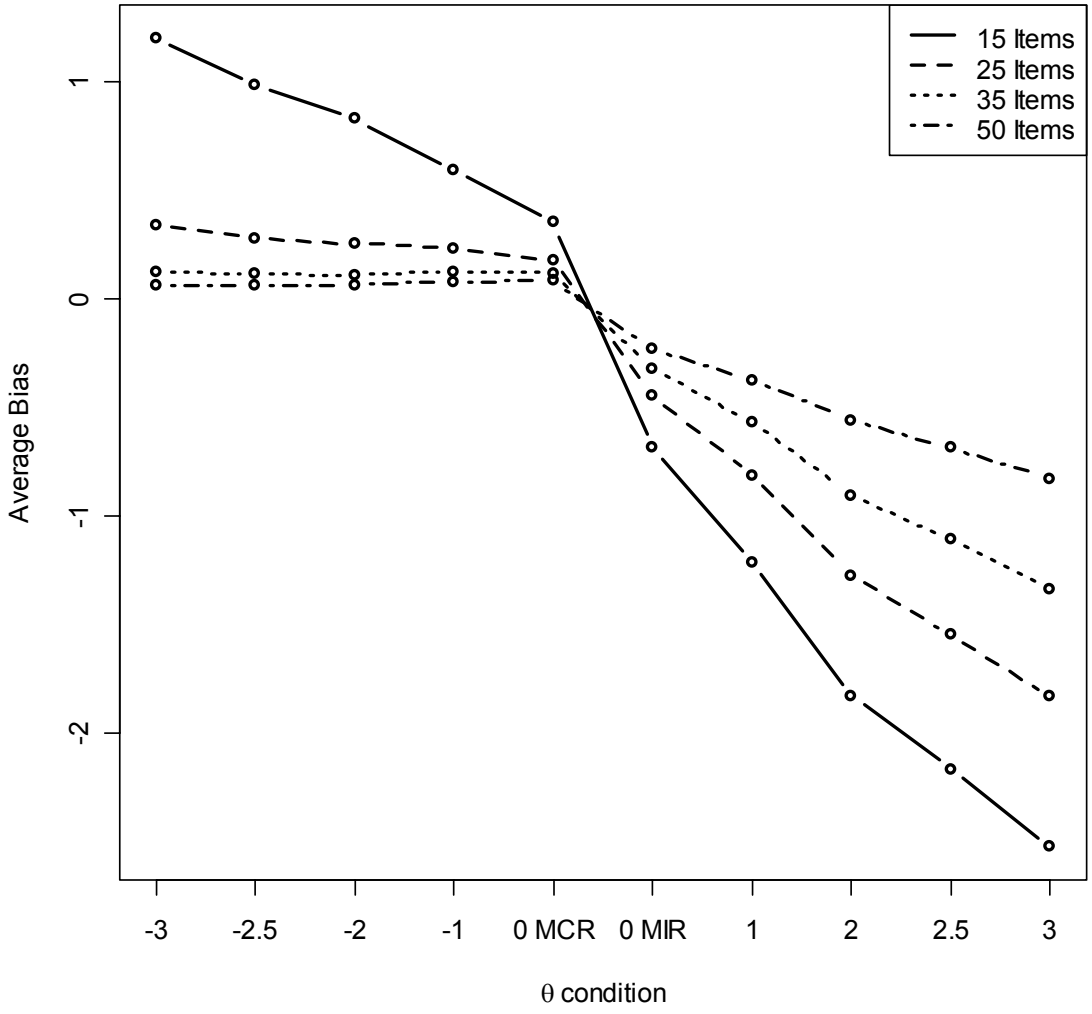
θ

Figure 2 and Table A2 show that the average bias in θ (without regard to θ estimation method) decreased as θ changed in absolute value from 3 to 0. This trend was most pronounced for θ values greater than 0. As shown by Figure 2, the recovery of θ improved as test length increased. For conditions with MCR (i.e., primarily the negative θ values), there was a large amount of positive bias when only 15 items were used in the CAT. Recovery improved after 25 items as all of the bias values became less than 0.5. After 35 and 50 item conditions the bias values were similar across θ for the MCR conditions.

The average bias values were greater in absolute value for conditions with MIR (i.e., primarily the positive θ values). As seen in Table A2, after 15 items the average bias for $\theta = 3$ was –2.528, despite the average being computed across misfit conditions. The bias decreased as test length increased, but even after 50 items the average bias for $\theta = 3$ was

–0.831.

Figure 2
Average Bias Across θ for Different CAT Lengths

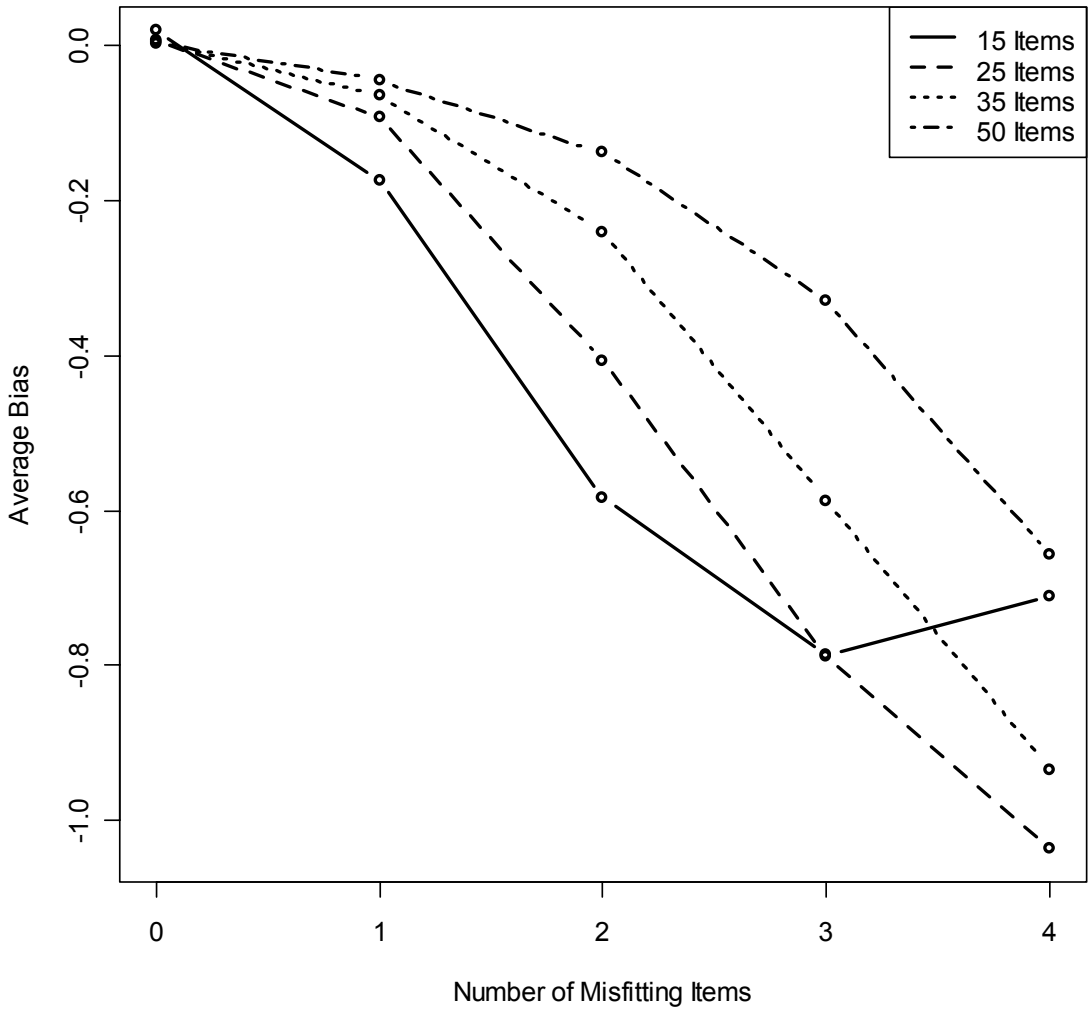


Misfit

Figure 3 and Table A3 displays the average bias values for different test lengths across the five different misfit conditions. Figure 3 shows that there was greater negative bias than positive bias in this study, and that this bias generally increased as the number of misfitting items increased. This can be attributed to the greater effect of MIR compared to MCR. As shown by Figure 3, the effect of misfit on recovery differed for the 15-item condition compared to the 25-, 35-, and 50-item conditions. For 15 items the average bias leveled off after two items of misfit. After 15 items there was positive bias for the MCR conditions, that when averaged against the negative bias resulted in the leveling off. The positive bias was smaller for the 25-, 35-, and 50-item conditions, but the bias continued to increase in absolute value as the number misfitting items increased.

Figure 3

Average Bias Across Misfit Conditions for Different CAT Lengths

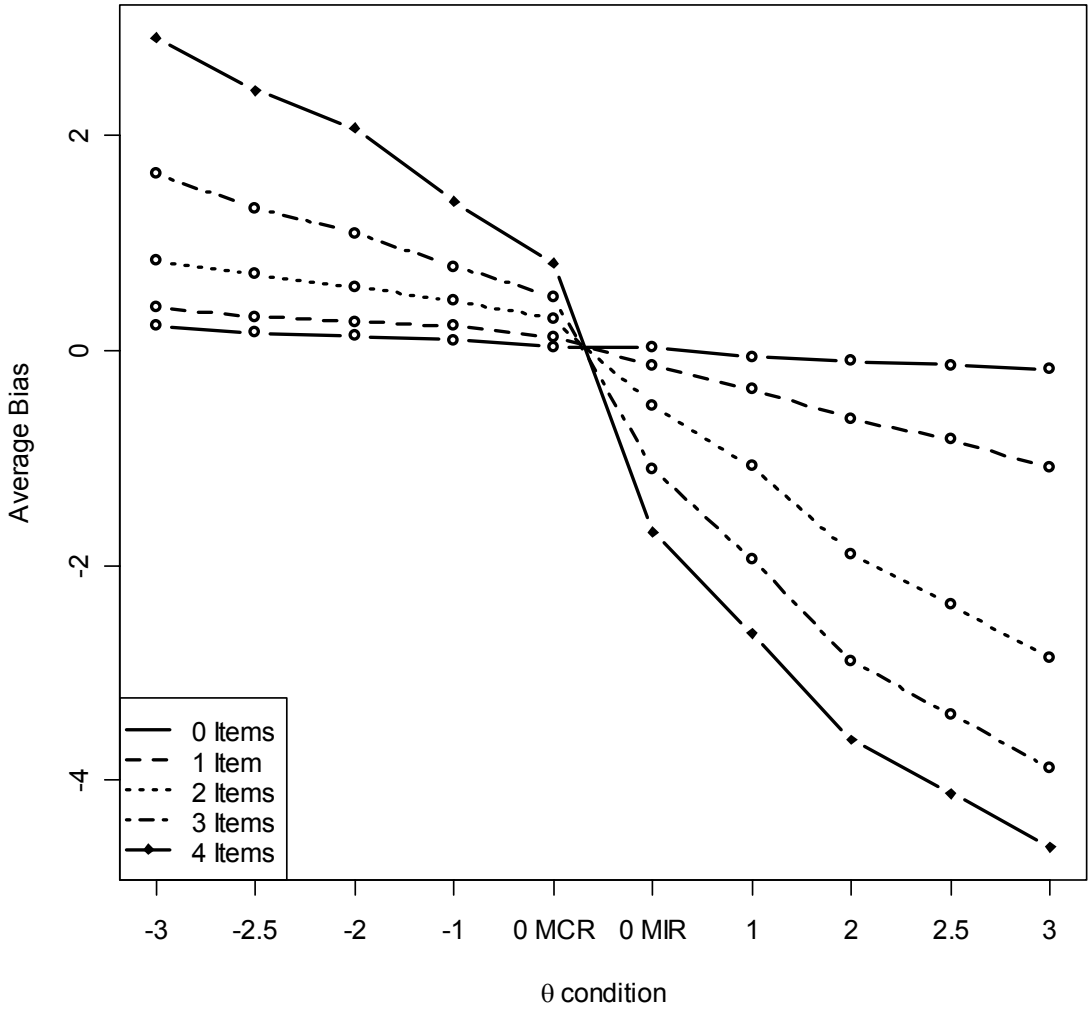


θ × Misfit

The η^2 values for the θ × misfit interaction ranged from .269 after 25 items to .286 after 15 items were administered. As seen in Figure 4 and Table A4, the θ × misfit interaction resulted from the increased bias in θ estimation as the number of misfitting items increased. As seen in Table A4, the average bias for the zero-misfit condition remained less than 0.06 in absolute value across θ. The bias increased both as the number of misfitting items increased and as θ increased in absolute value. In addition, whether misfit was operationalized as correct or incorrect responses changed the sign and severity of the bias. As seen in Figure 4, the effect of MIR on the bias values was greater than MCR.

Figure 4

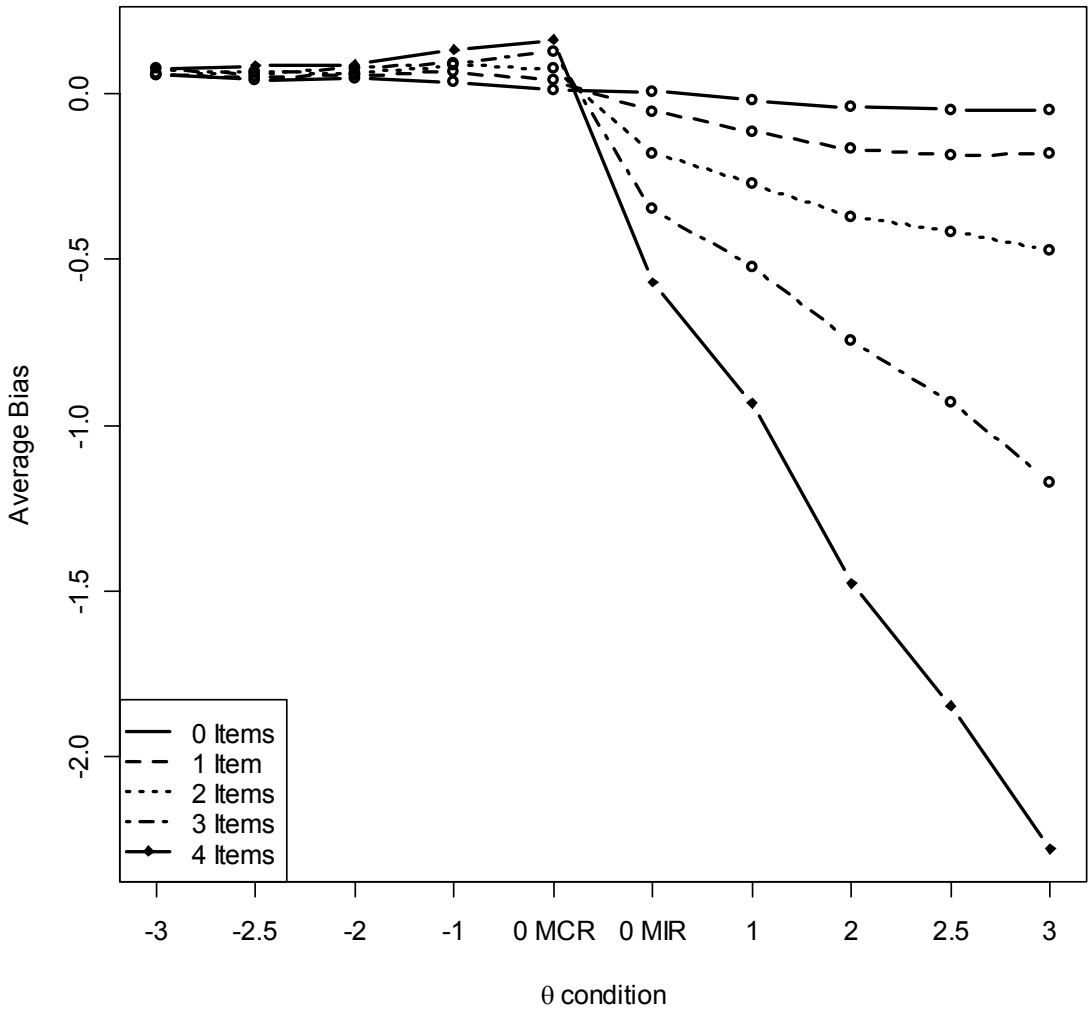
Average Bias for the θ × Misfit Interaction After 15 Items



As shown by Figure 5, the bias values for the MCR conditions remained near zero across the misfit conditions for θ values of –3 to –2 when 50 items were administered. For θ values of –1 and 0 the bias became slightly more positive as the number of misfitting items increased. Figure 5 reveals that there was still negative bias present after 50 items for the MIR conditions. Even when the simulee responded incorrectly to one item, θ did not recover to zero bias after 50 items.

Figure 5

Average Bias for the θ × Misfit Interaction After 50 Items



Other Notable Interactions in the ANOVA

Bock and Mislevy (1982) showed that EAP estimation was biased toward the mean of the prior, so the small effect size for the estimation method condition does not necessarily mean that estimation method had no effect on θ recovery. In addition, the effect of K-L selection and WLE on recovery of θ was shown to be non-negligible until 10−15 items were administered in a CAT (Cheng & Liou, 2000). In order to examine the effect of estimation method and item selection on θ recovery it would be prudent to not collapse across θ, as the direction and size of the bias changed conditional on θ.

θ × Estimation × Selection

Results after 15 items. The results after 15 items across FI item selection and K-L selection were nearly identical, as shown by Figure 6 and Table A6. There was a large negative bias present for each of the MIR conditions. A difference between K-L and FI that emerged was for WLE θ estimation. WLE was the most biased estimator for conditions with MIR with FI selection, but less biased than MLE with K-L selection.

Interestingly, for the MCR conditions WLE had the most bias of the θ estimators with K-L selection and the least bias with FI selection for the θ of –3 and –2.5 conditions. The bias of EAP estimation was lower than MLE or WLE after 15 items.

Results after 50 items. There were fewer differences between K-L and FI after 50 items were administered, as shown by Table A7 and Figure 7. For θ conditions with MIR, WLE was the most biased of the three estimation methods when FI selection was used, as shown by Figure 7. WLE was the least biased θ estimation method when K-L item selection was used. There were no differences across item selection for MLE or EAP. Despite the administration of 50 items, each θ method remained quite biased for the MIR

conditions.

No differences emerged between the three θ estimation methods across item selection for the MCR conditions. EAP was positively biased for the MCR conditions due to the effect of the prior. WLE and MLE were not as biased as EAP, but still showed a bias of about 0.1 for the $\theta = -1$ and 0 MCR conditions.

Figure 6
Average Bias for the θ × Estimation × Selection Interaction After 15 Items

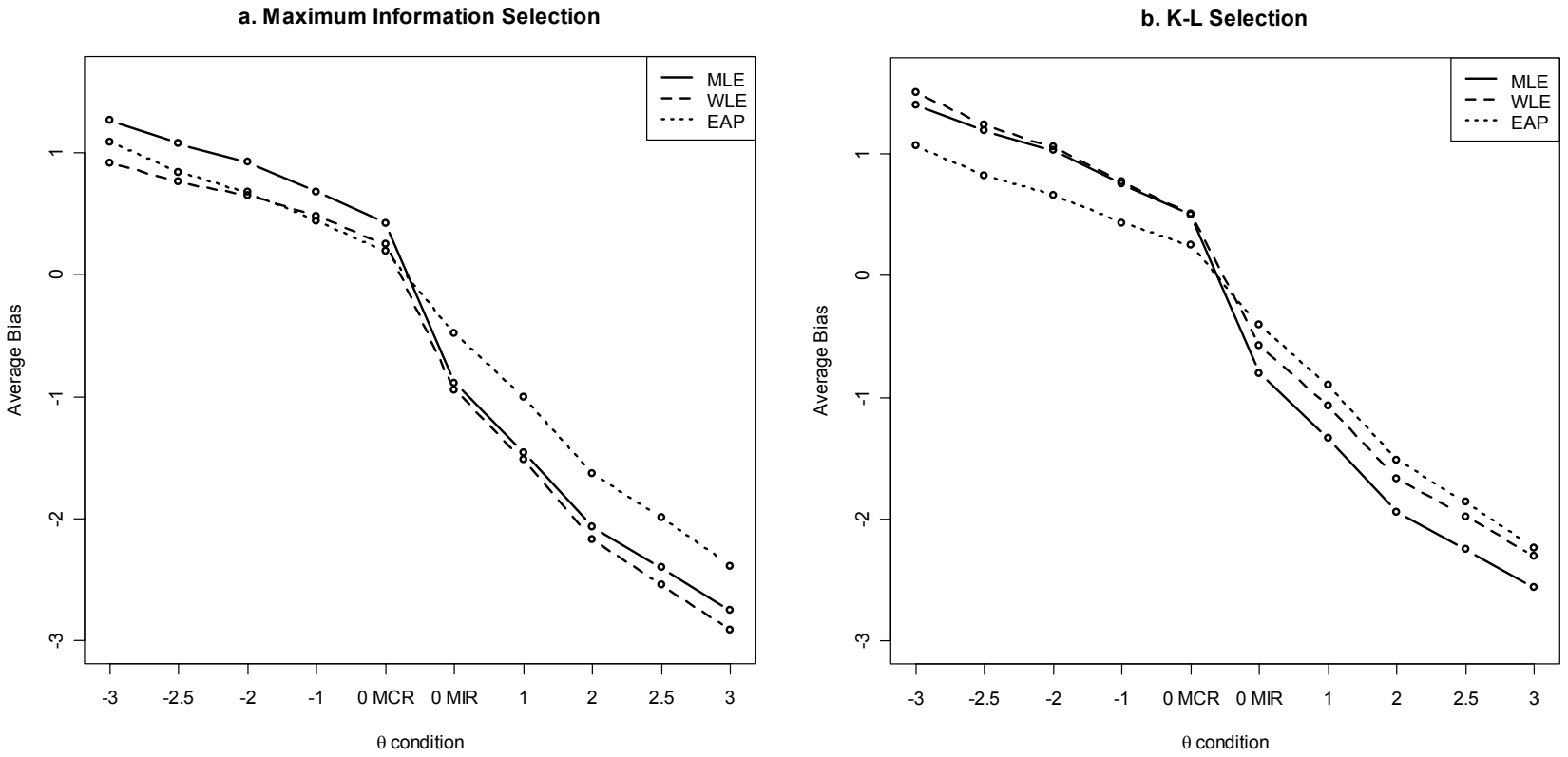
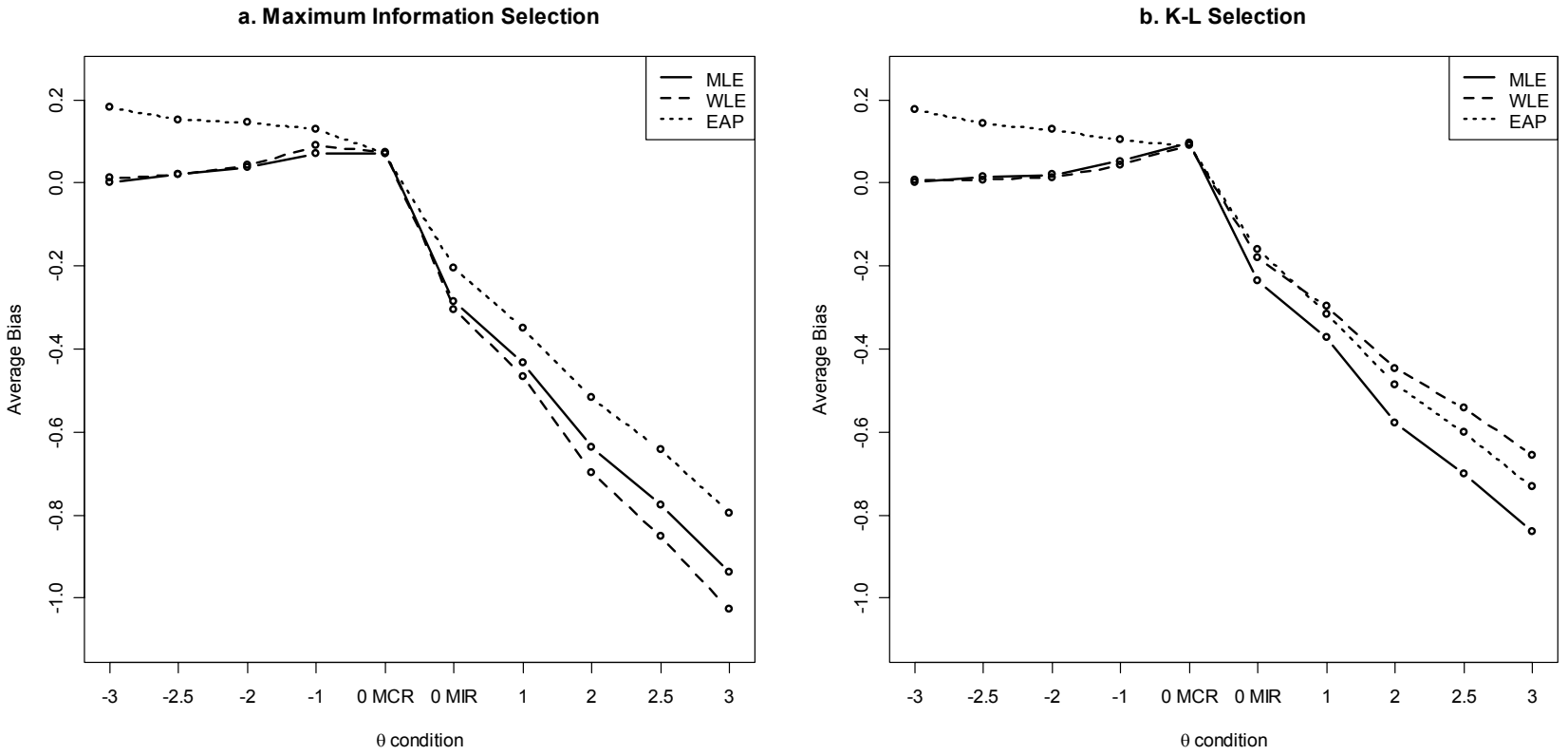


Figure 7
Average Bias for the θ × Estimation × Selection Interaction After 50 Items



Results for Each Cell of the Design

Bias

Conditions Without Misfit

The average bias values for each cell in the design are presented in Tables A8−A47. It was observed that, as expected, EAP was biased toward the mean of the prior. This bias increased as θ moved away from zero. As seen in Tables A8−A47, EAP had lower bias than MLE or WLE when θ was at the mean of the prior (0).When there were no misfitting items, WLE and MLE had nearly equal bias values.

The average bias was calculated from 6 to 50 items for the MLE θ estimates that converged. These values are plotted for $\theta = 3$, 1, –1, and –3 in Figures 8–11. As seen in Figure 8, it took MLE over 10 items to yield unbiased θ estimates when $\theta = 3$. For $\theta = 1$ (Figure 9) the θ estimates did not become unbiased until 15 items were administered. A similar trend was observed for $\theta = -1$ and –3 (Figures 10 and 11), as the MLE estimates did not become unbiased until about 20 items were administered. This was evident for $\theta = -3$ as all three estimation methods had bias in excess of 0.2 after 10 items.

The performance of WLE was quite similar to MLE. WLE θ estimates were slightly lower (and less biased) than MLE when $\theta = 3$, 1, and –1. For $\theta = -3$, it can be seen in Figure 11b that WLE was more biased than MLE when K-L selection was used. The bias of EAP toward the mean of the prior can be seen in Figures 8–11. As expected, the bias of EAP was greater for $\theta = \pm 3$ than it was for $\theta = \pm 1$. The bias of EAP formed a non-linear function where the bias rapidly decreased until about 20 items into the CAT, at which point the trend began to asymptote.

There were few differences across item selection methods. No consistent differences

between item selection methods were observed when $\theta = 1$ or 3. It was observed that θ estimates for K-L selection when $\theta = -1$ or -3 were more biased than when FI was used.

MIR

It was observed for the $\theta \times$ estimation $\times$ selection interaction that EAP provided less biased θ estimates than MLE or WLE. These results were contingent on test length and number of misfitting items. As seen in Tables A13, A23, A33, and A43, EAP provided less biased θ estimates than MLE or WLE for 1–4 misfitting items and a θ of 0.

For test lengths of 15, 25, 35, and 50 items, MLE provided less biased θ estimates than EAP when there was one misfitting item. As seen in Tables A36−A37 and A44−A47, MLE provided less biased θ estimates than EAP when there were two misfitting items. It was observed that MLE was less biased than EAP after 35 or 50 items, while EAP was less biased than MLE after 15 or 25 items. EAP still provided less biased θ estimates than MLE or WLE when there were three or four misfitting items. These results implied that the reduction in bias from use of a prior dissipated as the CAT length increased.

1 misfitting item. To examine this phenomenon more closely, average bias values for $\theta = 3$ and 1 were obtained for CAT lengths from 6 to 50 items. Figures 12 and 13 display the average bias values for the one misfitting item condition for both item selection procedures. The θ estimates were less biased for $\theta = 1$ than for $\theta = 3$. However, there was still bias present for $\theta = 1$ after 50 items were administered.

MLE became less biased than EAP when 10 items were administered with FI for both $\theta = 3$ and 1. WLE did not become less biased than EAP until after 16 items for both $\theta = 3$ and 1, and was consistently more biased than MLE for both item selection methods. MLE and WLE performed better than EAP after just 6 items for both $\theta = 3$ and 1 when K-L

selection was used.

2 misfitting items. Figures 14 and 15 display the recovery of θ from 6 to 50 items. The results indicated that the bias of the θ estimates was lower when $\theta = 1$ compared to $\theta = 3$. For FI selection and $\theta = 3$ (Figure 14), EAP was less biased than MLE until 34 items were administered. It took WLE until 41 items to become less biased than EAP for FI selection when $\theta = 3$. When $\theta = 1$ (Figure 15), EAP remained less biased than MLE or WLE until 45 items were administered using FI selection.

When K-L information was used to select items for $\theta = 3$, WLE became less biased than EAP after just 21 items, while it took MLE 30 items. In addition, WLE was consistently less biased than MLE with K-L selection but more biased than MLE with FI selection. A similar pattern emerged for $\theta = 1$, as WLE was less biased than it was for FI selection.

3 and 4 misfitting items. Figures 16 and 18 display the average bias values after 3 and 4 misfitting responses for $\theta = 3$. As seen in Figure 16, when K-L selection was used, WLE became less biased than both EAP and MLE after 40 items were administered. One trend observed for FI selection was that WLE diverged (became more biased) from MLE between 20 and 50 items. Interestingly, the opposite trend was observed for K-L selection as WLE became progressively less biased compared to MLE as test length increased.

Similar results were observed for $\theta = 1$, as shown by Figures 17 and 19. The bias of MLE and EAP were the same across item selection method. However, WLE was found to be consistently less biased when K-L selection was used compared to FI. As seen before, the bias of the θ estimates was less when $\theta = 1$ compared to $\theta = 3$.

MCR

The performance of the θ estimation methods in terms of bias can be best summarized graphically. Figures 20–27 displays the recovery of $\theta = -1$ and -3 for the 1 to 4 misfitting items conditions after 6 to 50 items were administered. It can be observed in Figures 23b and 25b, as well as 22, 24, and 26, that the bias curves for MLE were not quite smooth.

1 misfitting item. As seen in Figures 20a and 21a, WLE provided the most unbiased θ estimates when FI selection was used. For $\theta = -3$ that WLE was consistently more biased than MLE when K-L selection was used, as seen in Figure 24b. When $\theta = -1$, WLE was more biased than MLE until 15 items were administered in the CAT using K-L selection. EAP was the most biased θ estimation method when there was one misfitting item. Interestingly, as shown in Tables A38 and A41, the bias for $\theta = -1$ after 50 items was 0.064 while the bias for $\theta = -3$ was –0.012.

2 to 4 misfitting items. One trend observed for conditions with 2 to 4 misfitting items and $\theta = -3$, was that EAP provided the least biased θ estimates when a short CAT was used with K-L selection. When FI selection was used, WLE was less biased than EAP for the 2-item MCR condition, as shown by Figure 22a. EAP was less biased than WLE for the 3- and 4-item MCR conditions with FI selection. The number of items required for MLE to provide less biased estimates than EAP was dependent on the number of misfitting items. As seen in Figures 22a, 24a, and 26a, it took test lengths of 14, 20, and 26 items for MLE with FI selection to yield less biased θ estimates than EAP for the 2, 3, and 4 misfitting item conditions, respectively.

Similar results were obtained for $\theta = -1$ as seen in Figures 23, 25, and 27. It was found that EAP was less biased than MLE or WLE until a specific number of items were

administered. The number of items required for MLE to become less biased than EAP increased as the number of misfitting items increased from 2 to 4.

The effect of item selection method was quite small. It was observed in Figures 22–27 that K-L item selection provided slightly more biased θ estimates than FI selection for the first 15 items in the CAT. As such, it took longer for MLE to provide less biased estimates than EAP when K-L selection was used compared to FI selection.

The performance of WLE again was dependent on item selection method. For FI selection, WLE provided θ estimates that were less biased than MLE when there was MCR present. The trend reversed for K-L selection as WLE estimates became slightly more biased than MLE estimates.

Figure 8

Average Bias Across CAT Lengths for the 0-Item Misfit Condition for $\theta = 3$

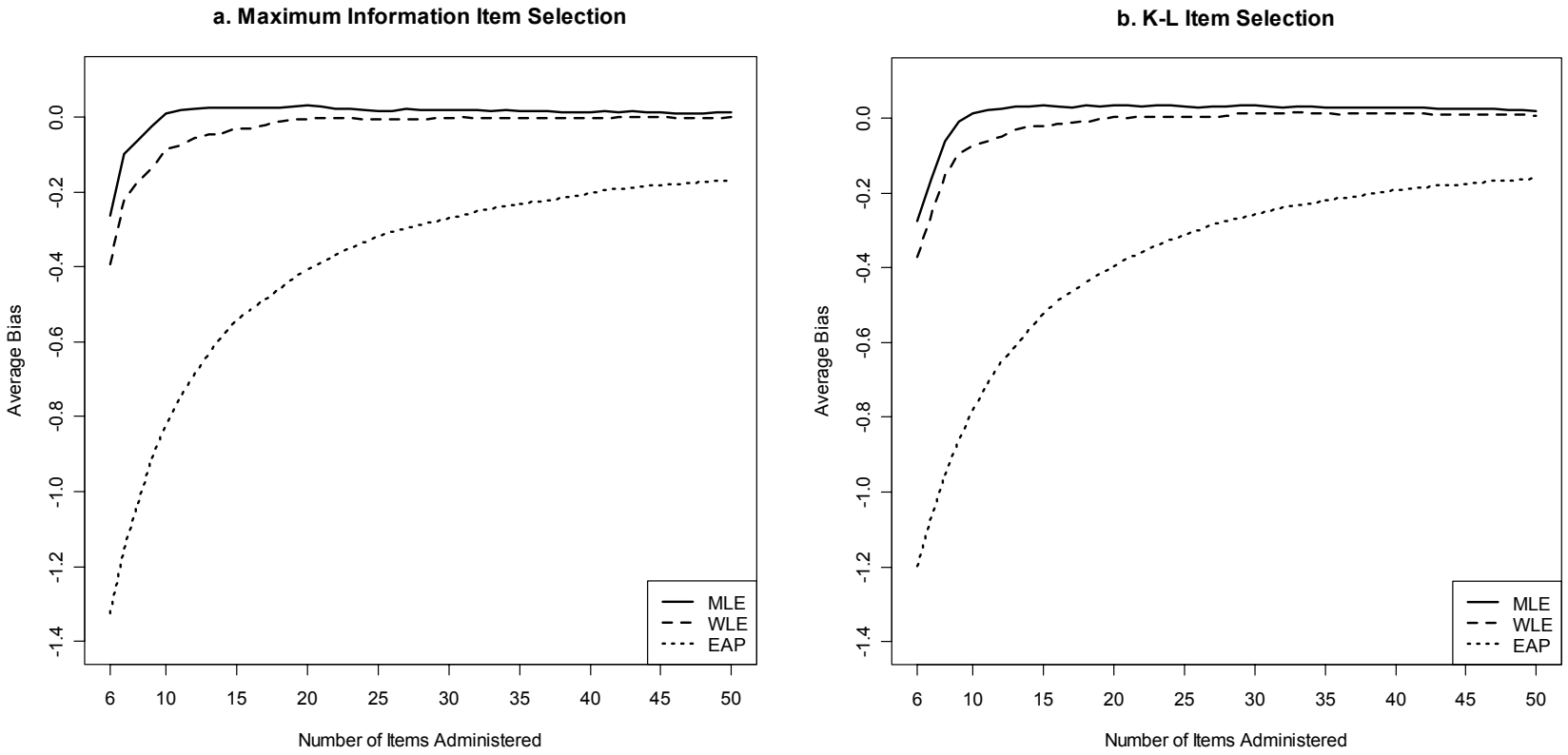


Figure 9

Average Bias Across CAT Lengths for the 0-Item Misfit Condition for $\theta = 1$

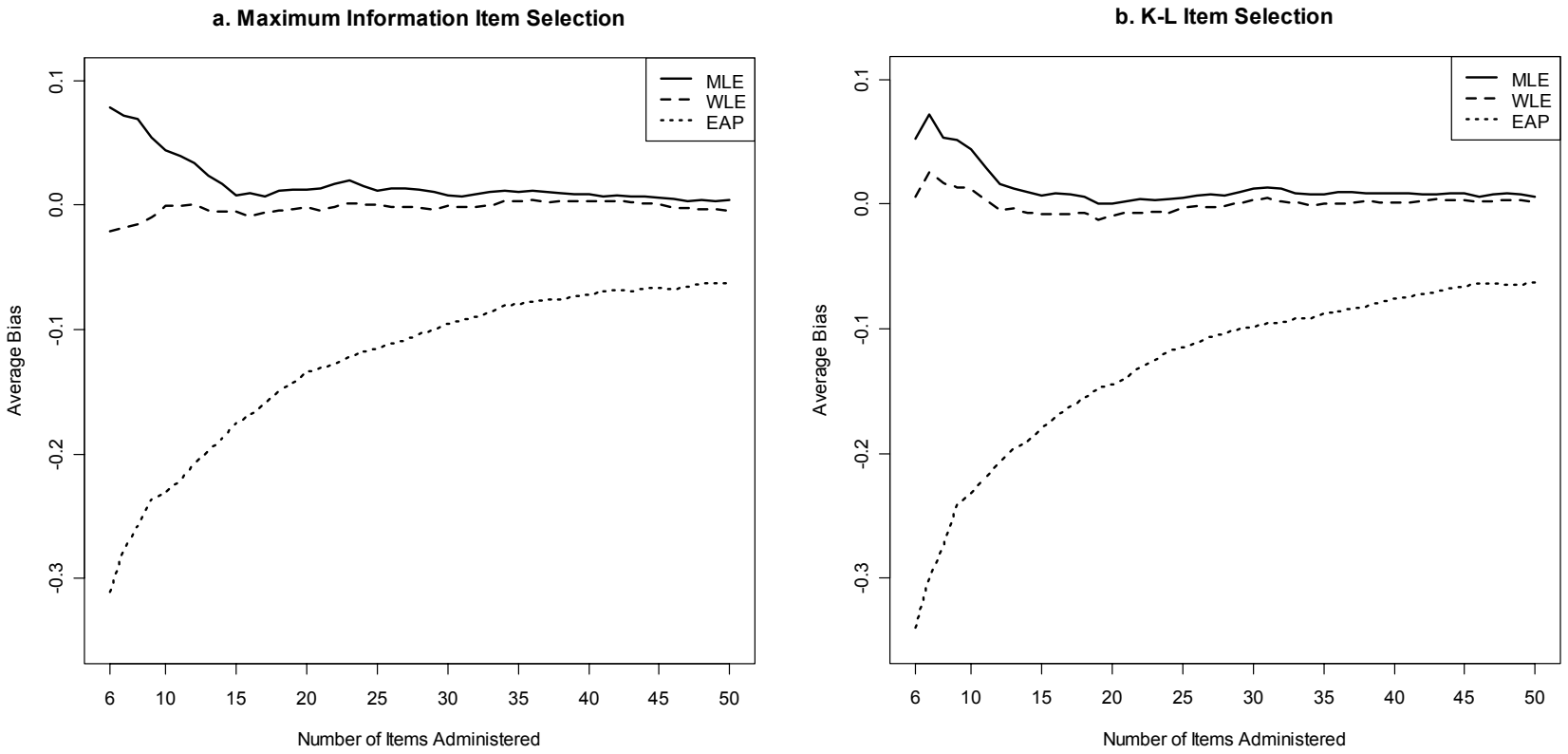


Figure 10

Average Bias Across CAT Lengths for the 0-Item Misfit Condition for θ = –1

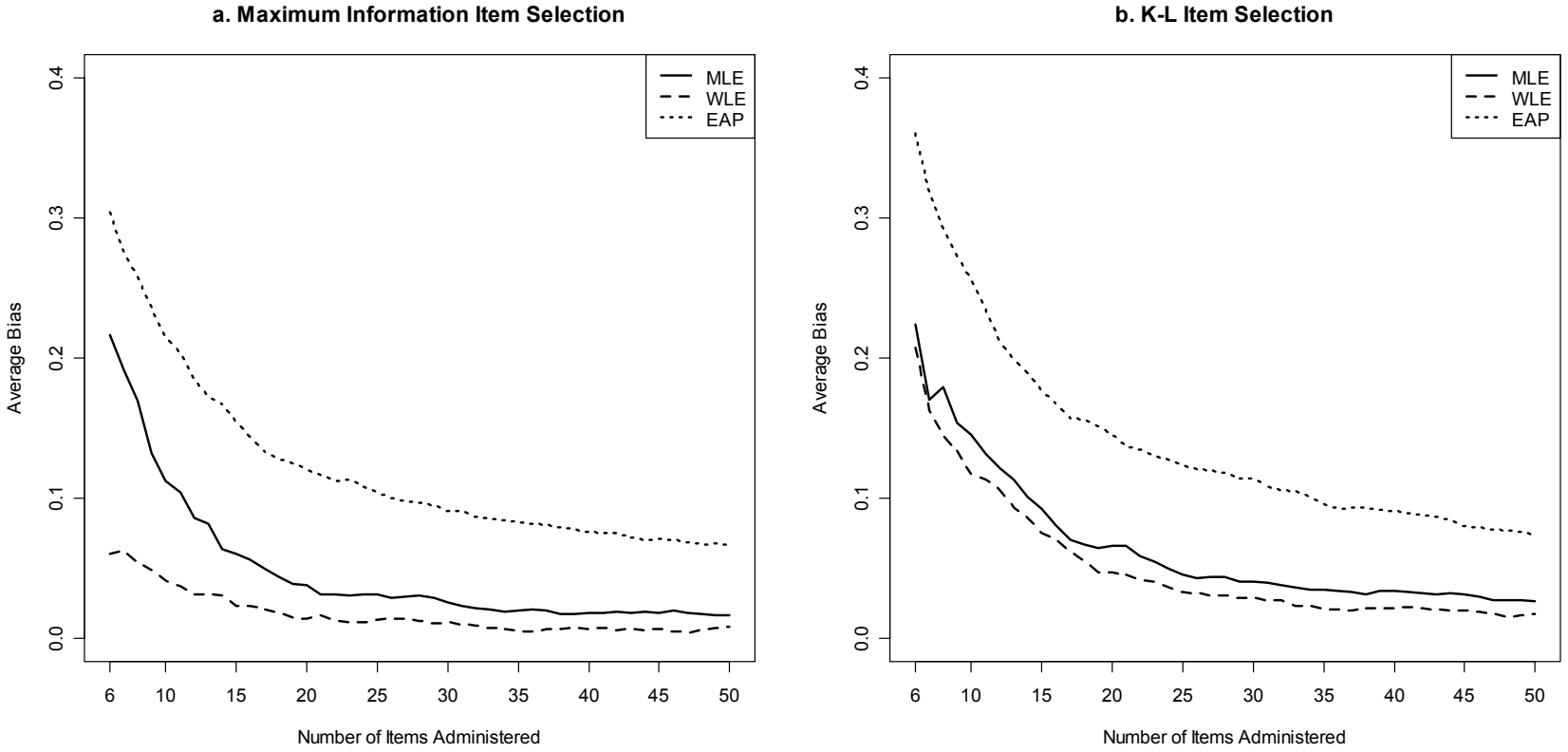


Figure 11

Average Bias Across CAT Lengths for the 0-Item Misfit Condition for $\theta = -3$

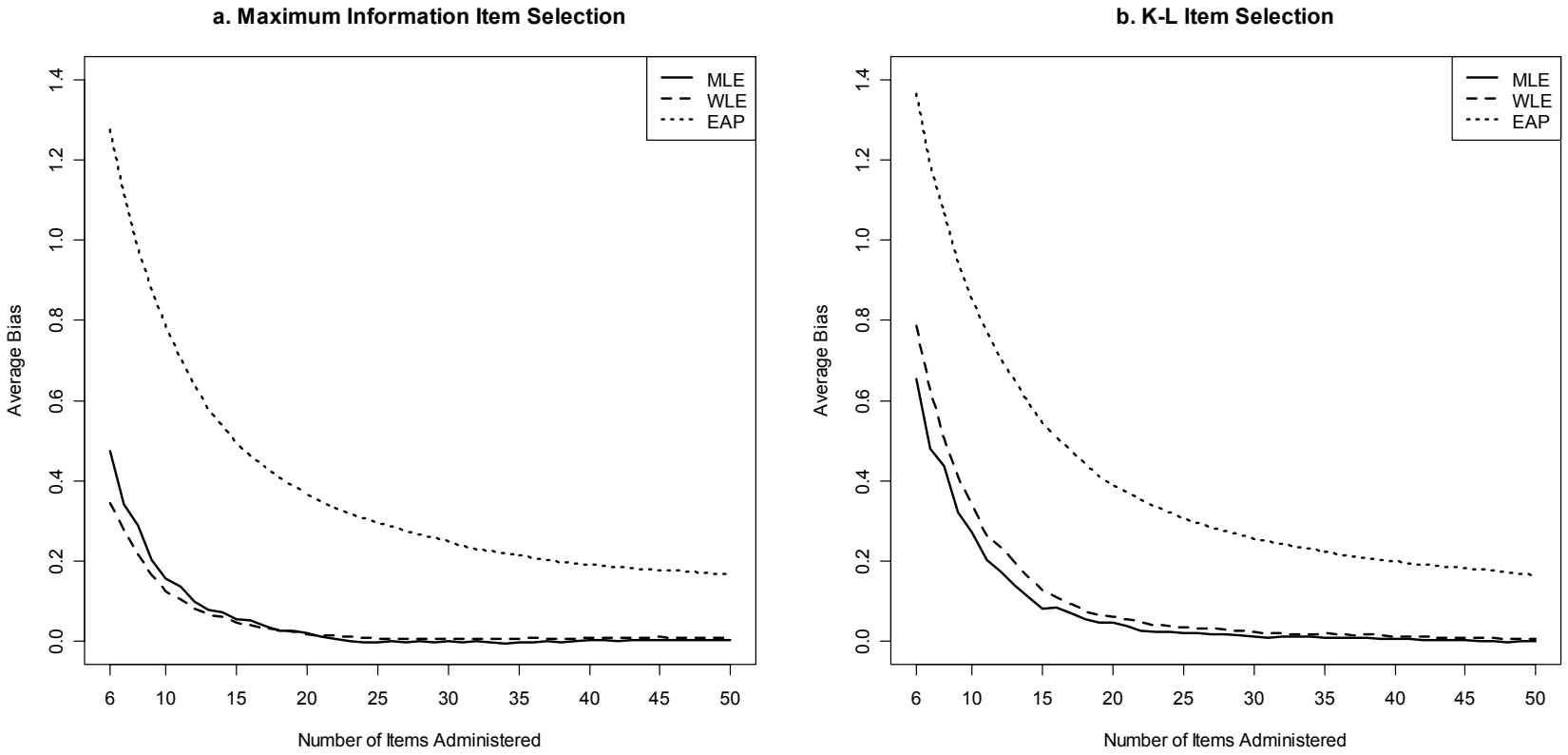


Figure 12

Average Bias Across CAT Lengths for the 1-Item Misfit Condition for θ = 3 (MIR)

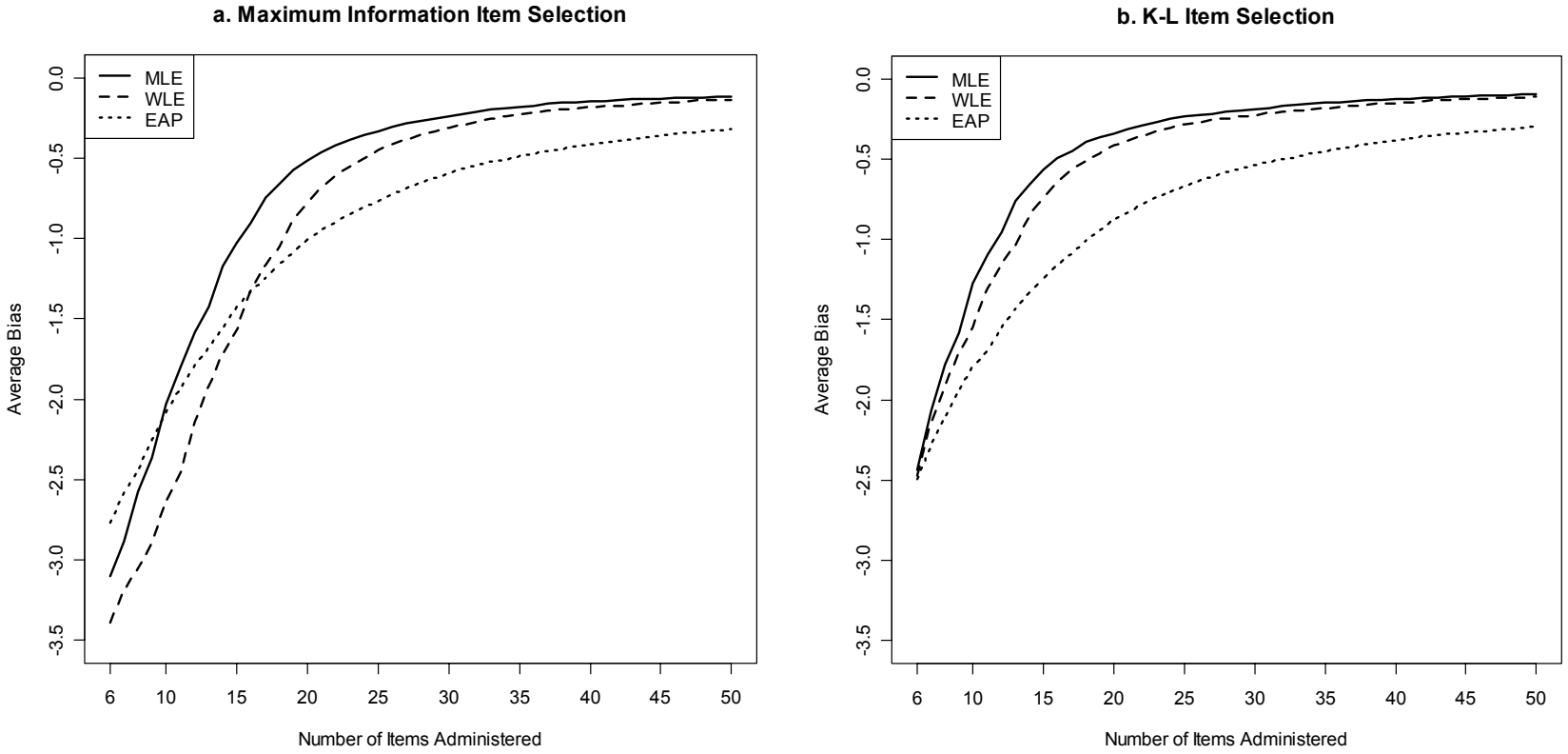


Figure 13

Average Bias Across CAT Lengths for the 1-Item Misfit Condition for θ = 1 (MIR)

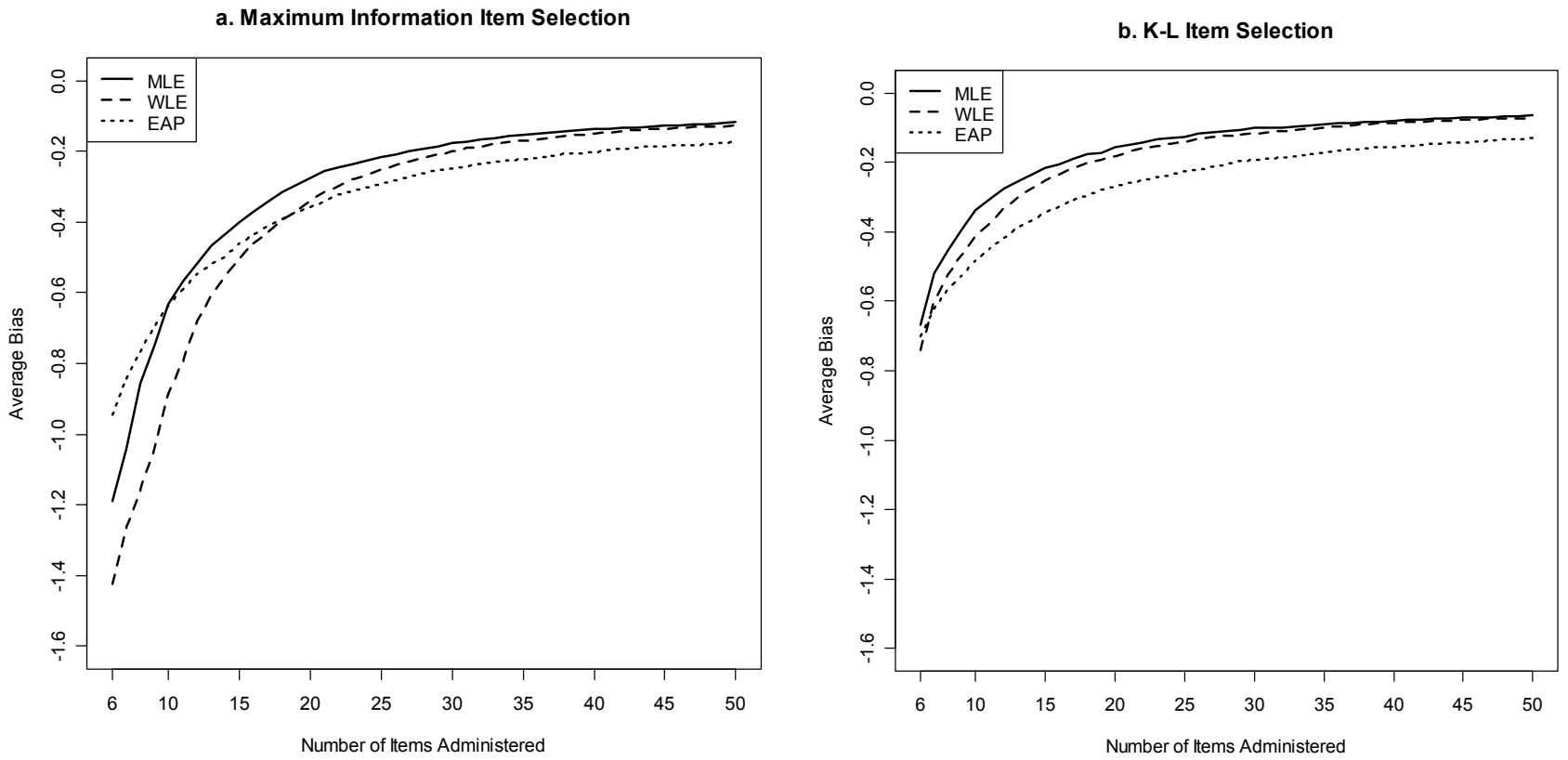


Figure 14

Average Bias Across CAT Lengths for the 2-Item Misfit Condition for θ = 3 (MIR)

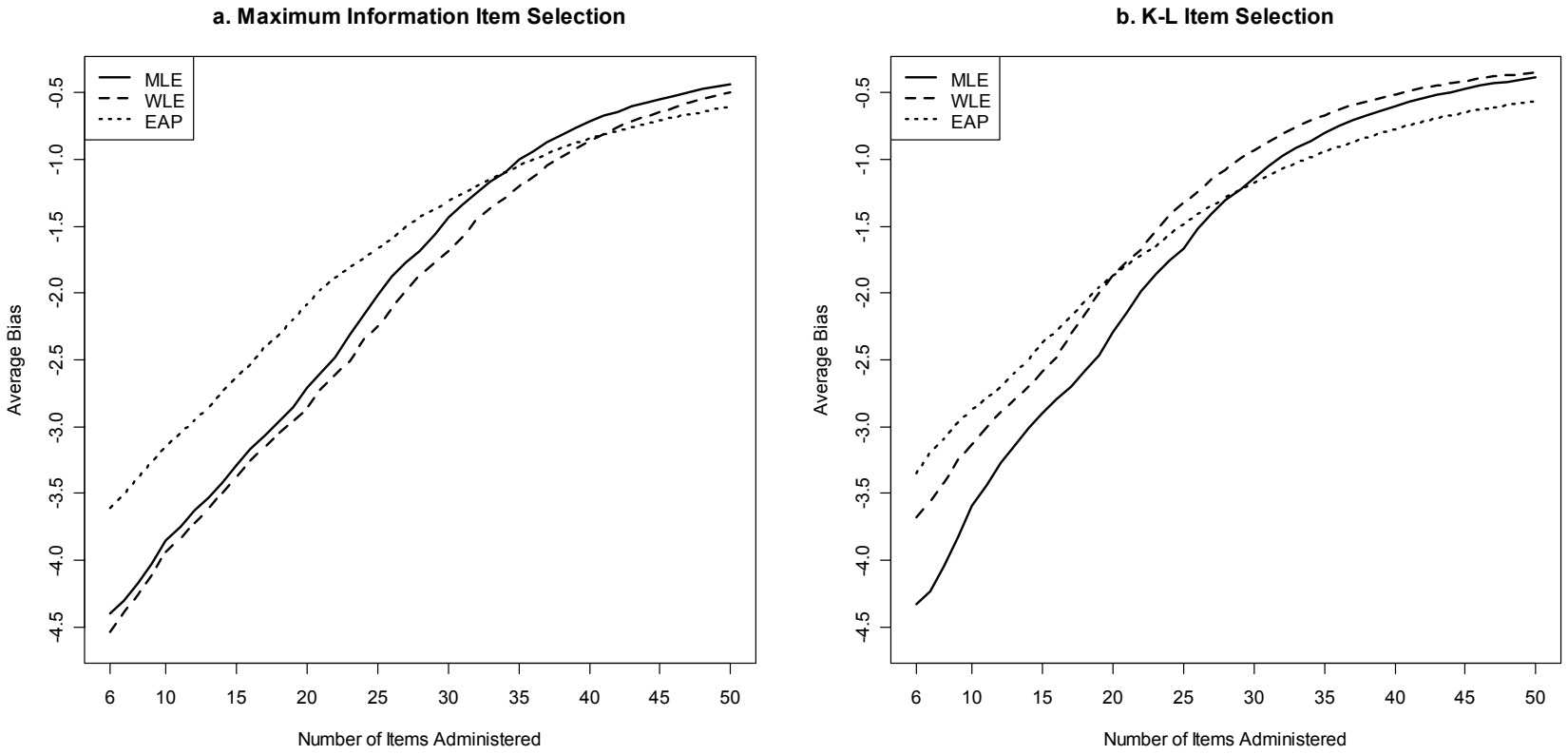


Figure 15

Average Bias Across CAT Lengths for the 2-Item Misfit Condition for θ = 1 (MIR)

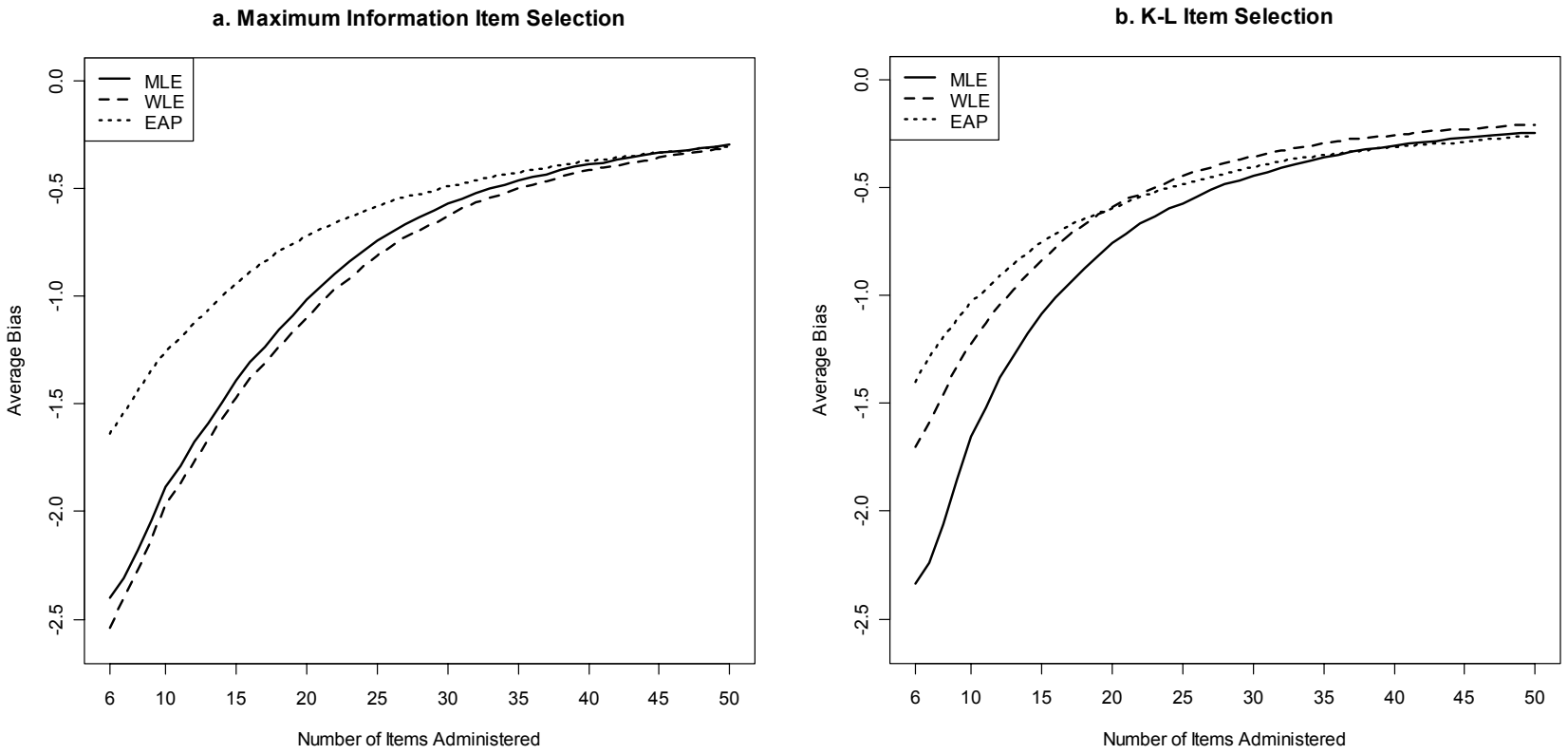


Figure 16

Average Bias Across CAT Lengths for the 3-Item Misfit Condition for θ = 3 (MIR)

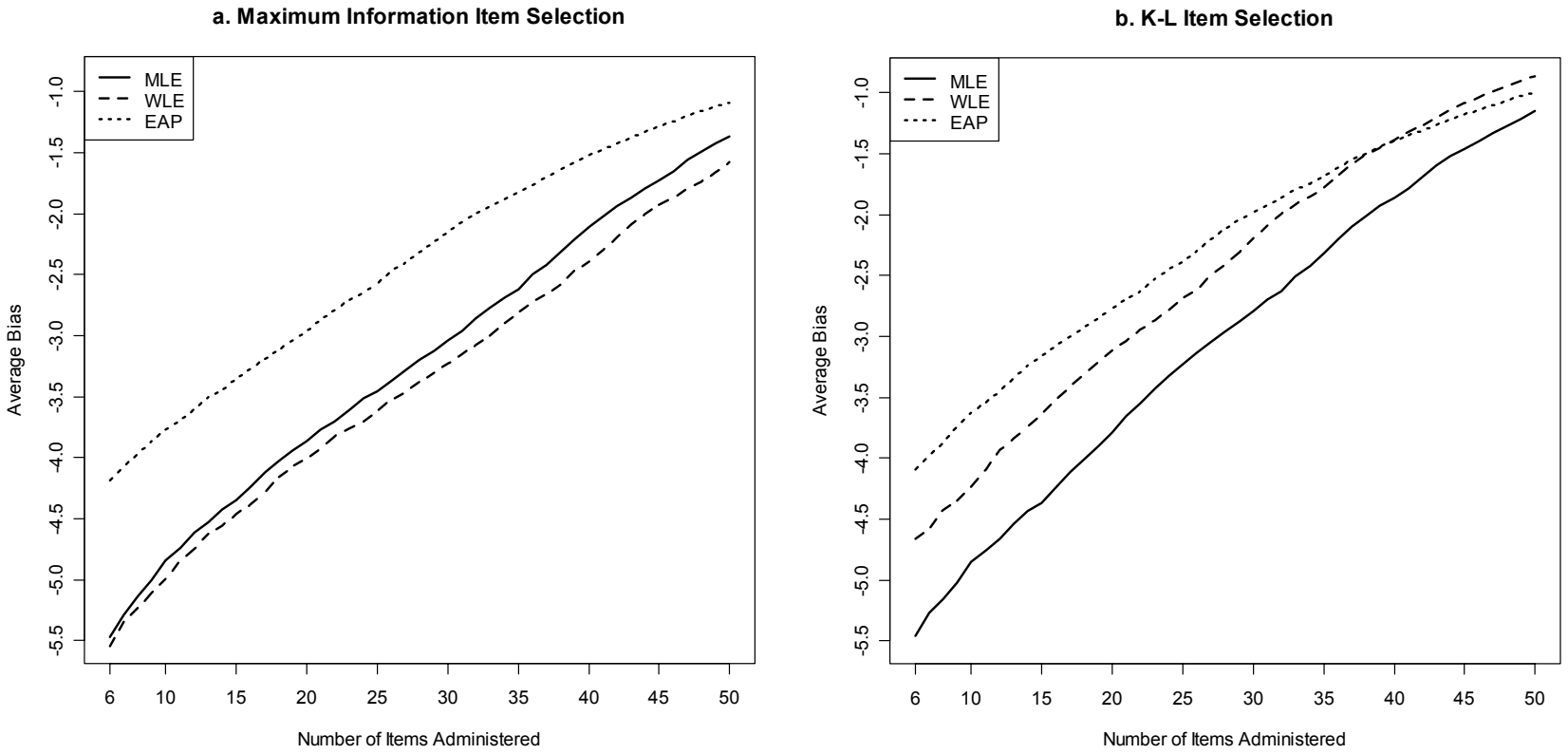


Figure 17

Average Bias Across CAT Lengths for the 3-Item Misfit Condition for θ = 1 (MIR)

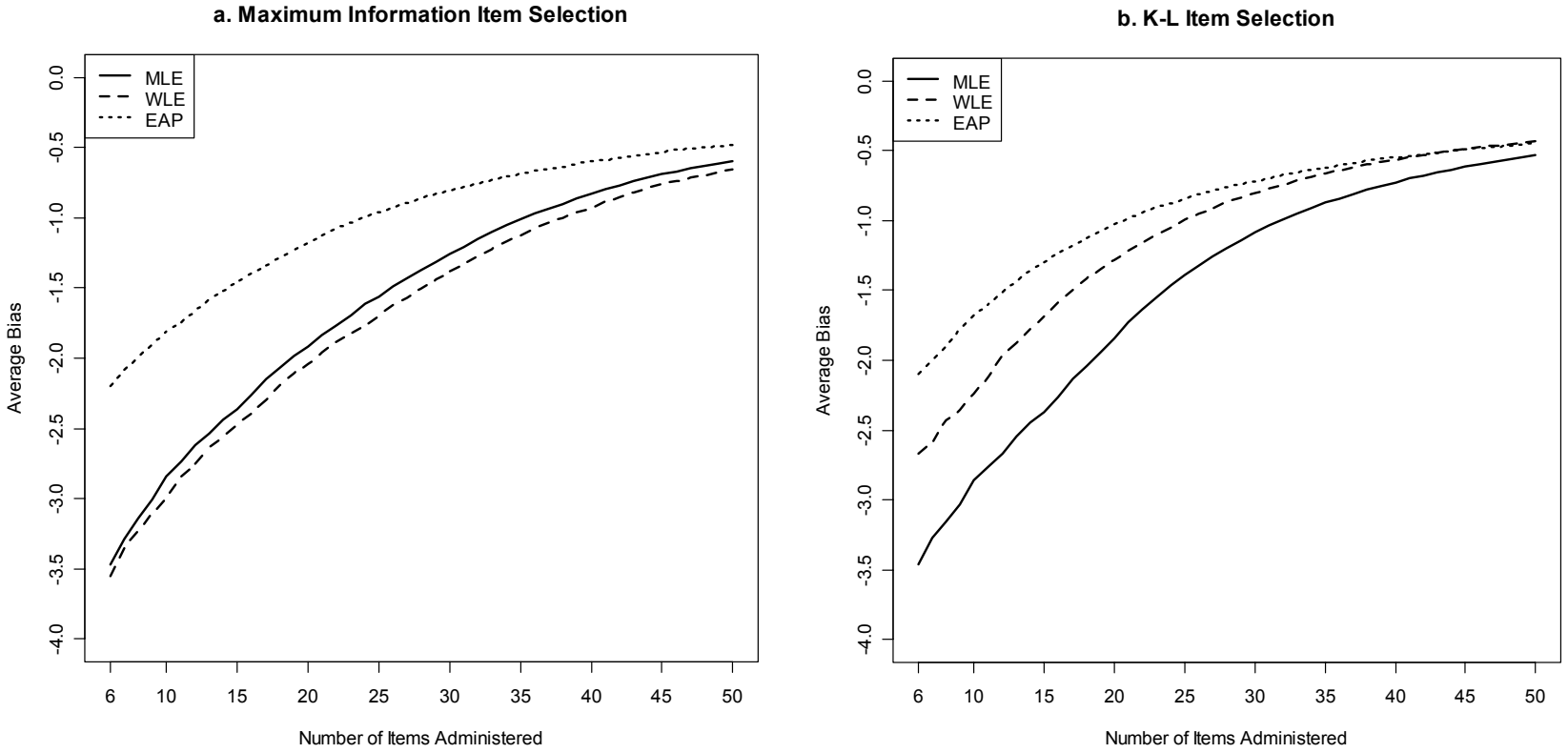


Figure 18

Average Bias Across CAT Lengths for the 4-Item Misfit Condition for θ = 3 (MIR)

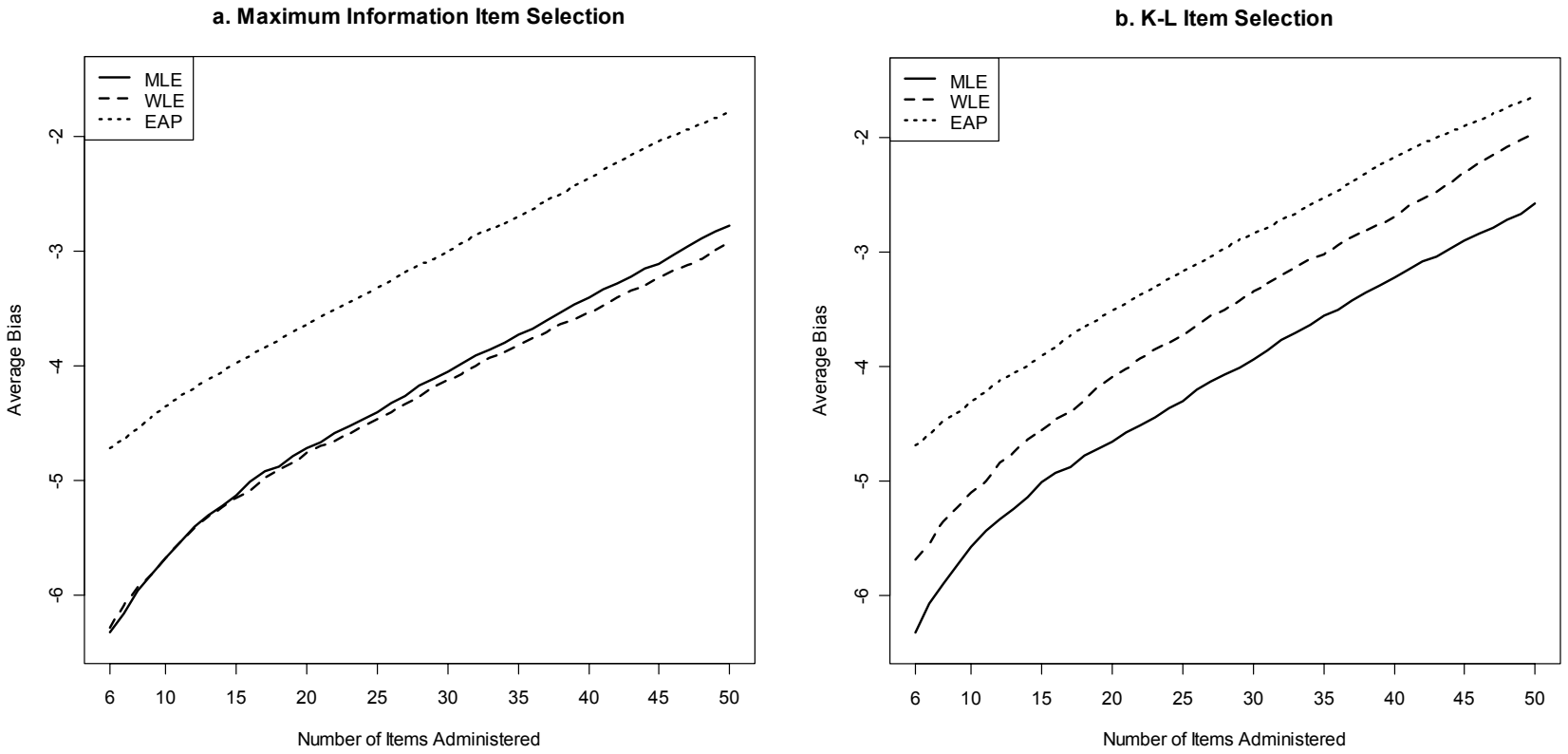


Figure 19

Average Bias Across CAT Lengths for the 4-Item Misfit Condition for θ = 1 (MIR)

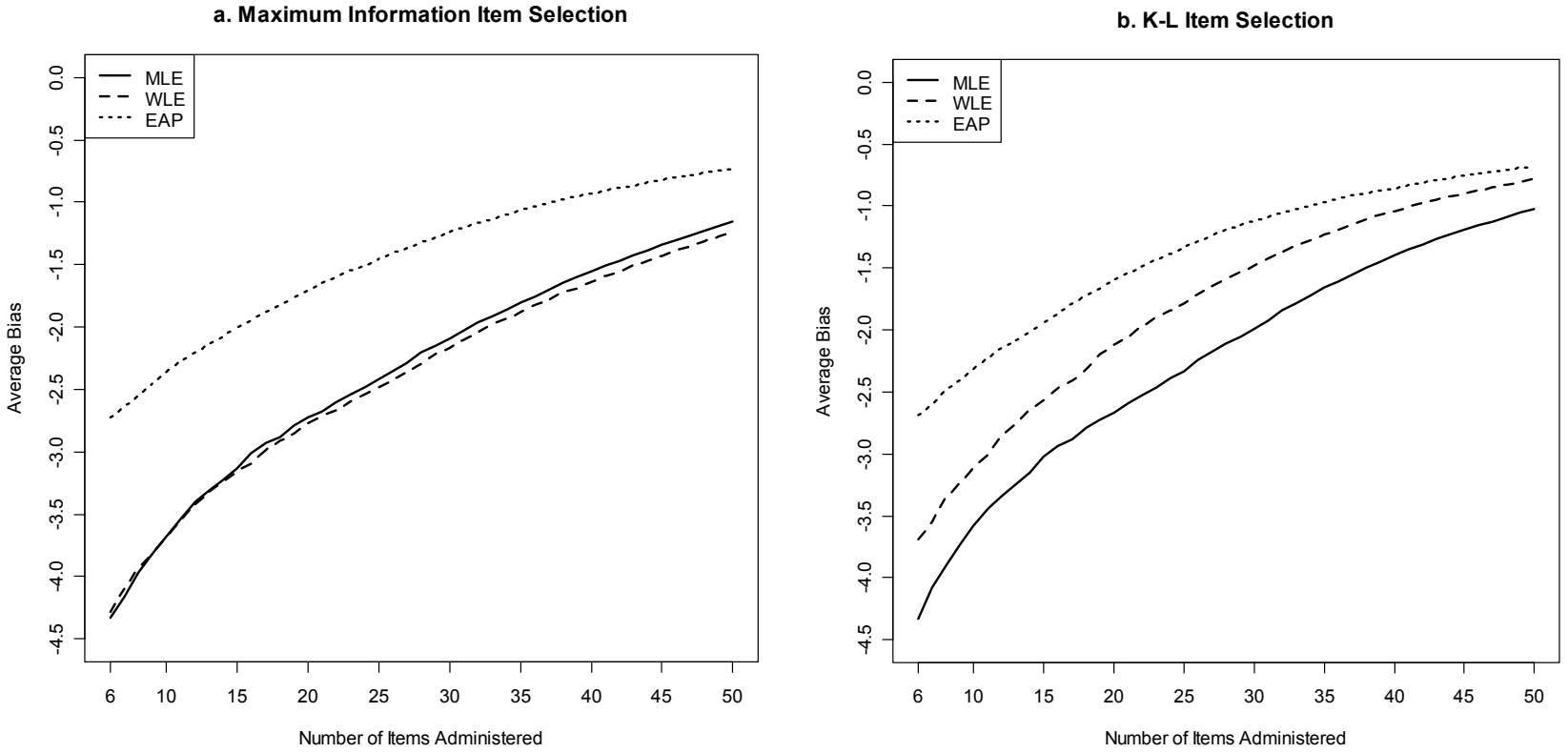


Figure 20

Average Bias Across CAT Lengths for the 1-Item Misfit Condition for θ = –3 (MCR)

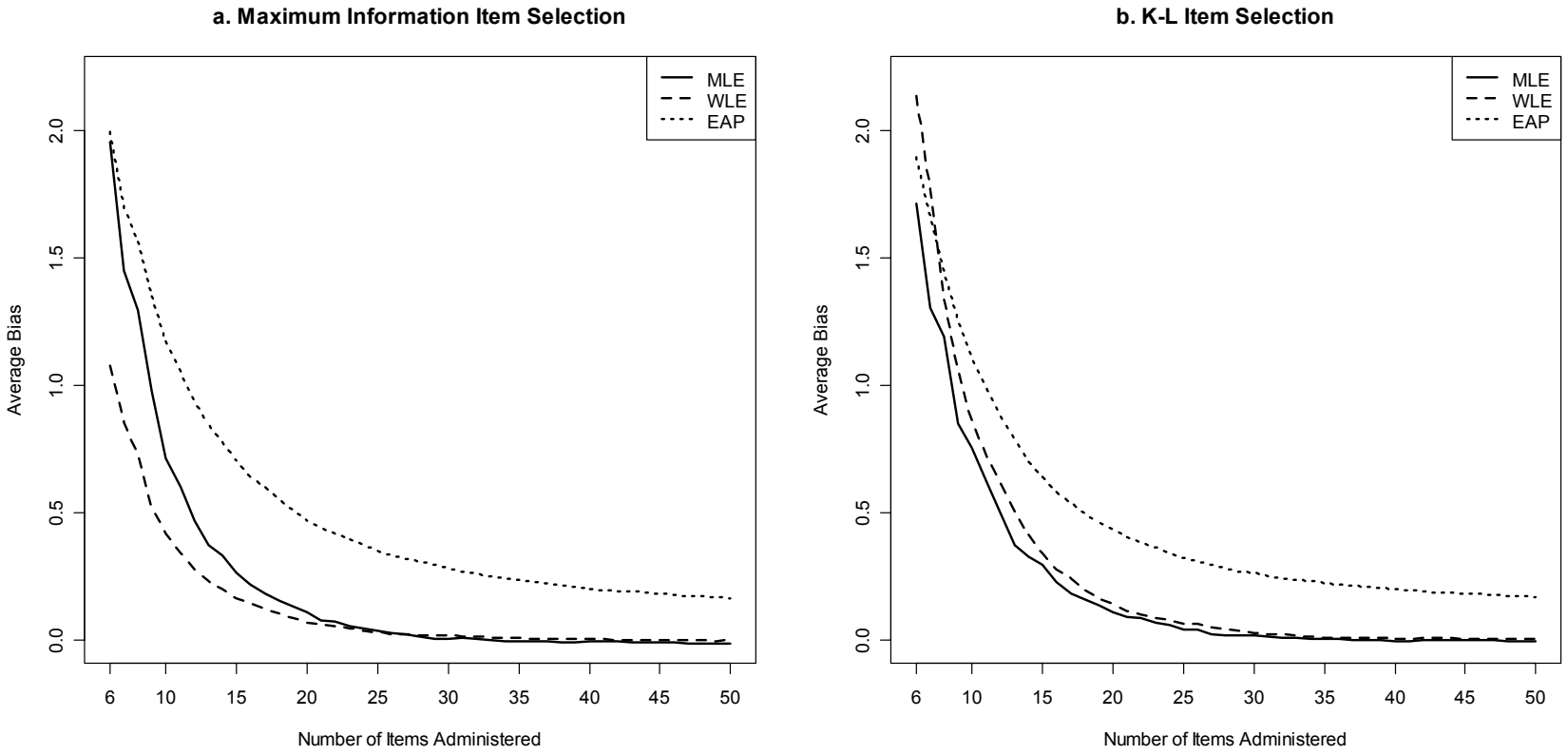


Figure 21

Average Bias Across CAT Lengths for the 1-Item Misfit Condition for θ = –1 (MCR)

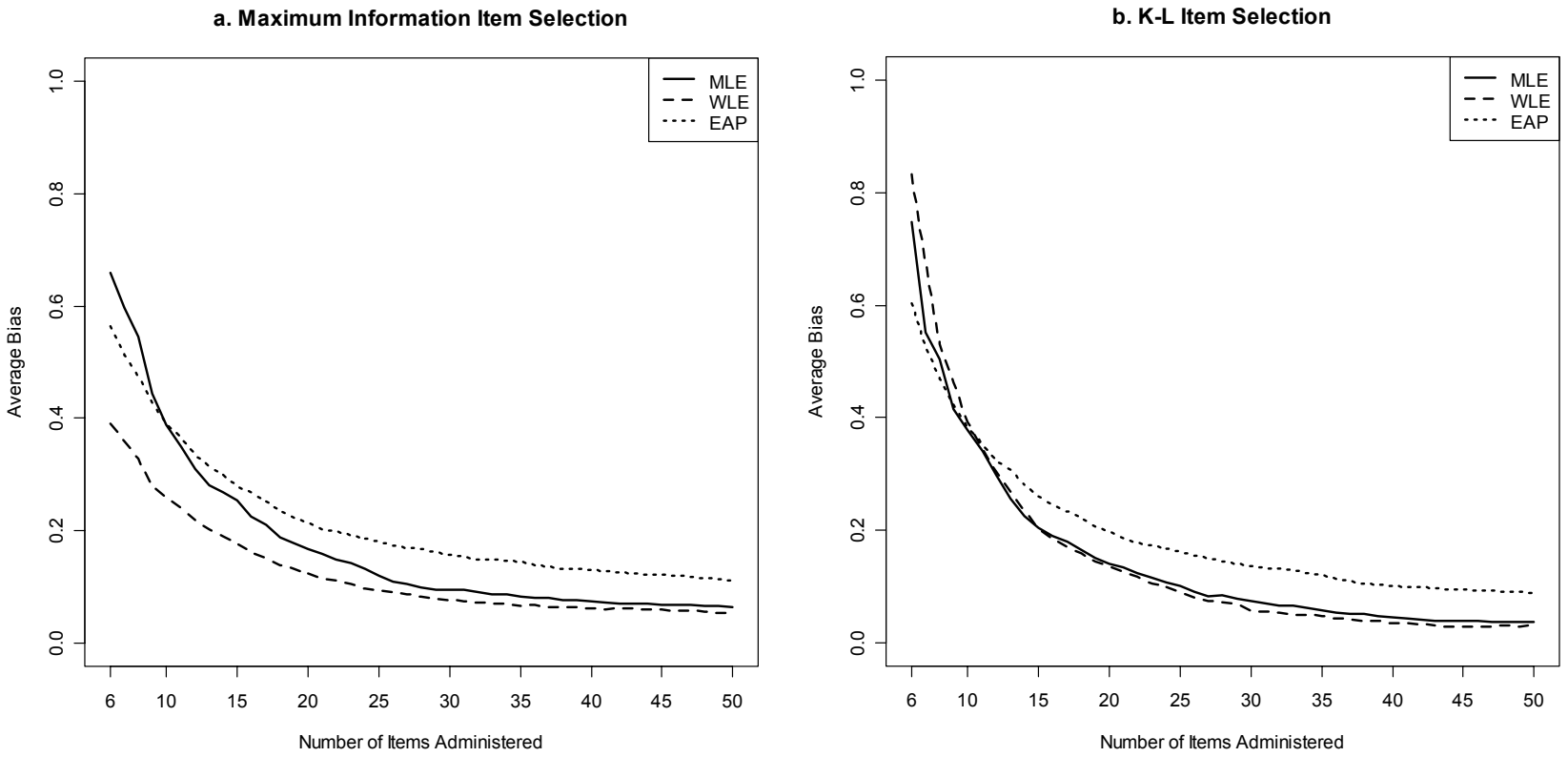


Figure 22

Average Bias Across CAT Lengths for the 2-Item Misfit Condition for θ = –3 (MCR)

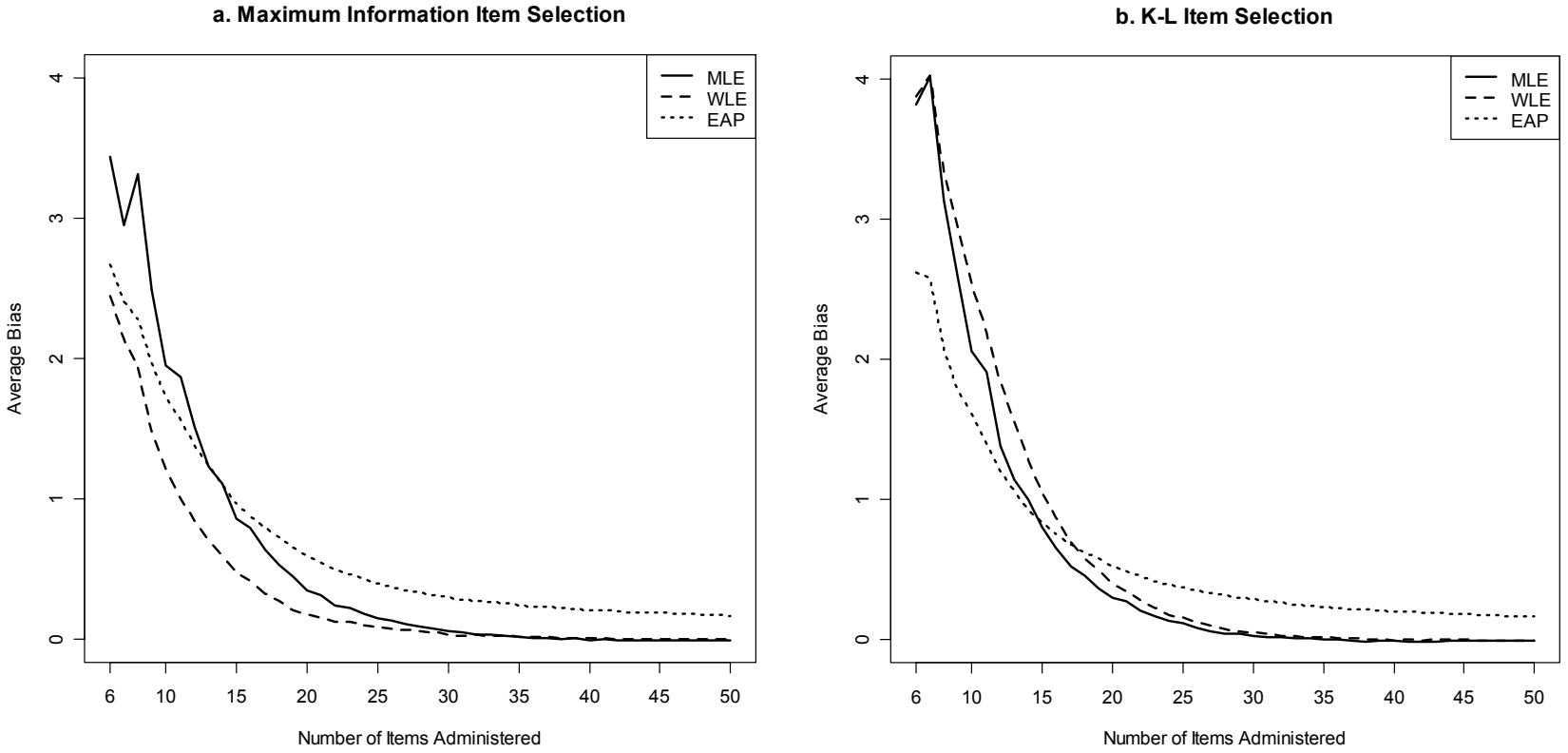


Figure 23

Average Bias Across CAT Lengths for the 2-Item Misfit Condition for θ = –1 (MCR)

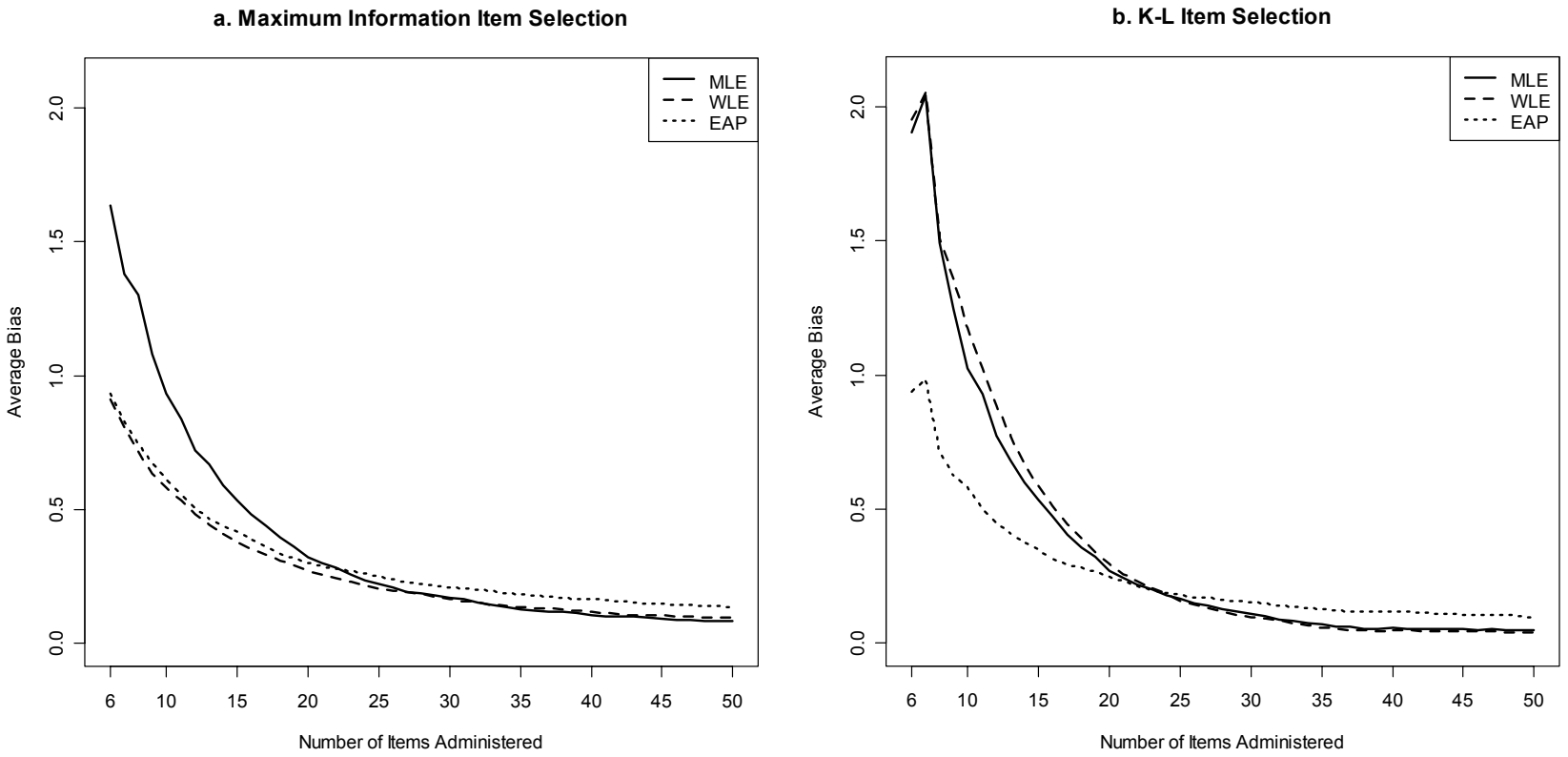


Figure 24

Average Bias Across CAT Lengths for the 3-Item Misfit Condition for θ = –3 (MCR)

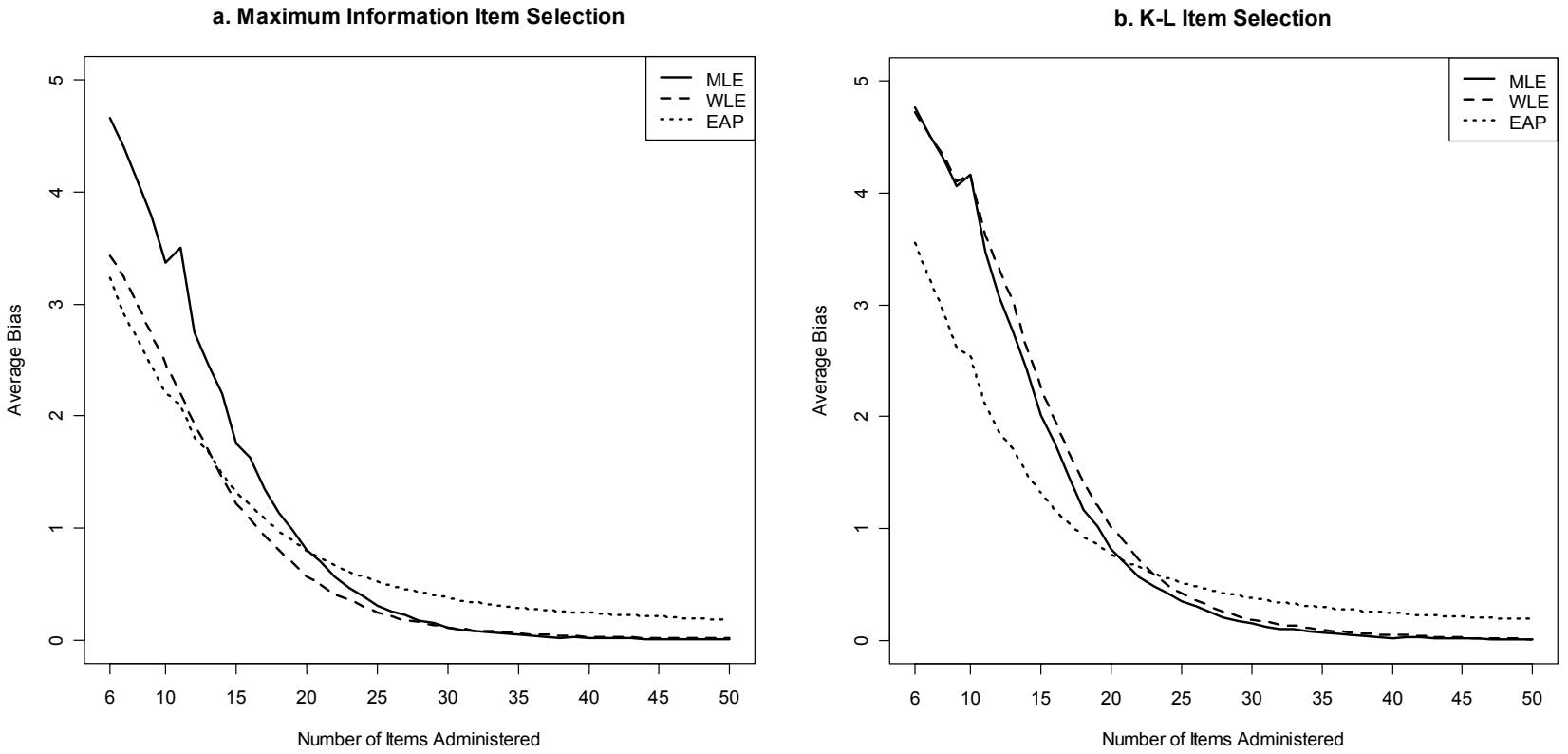


Figure 25

Average Bias Across CAT Lengths for the 3-Item Misfit Condition for θ = –1 (MCR)

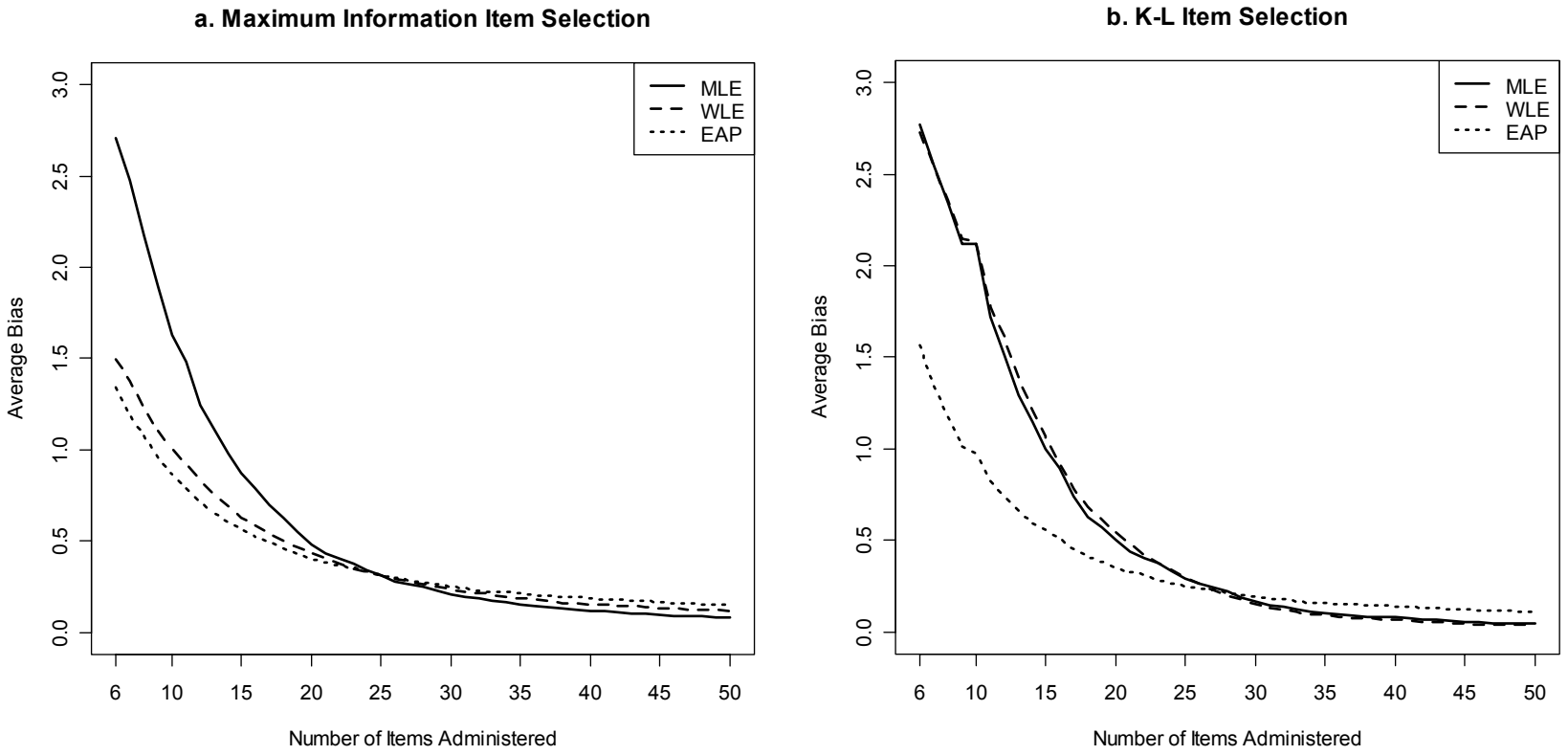


Figure 26

Average Bias Across CAT Lengths for the 4-Item Misfit Condition for θ = –3 (MCR)

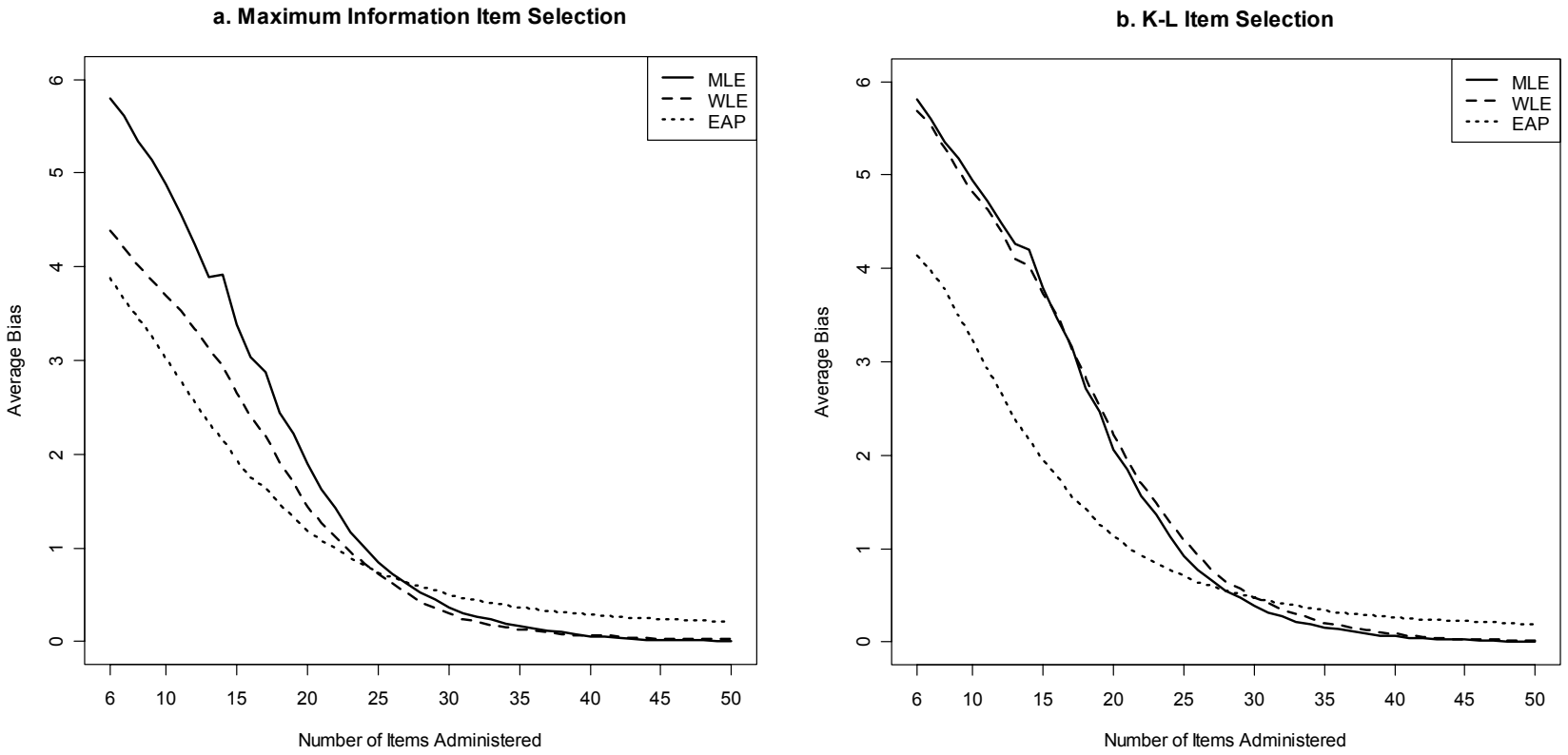
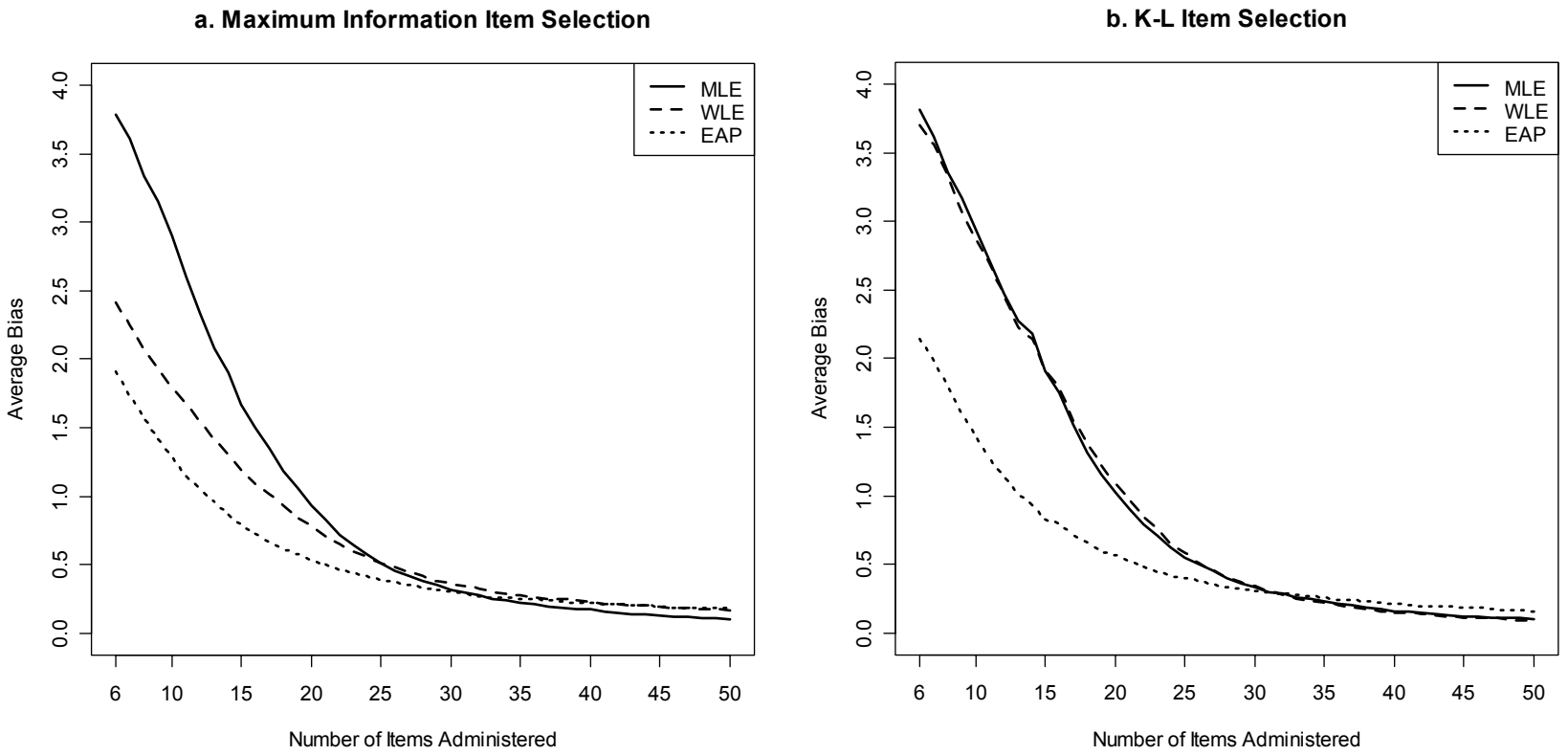


Figure 27

Average Bias Across CAT Lengths for the 4-Item Misfit Condition for θ = –1 (MCR)



Initial Items Selected in the CAT

The first item selected in the CAT differed between FI selection and K-L selection. An item with a *b* parameter of –0.440 was selected for FI selection, whereas, K-L selected an initial item with a *b* of 0.484. Both item selection procedures used an initial θ of 0.0.

Effect on EAP. To better understand these differences, the item parameters for the first five items selected during the CAT, and the θ estimates for each method, are provided. As seen in Tables 5 and 6, for the MIR and MCR conditions, respectively, the EAP θ estimates were not as extreme in absolute value as the WLE estimates. The difference in the difficulty of the first item did not affect the EAP θ estimate as much as it did WLE.

Effect on WLE. WLE was quite sensitive to the difficulty of the initial item. When FI selection was used for the MIR condition (Table 5), $\hat{\theta} = -1.108$ after the first item, while $\hat{\theta} = -0.111$ for K-L selection. This trend continued, as the bias was almost one unit higher when FI was used instead of K-L. For the MCR condition (Table 6), when FI selection was used $\hat{\theta} = 0.040$ after the first item, but $\hat{\theta} = 0.980$ for K-L selection. The discrepancy between the results for FI ($\hat{\theta} = 1.878$) and K-L ($\hat{\theta} = 3.036$) became even more pronounced after the fourth item was administered. These discrepancies might account for the observed differences in the results across item selection methods when WLE was used to estimate θ.

Table 5

Item Parameters for the First Five Items Selected and θ Used to Select the Item for the 4-Item MIR Conditions

Selection	MLE				WLE				EAP			
and Item #	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$
Max. Info.												
1	1.118	−.440	.196	0	1.118	−.440	.196	0	1.118	−.440	.196	0
2	1.301	−1.535	.224	−1	1.301	−1.535	.224	−1.108	1.167	−.831	.198	−.754
3	1.338	−2.274	.206	−2	1.338	−2.274	.206	−2.068	1.301	−1.535	.224	−1.202
4	1.570	−3.179	.194	−3	1.570	−3.179	.194	−2.815	1.221	−1.691	.166	−1.664
5	1.005	−3.405	.193	−4	1.126	−3.176	.170	−3.644	1.338	−2.274	.206	−1.959
K-L												
1	1.086	.484	.192	0	1.086	.484	.192	0	1.086	.484	.196	0
2	1.301	−1.535	.224	−1	1.118	−.440	.196	−.111	1.167	−.831	.198	−.429
3	1.338	−2.274	.206	−2	1.301	−1.535	.224	−1.071	1.301	−1.535	.224	−1.036
4	1.570	−3.179	.194	−3	1.338	−2.274	.206	−2.077	1.221	−1.691	.166	−1.594
5	1.005	−3.405	.193	−4	1.527	−3.179	.194	−2.817	1.338	−2.274	.206	−1.924

Table 6

Item Parameters for the First Five Items Selected and θ Used to Select the Item for the 4-Item MCR Conditions

Selection	MLE				WLE				EAP			
and Item #	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$
Max. Info.												
1	1.118	−.440	.196	0	1.118	−.440	.196	0	1.118	−.440	.196	0
2	1.151	.969	.175	1	1.005	−.289	.198	.040	1.086	.484	.192	.322
3	1.243	1.878	.213	2	1.086	.484	.192	.408	1.151	.969	.175	.681
4	1.250	2.600	.197	3	1.151	.969	.175	1.048	1.031	1.155	.161	1.042
5	1.115	3.183	.222	4	1.243	1.878	.213	1.569	1.053	1.039	.202	1.337
K-L												
1	1.086	.484	.192	0	1.086	.484	.192	0	1.086	.484	.192	0
2	1.151	.969	.175	1	1.151	.969	.175	.980	1.151	.969	.175	.458
3	1.243	1.878	.213	2	1.243	1.878	.213	1.534	1.031	1.156	.161	.905
4	1.250	2.600	.197	3	1.250	2.600	.197	2.326	1.053	1.039	.202	1.264
5	1.115	3.183	.222	4	1.147	3.036	.178	3.047	1.243	1.878	.213	1.494

Empirical SE

Conditions Without Misfit

The empirical standard errors for each of the cells in the design are presented in Tables A8–A47. A general trend observed was that the difference between MLE and WLE became smaller as test length increased from 15 to 50 items. As expected, EAP had the lowest SE values regardless of test length or θ. In addition, the SE values were higher early in the CAT for θ values less than 0 than for θ values greater than 0.

When 15 to 35 items were administered in the CAT, the performance of WLE versus MLE differed depending on the value of θ when FI was used to select items. After 15 items were administered, WLE had lower SEs than MLE for $\theta = -3$ to 0, while MLE had lower SEs than WLE for $\theta = 1$ to 3. For test lengths of 25 and 35 items, WLE had lower empirical SEs than MLE for $\theta = -3$ to –1, while WLE and MLE had nearly equal SEs for $\theta = 0$ to 3. When 50 items were administered, there were no systematic differences in the SEs between WLE and MLE.

The differences between the SEs for WLE and MLE became smaller when K-L information was used to select items in the CAT. This result is best summarized graphically. The empirical standard errors were computed for test lengths from 6 to 50 items for $\theta = \pm 3$ and ± 1 and are shown by Figures 28 and 29. EAP had the lowest SEs for both item selection methods. For K-L selection the standard errors were consistently as follows: MLE > WLE > EAP.

For $\theta = 3$ (Figure 28) the standard errors for WLE with FI selection were higher than MLE until 20 items were administered in the CAT. When K-L selection was used WLE had slightly lower SEs than MLE. For $\theta = -1$ and -3 (Figures 30 and 31) the SEs for

MLE, WLE, and EAP were higher when K-L selection was compared to FI. For $\theta = -3$, the standard errors of the WLE and MLE θ estimates were quite similar across test lengths when K-L was used to select items. For FI selection, the SEs of WLE were lower than MLE across all test lengths.

The SEs for $\theta = 3$ (Figure 28) were initially lower than the SEs for $\theta = 1$ (Figure 29). This difference dissipated as test length increased. After 50 items were administered, the SEs for $\theta = 1$ were less than the SEs for $\theta = 3$, as seen in Tables A44 and A47. A different trend emerged for $\theta = -3$ (Figure 31) as the SEs were consistently higher than the SEs for $\theta = -1$ (Figure 30).

Figure 28
Empirical SE Across CAT Lengths for the 0-Item Misfit Condition for θ = 3 (MIR)

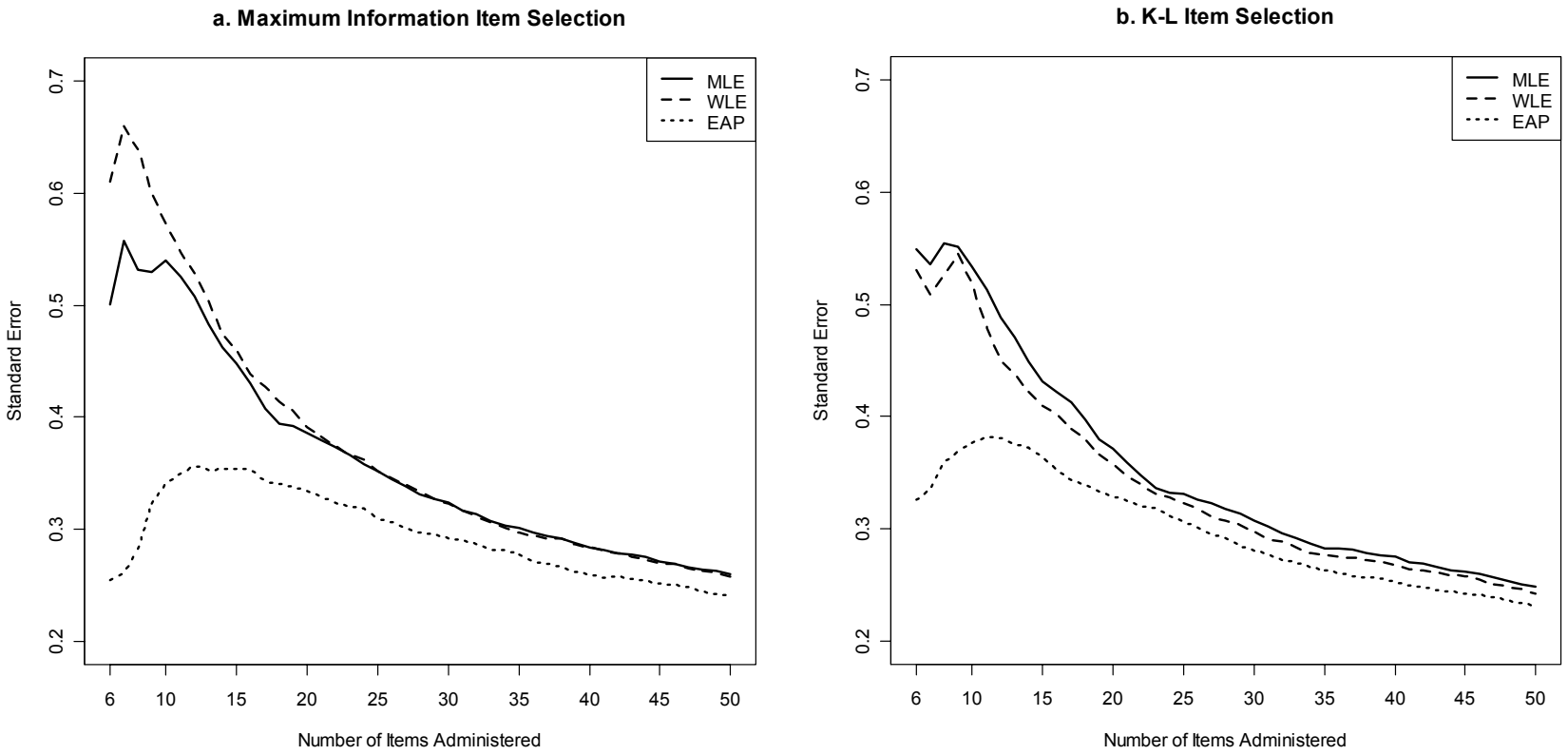


Figure 29
Empirical SE Across CAT Lengths for the 0-Item Misfit Condition for θ = 1

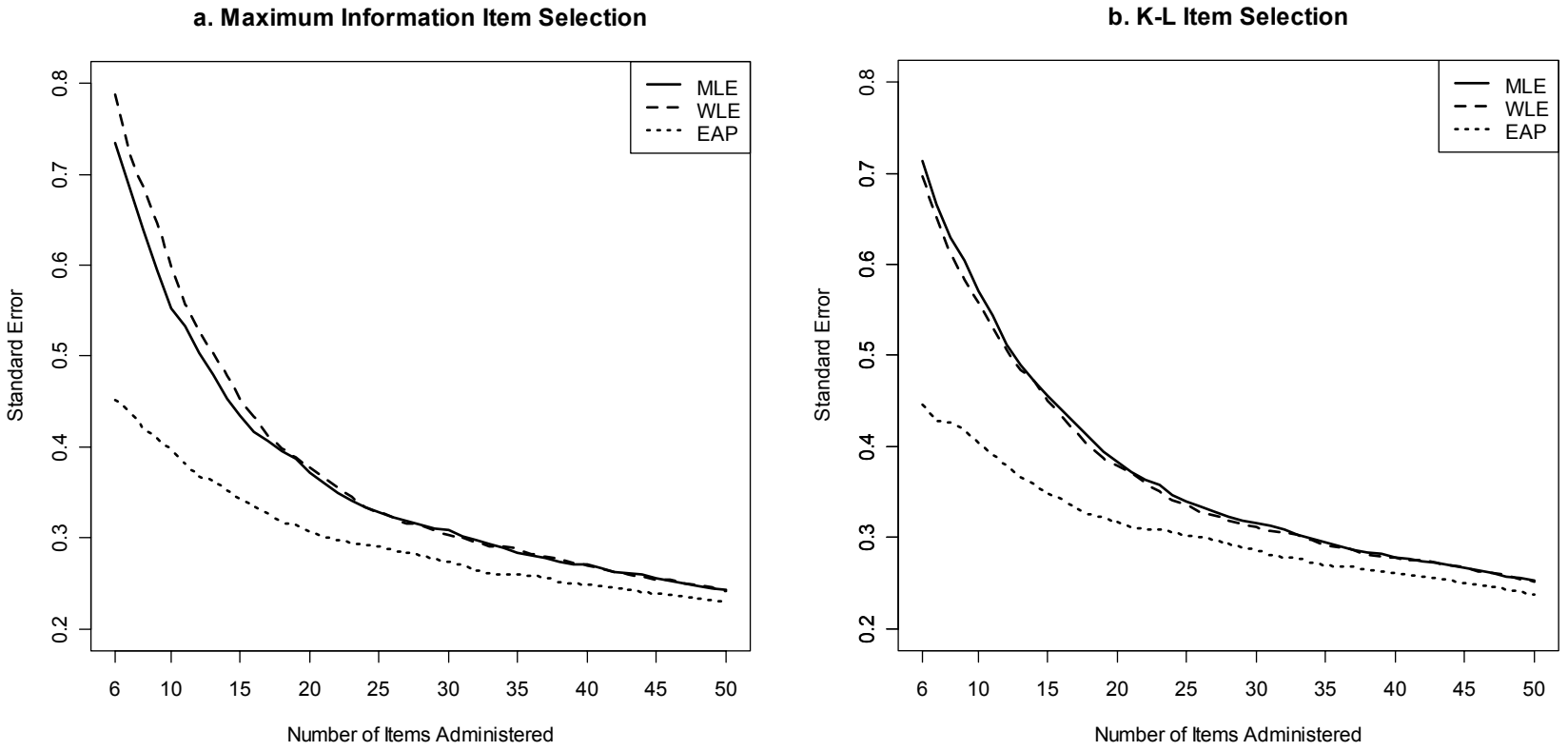


Figure 30
Empirical SE Across CAT Lengths for the 0-Item Misfit Condition for θ = –1

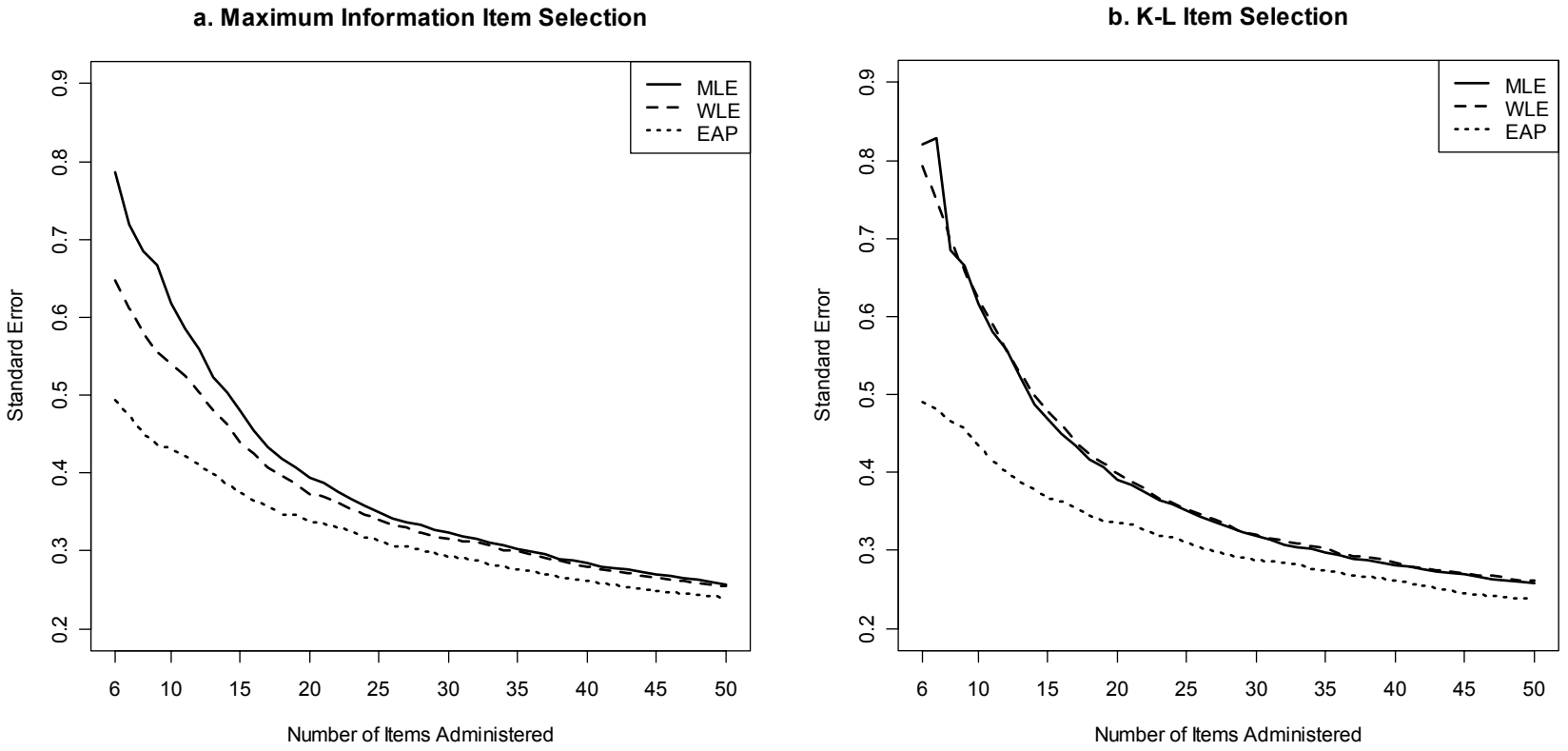
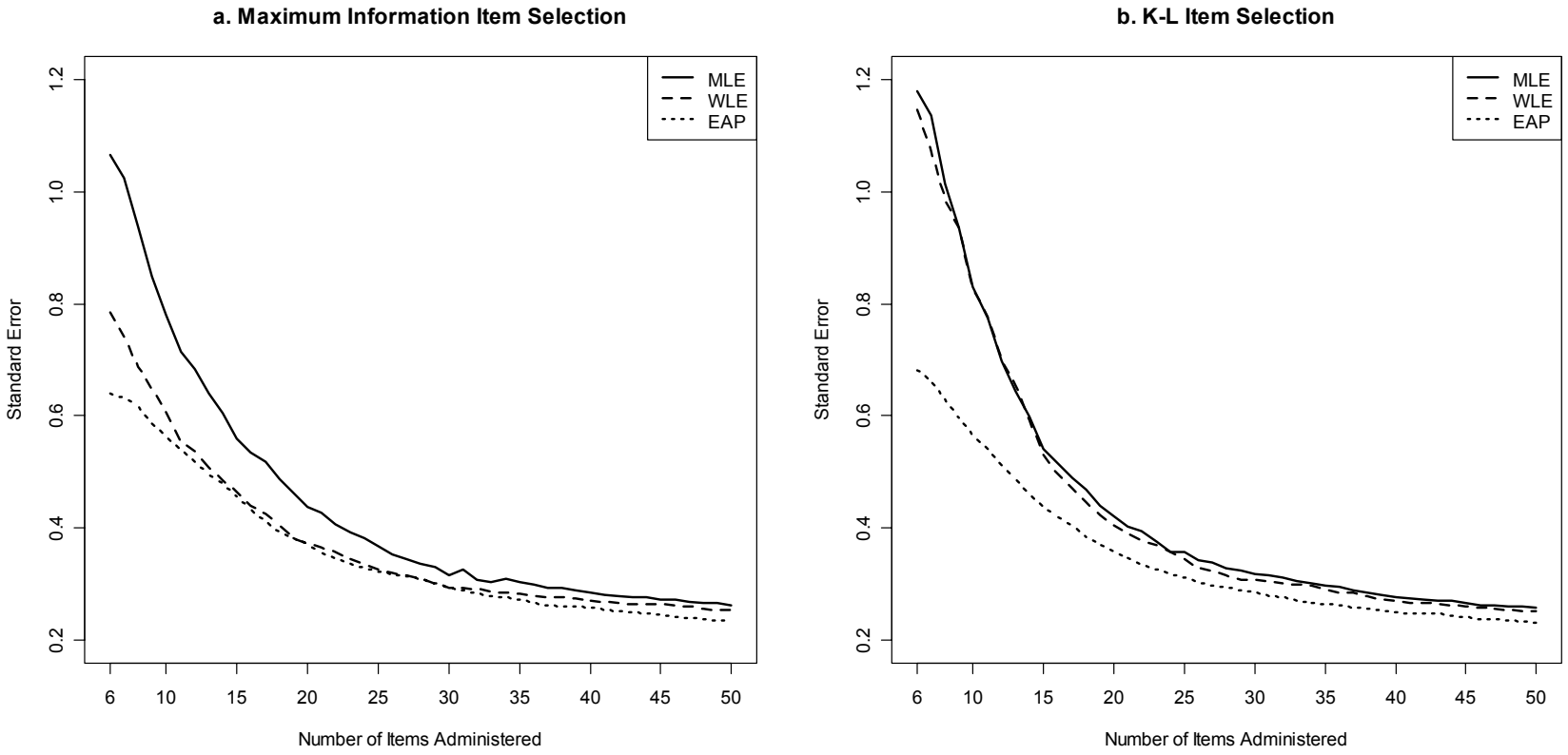


Figure 31
Empirical SE Across CAT Lengths for the 0-Item Misfit Condition for $\theta = -3$



MIR

Differences between the θ estimation methods were examined graphically by the number of misfitting items and test length. The empirical SEs for $\theta = 3$ and 1 were obtained for test lengths of 6 to 50 items across item selection and θ estimation method.

1 misfitting item. For $\theta = 3$ (Figure 32), the SEs for the three θ estimation methods increased until about 20 items were administered. It can be seen in Figure 33 ($\theta = 1$) that the SEs of the three θ estimation methods increased until about 10 to 15 items were administered. As with the no-misfit condition, it was observed that the SEs were ordered as follows: MLE > WLE > EAP. Few differences existed across item selection method; in general, K-L selection resulted in slightly higher SEs than FI selection.

2 misfitting items. The SEs for EAP estimation were greater than WLE or MLE early in the CAT when there were two misfitting items, as seen in Figures 34 and 35. The difference between EAP and WLE or MLE was most prevalent when FI selection was used. The SEs of MLE and WLE for the $\theta = 1$ condition (Figure 35) increased for the first 20 to 25 items when FI selection was used, then began to decrease. The SEs of EAP peaked earlier and began to decrease earlier across item selection methods.

When $\theta = 3$ (Figure 34), the SEs for all three estimation methods increased as the test length was increased from 6 to 40 items. From 40 items to 50 items the SEs began to decrease slightly. It seemed that the SEs for WLE were sensitive to item selection method, as they were greater for K-L selection than FI selection.

3 and 4 misfitting items. The SEs for all three θ estimation methods increased as CAT length increased, as seen in Figures 36–39. For the 3-item misfit condition with $\theta = 3$ (Figure 36), EAP had larger SEs early in the CAT for K-L selection, and throughout the

CAT for FI selection. When $\theta = 1$, EAP had larger SEs than MLE or WLE until about 35 items (FI, Figure 37a) or 25 items (K-L, Figure 37b) were administered. The SEs for WLE were larger when K-L information was used to select items. As seen in Figures 36–39, WLE was most sensitive to item selection method. The SEs for EAP were greater than MLE when there were four misfitting items and $\theta = 1$ or 3. WLE had larger SEs than EAP only for the 4 misfitting item condition (Figure 39b) when 42 or more items administered with K-L selection and $\theta = 1$.

Figure 32

Empirical SE Across CAT Lengths for the 1-Item Misfit Condition for θ = 3 (MIR)

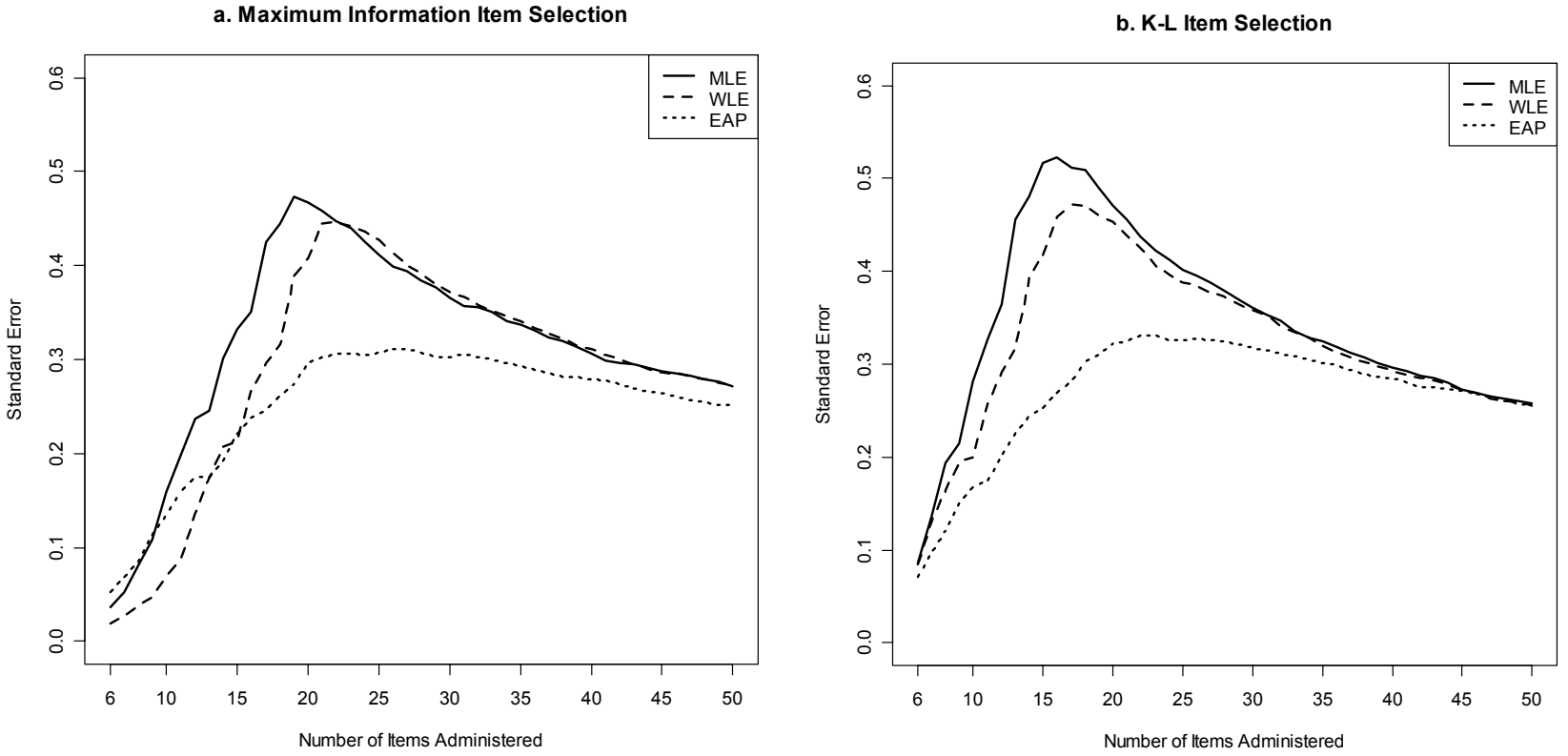


Figure 33
Empirical SE Across CAT Lengths for the 1-Item Misfit Condition for θ = 1 (MIR)

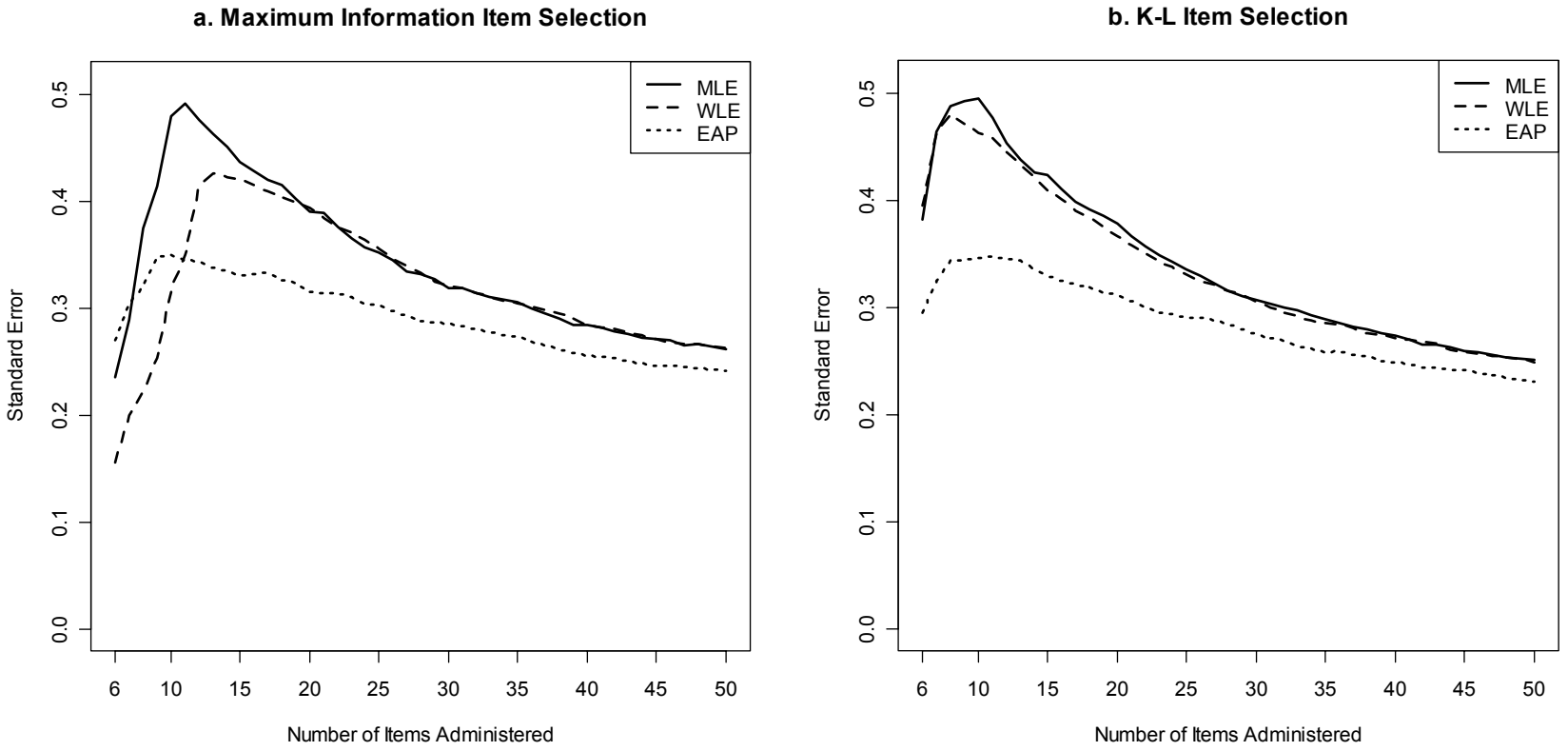


Figure 34

Empirical SE Across CAT Lengths for the 2-Item Misfit Condition for θ = 3 (MIR)

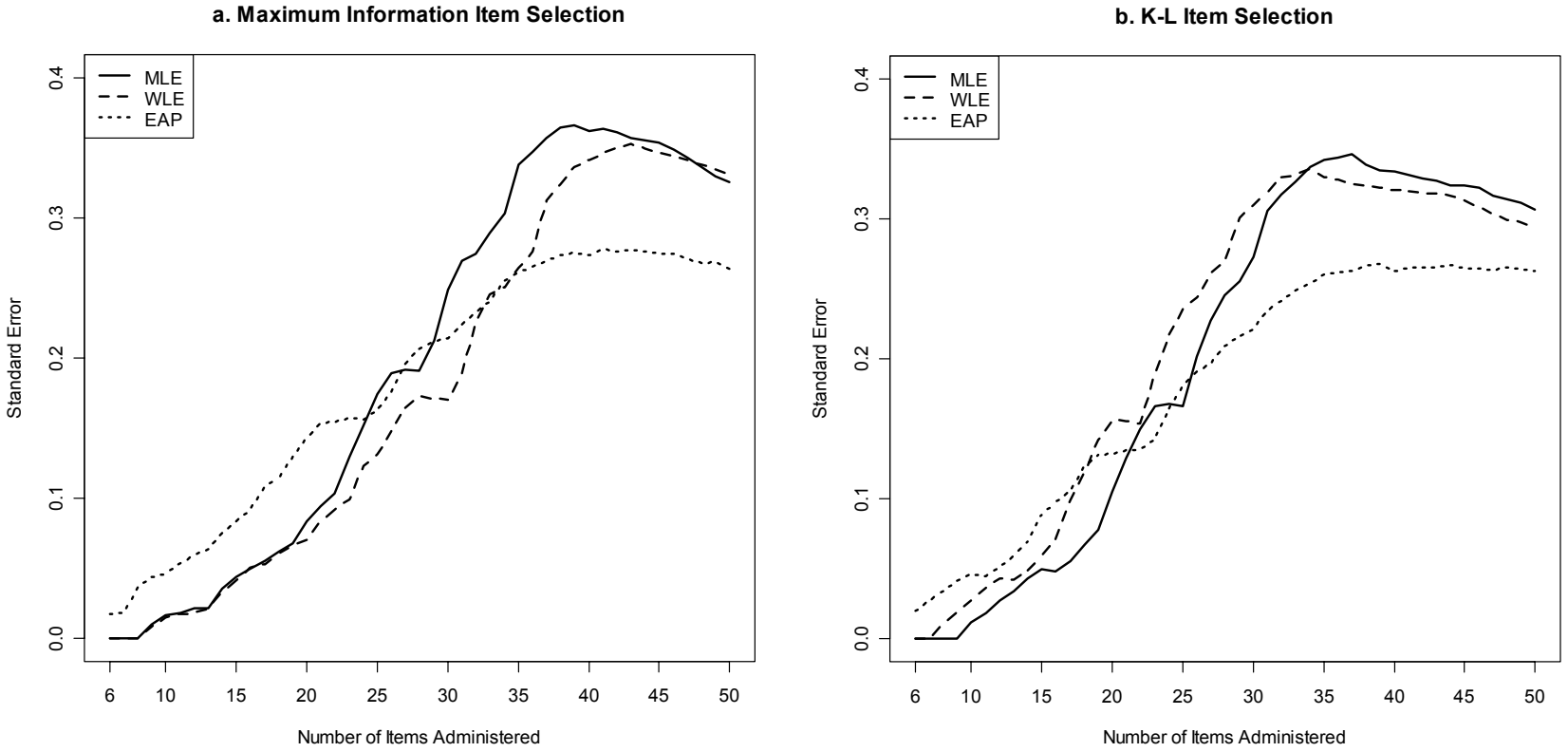


Figure 35

Empirical SE Across CAT Lengths for the 2-Item Misfit Condition for θ = 1 (MIR)

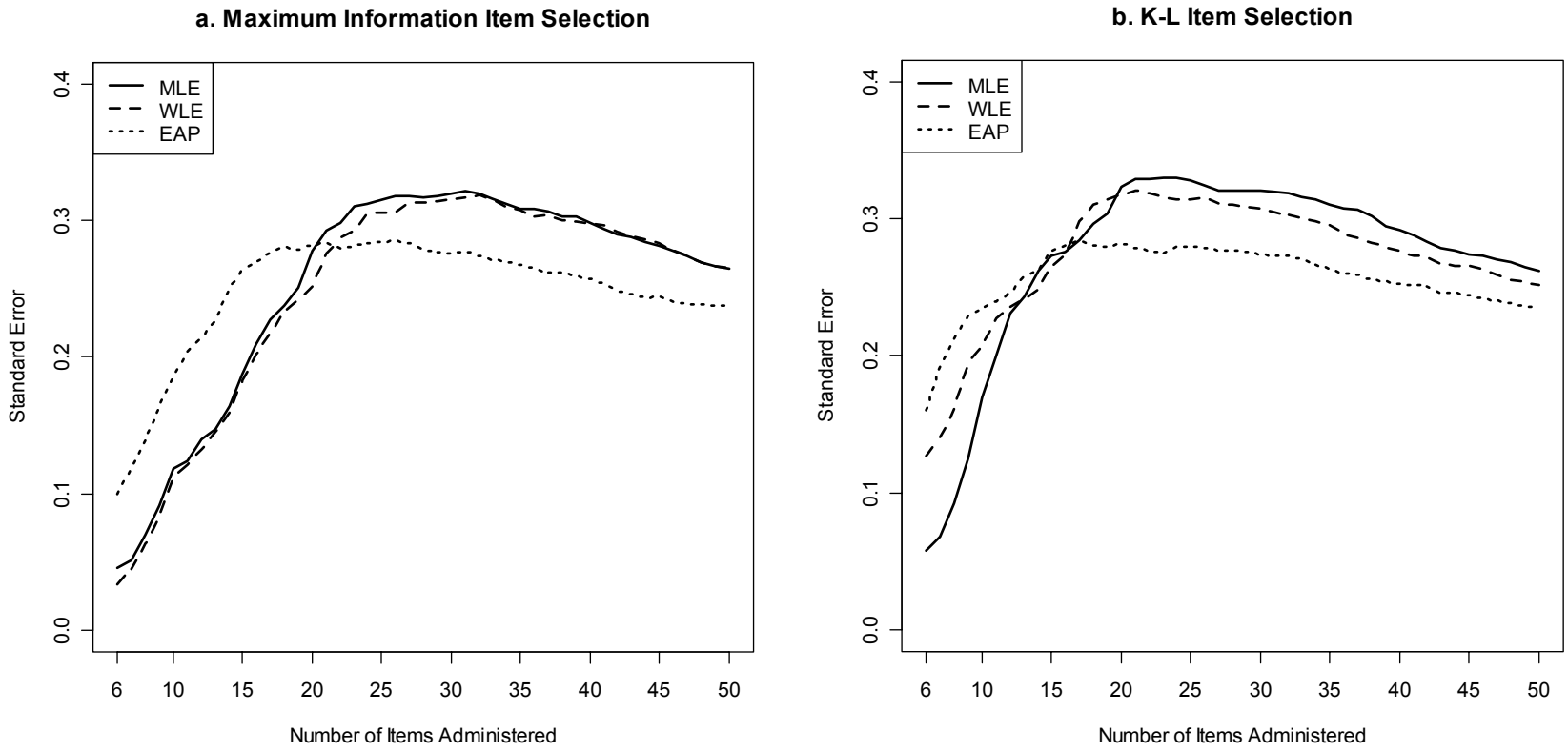


Figure 36

Empirical SE Across CAT Lengths for the 3-Item Misfit Condition for θ = 3 (MIR)

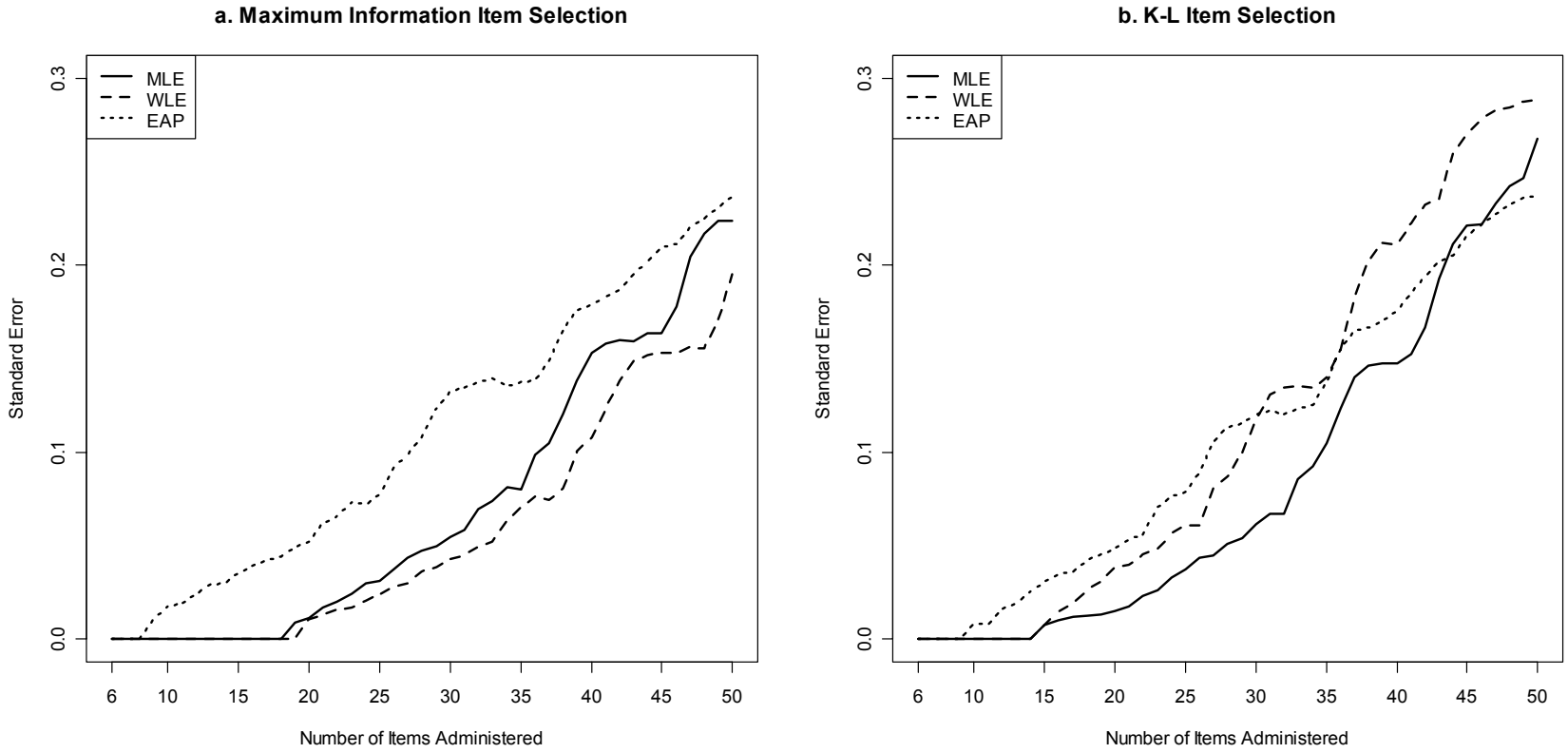


Figure 37

Empirical SE Across CAT Lengths for the 3-Item Misfit Condition for θ = 1 (MIR)

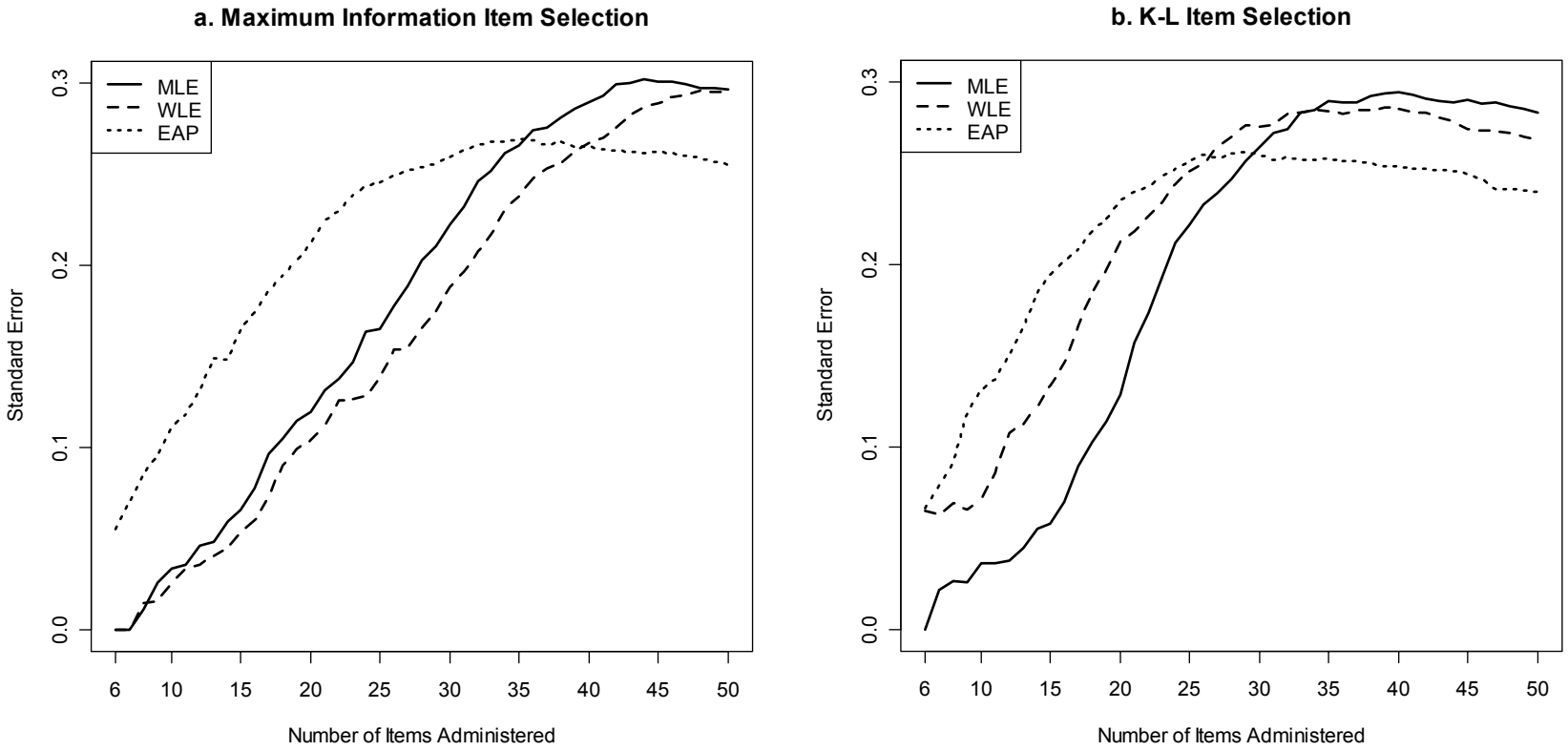


Figure 38

Empirical SE Across CAT Lengths for the 4-Item Misfit Condition for θ = 3 (MIR)

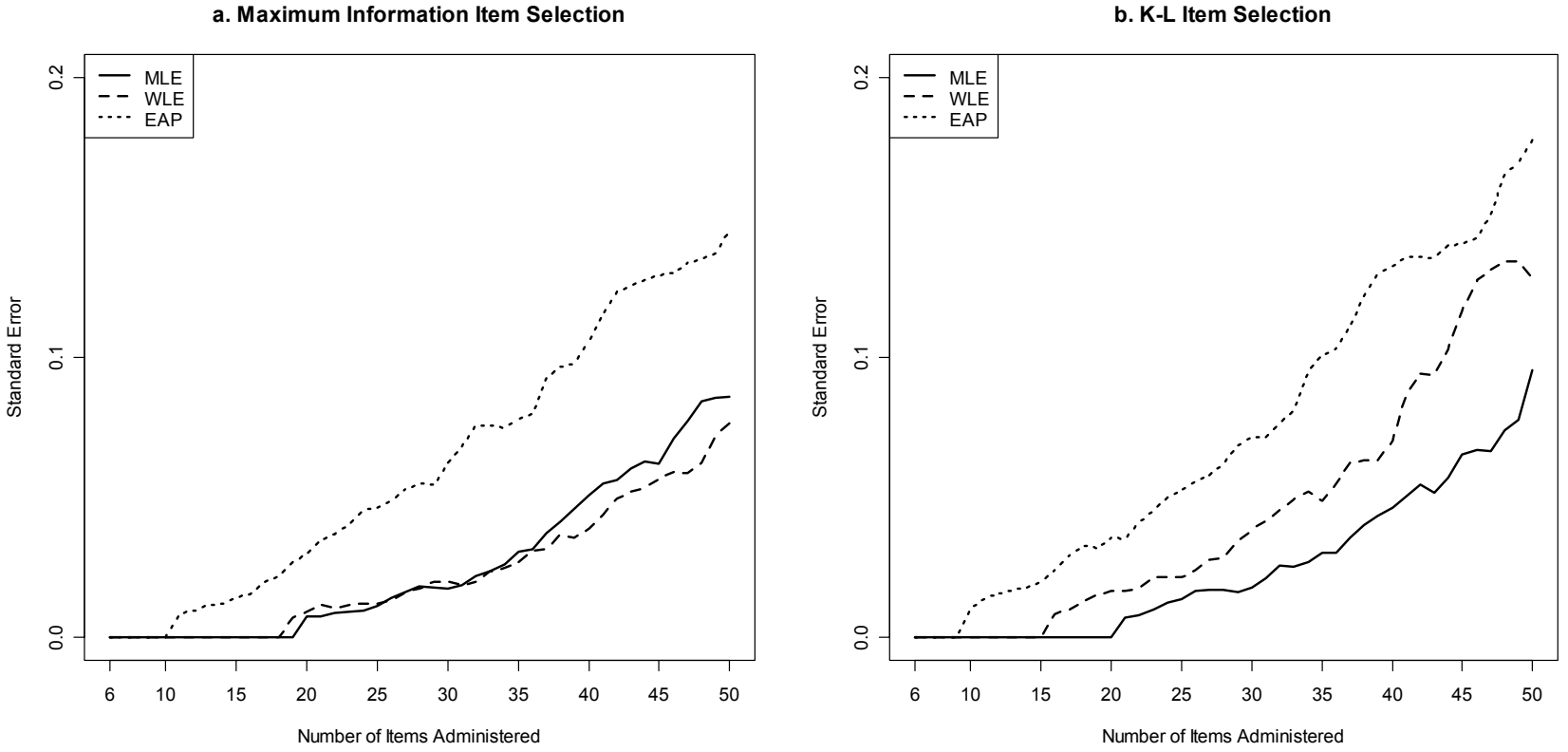
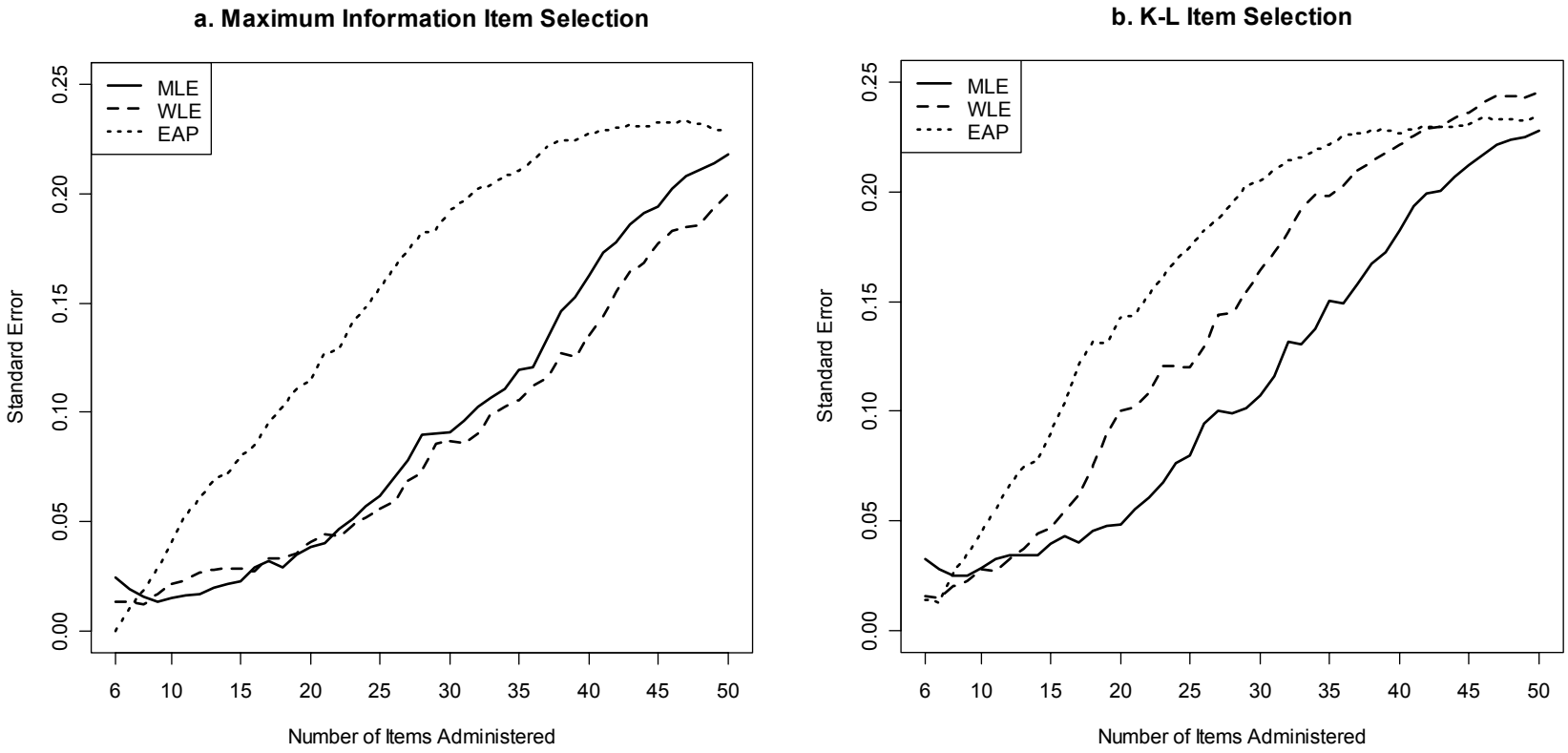


Figure 39

Empirical SE Across CAT Lengths for the 4-Item Misfit Condition for θ = 1 (MIR)



MCR

General trends. The empirical SEs are reported in Tables A8−A14, A18−A24, A28−A34, and A38−A44 for the MCR conditions. It was observed that the SEs of θ increased as more misfit was introduced. This result was consistent across the θ estimation and item selection conditions.

K-L selection resulted in larger SEs across θ estimation methods, as shown by Figures 40−47. In addition, the SEs for the methods were generally ordered as follows: MLE > WLE > EAP. This trend was consistent for conditions with 1 to 4 misfitting items. The SEs for the $\theta = -1$ conditions (Figures 41, 43, 45, and 47) were lower than the SEs for the $\theta = -3$ conditions (Figures 40, 42, 44, and 46).

Number of misfitting items. There was strong evidence that the SEs of the θ estimates peaked after different numbers of items were administered, as seen in Figures 40−47. For FI selection and $\theta = -3$, the SE functions for MLE peaked after 8, 10, 15, and 20 items for the one, two, three, and four misfit conditions, respectively. The maximum of the SEs for MLE were similar across the two (1.71), three (1.71) and four (1.734) misfitting item conditions. Although MLE had generally higher SEs than WLE under K-L information, the differences tended to become negligible after a given number of items, with the number of items increasing as the number of misfitting items increased.

Figure 40

Empirical SE Across CAT Lengths for the 1-Item Misfit Condition for θ = –3 (MCR)

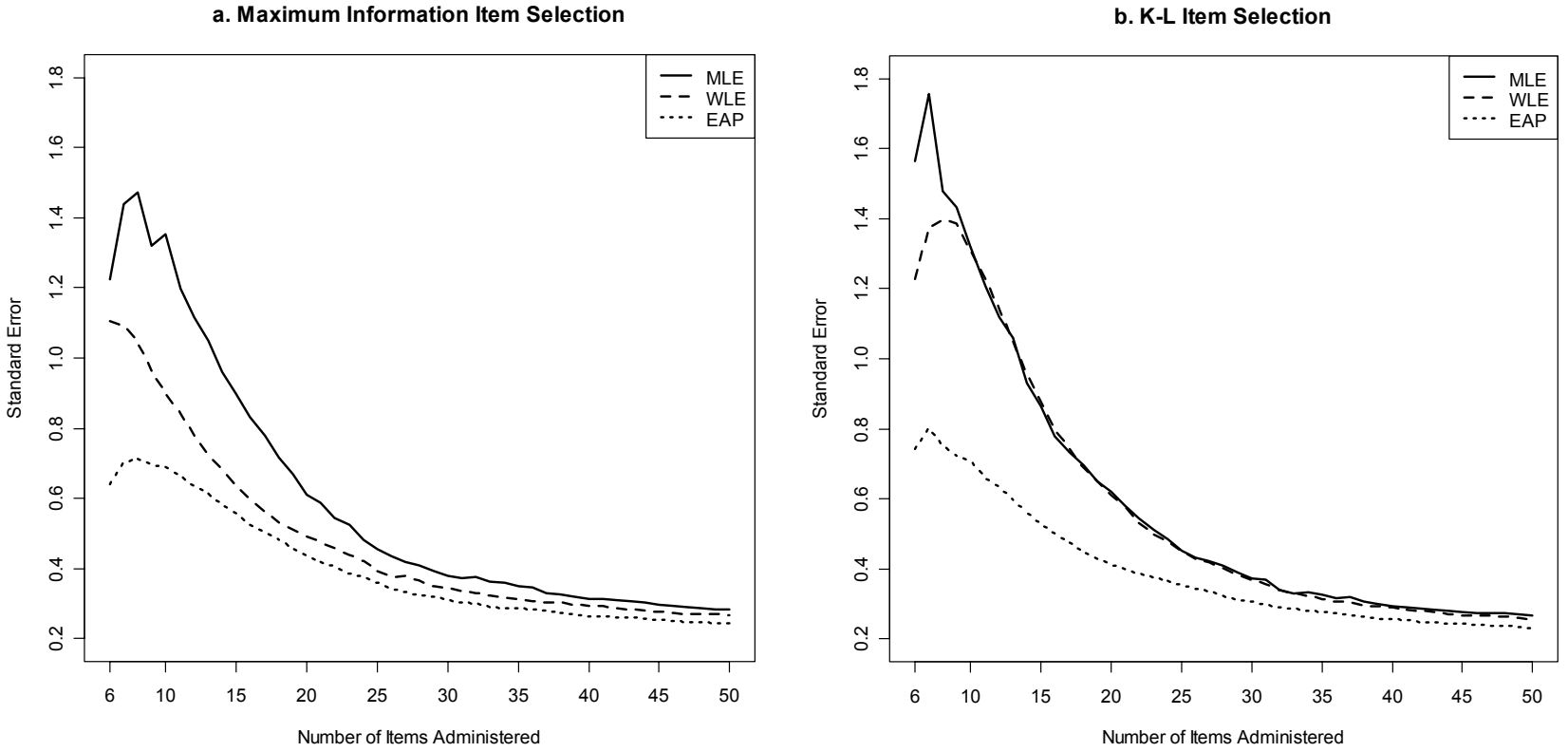


Figure 41

Empirical SE Across CAT Lengths for the 1-Item Misfit Condition for θ = –1 (MCR)

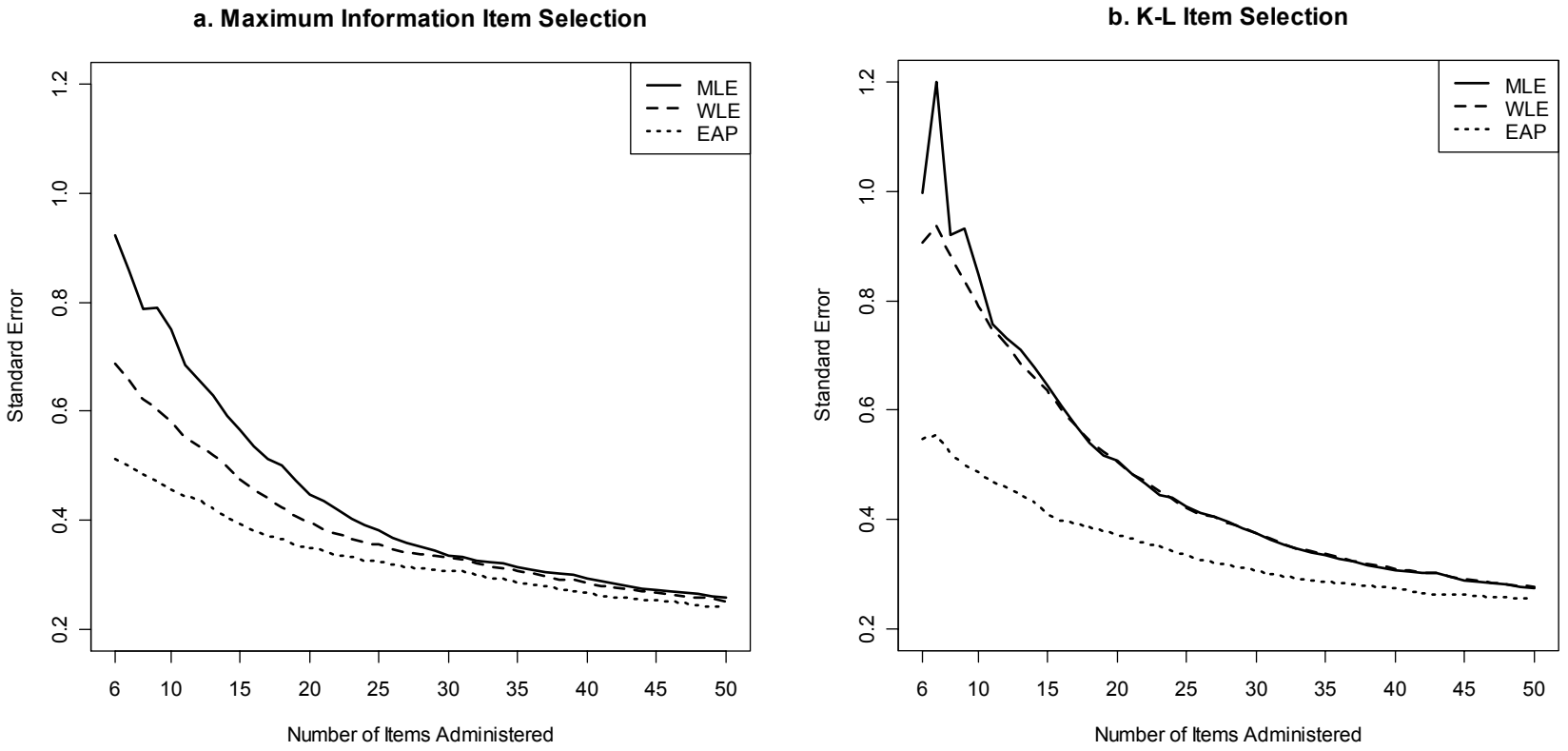


Figure 42

Empirical SE Across CAT Lengths for the 2-Item Misfit Condition for θ = –3 (MCR)

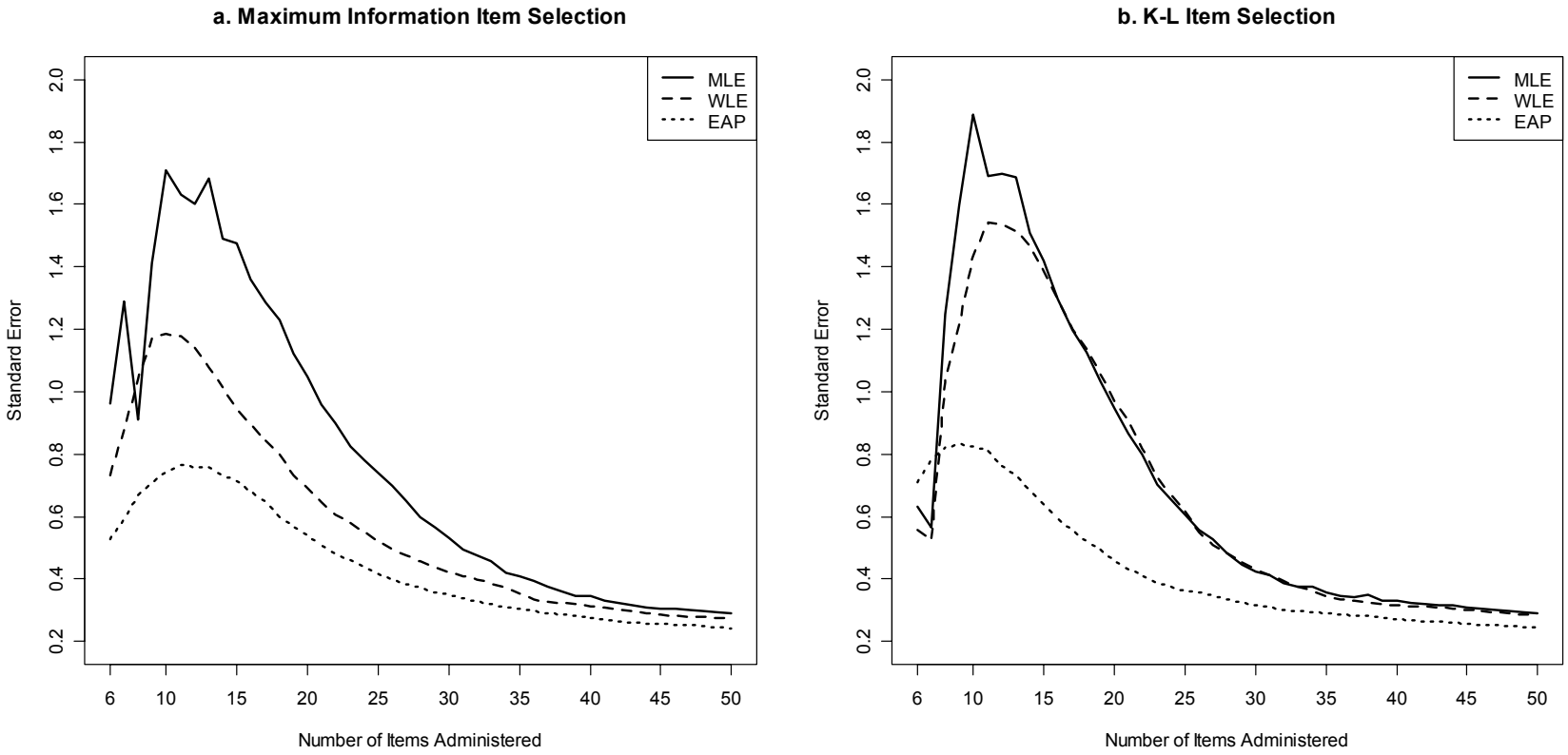


Figure 43
Empirical SE Across CAT Lengths for the 2-Item Misfit Condition for θ = –1 (MCR)

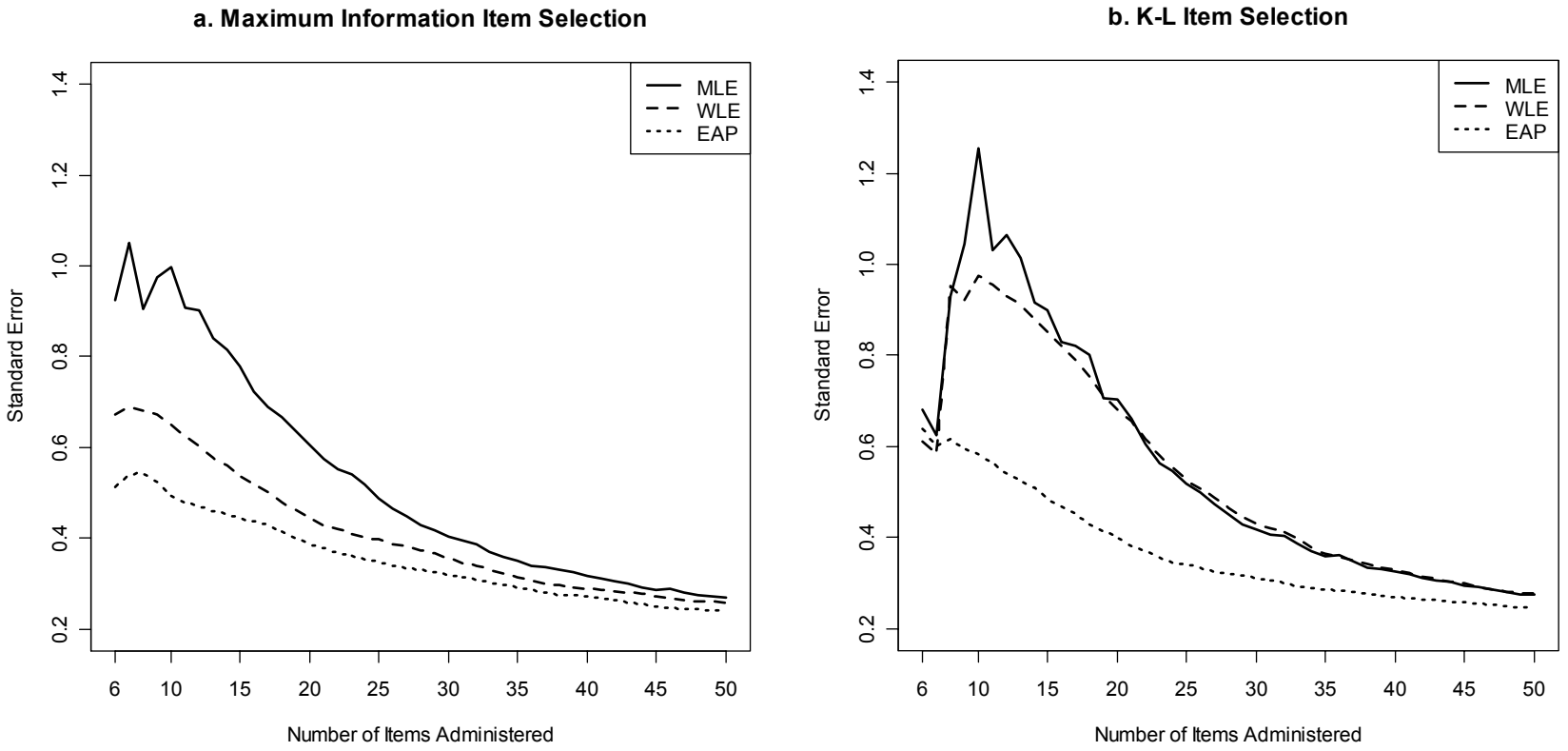


Figure 44

Empirical SE Across CAT Lengths for the 3-Item Misfit Condition for θ = –3 (MCR)

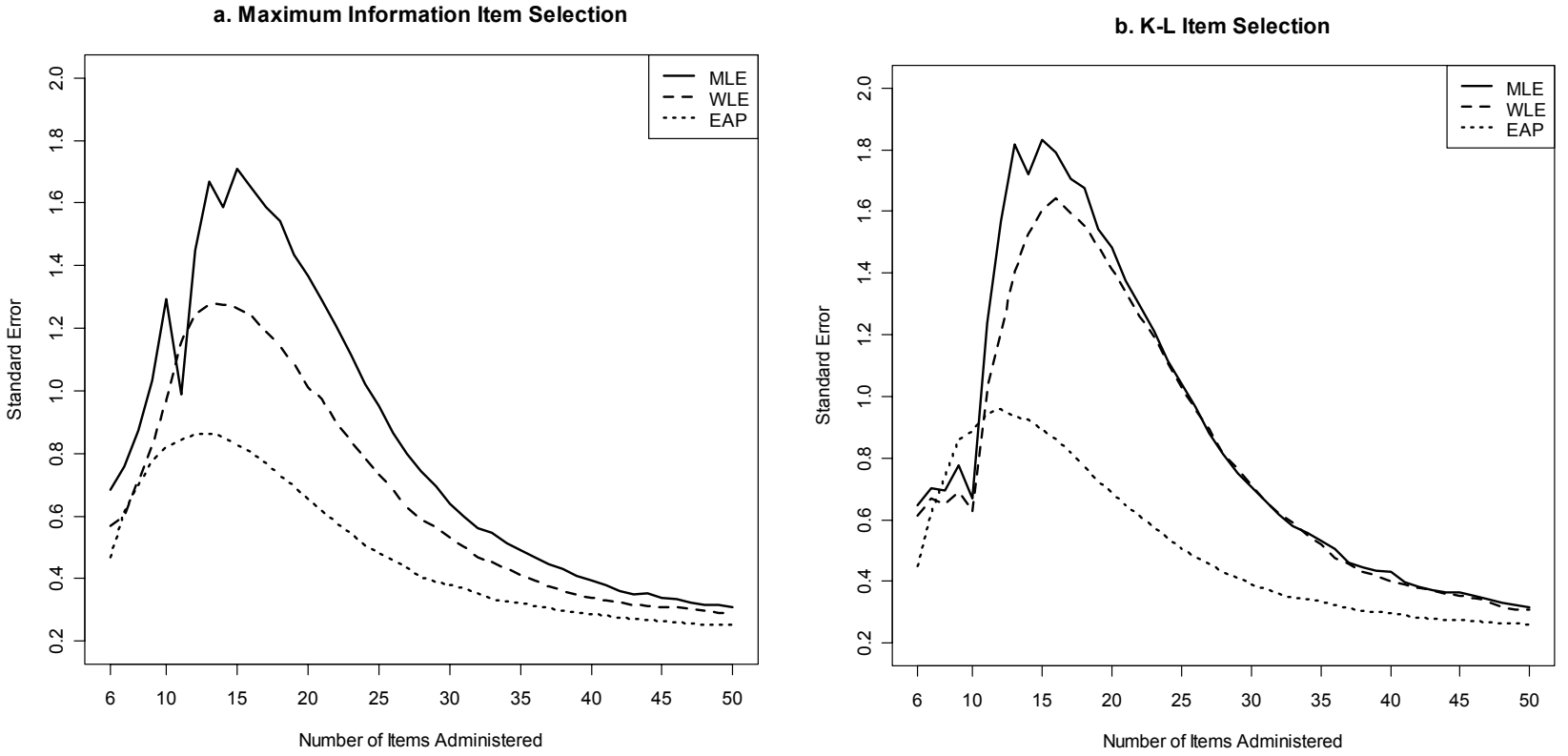


Figure 45
Empirical SE Across CAT Lengths for the 3-Item Misfit Condition for θ = –1 (MCR)

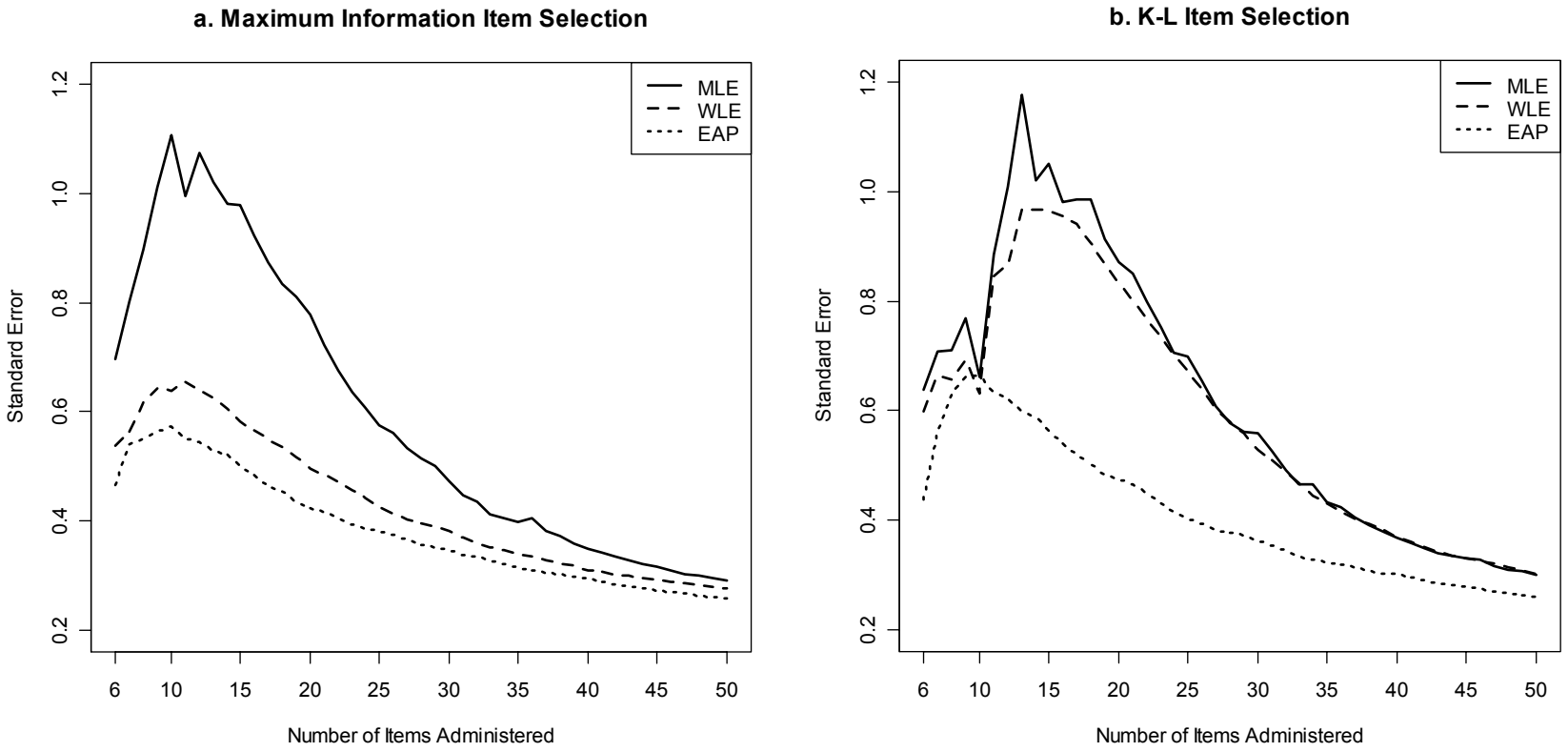


Figure 46

Empirical SE Across CAT Lengths for the 4-Item Misfit Condition for θ = –3 (MCR)

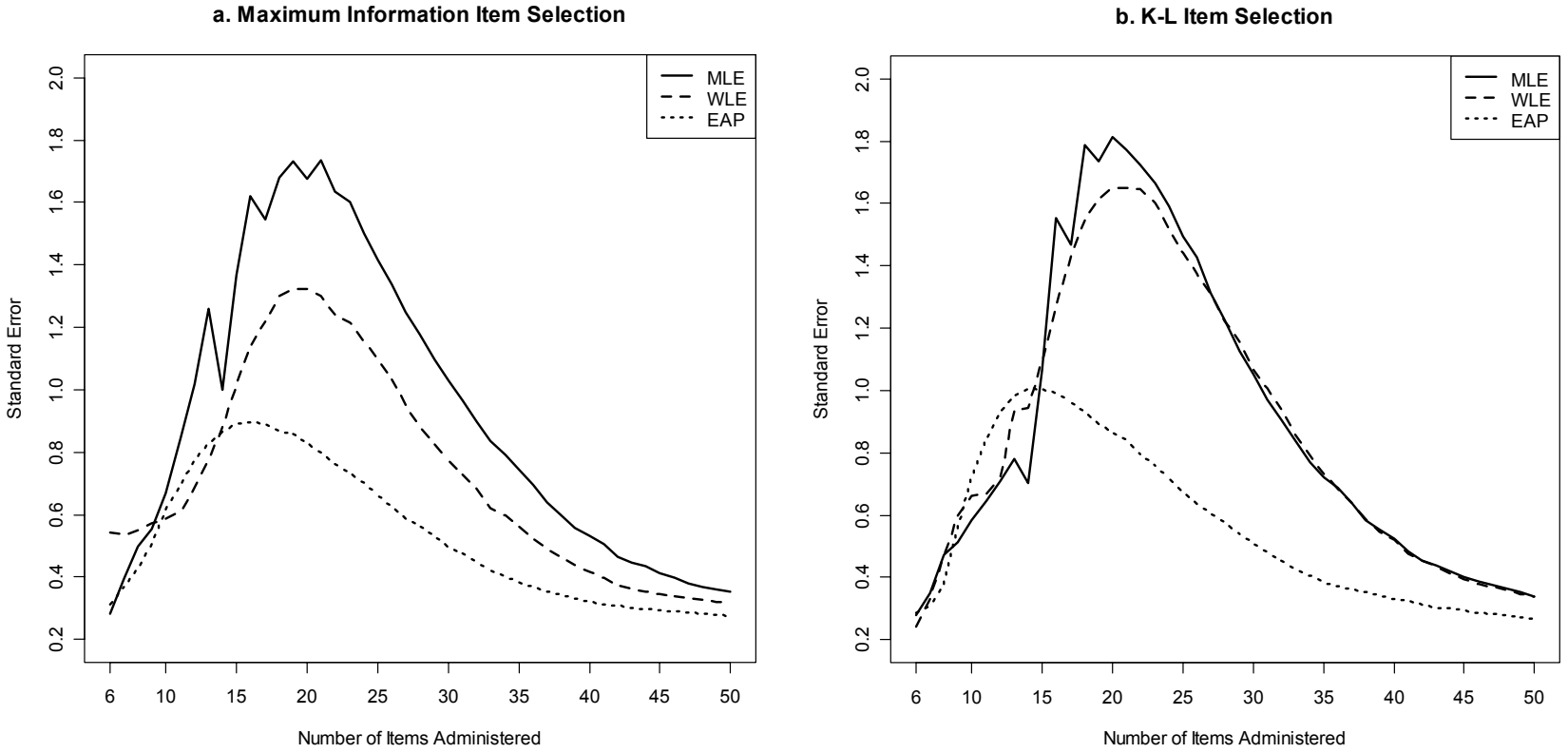
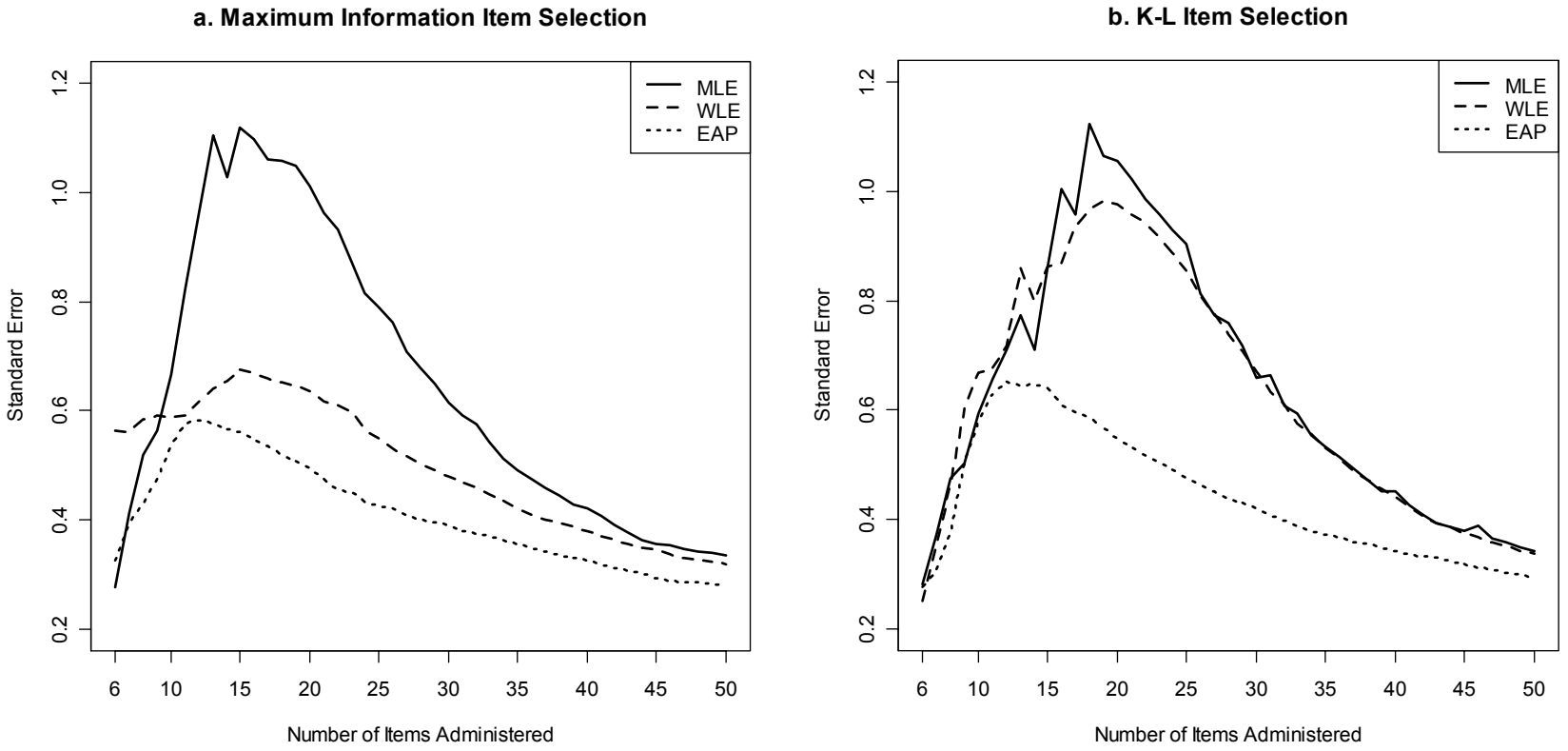


Figure 47
Empirical SE Across CAT Lengths for the 4-Item Misfit Condition for θ = –1 (MCR)



RMSE

Conditions Without Misfit

The RMSE values are shown in Tables A8−A47 for the different conditions in this study. EAP had larger RMSE values than MLE or WLE when θ was greater than 1.0 in absolute value. For conditions where $\theta = -1$, 0, or 1, the RMSE values for EAP were lower than MLE or WLE.

The RMSE values for $\theta = \pm 1$ and ± 3 were examined for test lengths of 6 to 50 items. As seen in Figures 49 and 50, the RMSEs for EAP when $\theta = \pm 1$ were lower than WLE or MLE. It was found that the RMSEs were lower for the $\theta = \pm 1$ conditions than the $\theta = \pm 3$ conditions. This discrepancy can be seen by comparing Figures 50 ($\theta = -1$) and 51 ($\theta = -3$). EAP was most sensitive to θ condition as seen in Figures 48–51.

As seen in Figures 48 and 51, EAP had the largest RMSE values of the three methods when $\theta = \pm 3$. The RMSE values for $\theta = -3$ were greater than they were for $\theta = 3$. The difference between $\theta = -3$ and $\theta = 3$ dissipated after 25 items were administered, as seen in Tables A12 and A21. After 6 items were administered in the CAT, it was observed that K-L selection resulted in larger RMSE values than FI selection. The difference between the item selection methods became less pronounced as test length increased, as seen in Tables A8−A47.

The results for WLE and MLE were dependent on item selection method and θ. For a θ of 3 or 1, WLE had larger RMSE values than MLE when FI was used, while MLE was greater than WLE when K-L selection was used. These discrepancies were quite small compared to the differences between MLE or WLE and EAP.

Figure 48
RMSE Across CAT Lengths for the 0-Item Misfit Condition for θ = 3

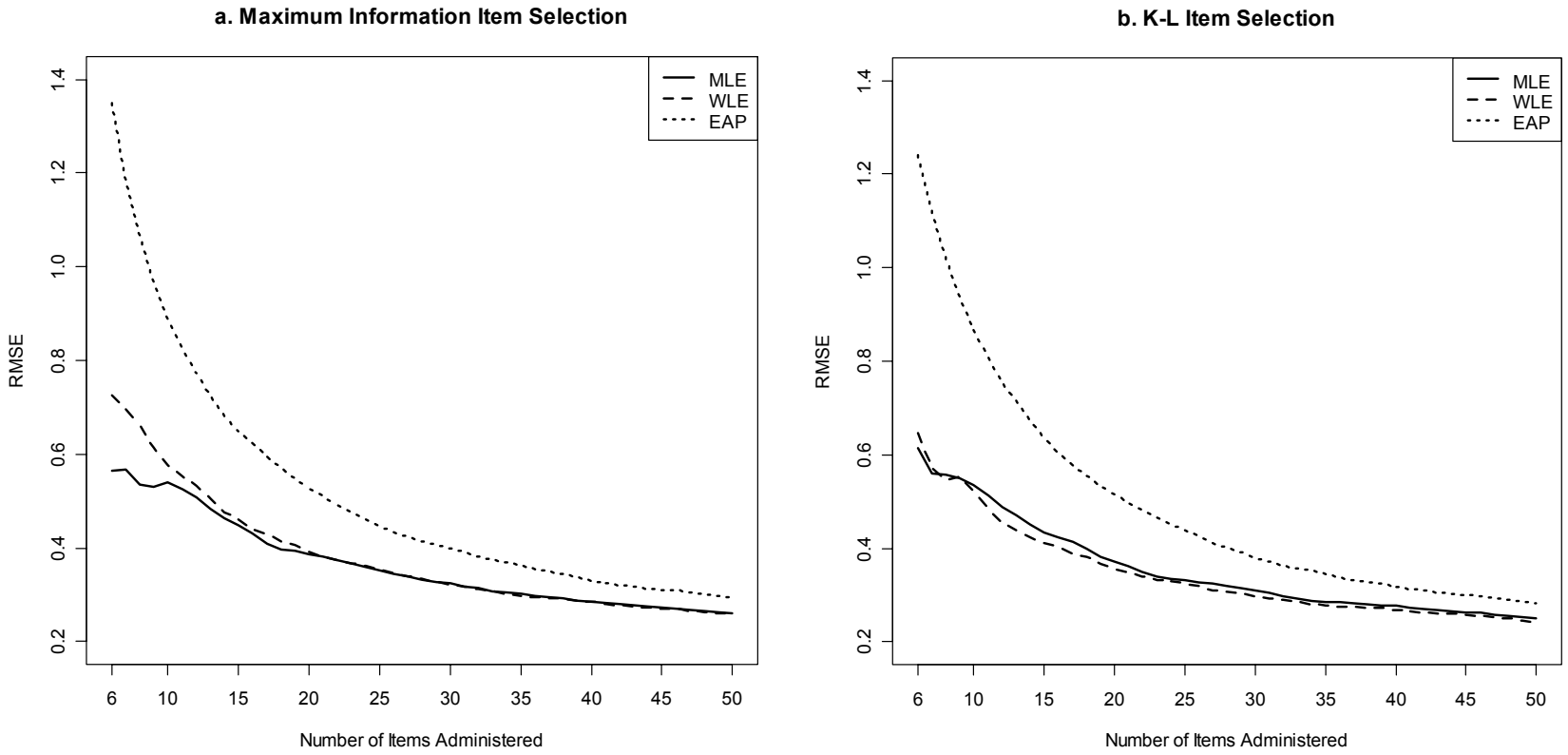


Figure 49
RMSE Across CAT Lengths for the 0-Item Misfit Condition for θ = 1

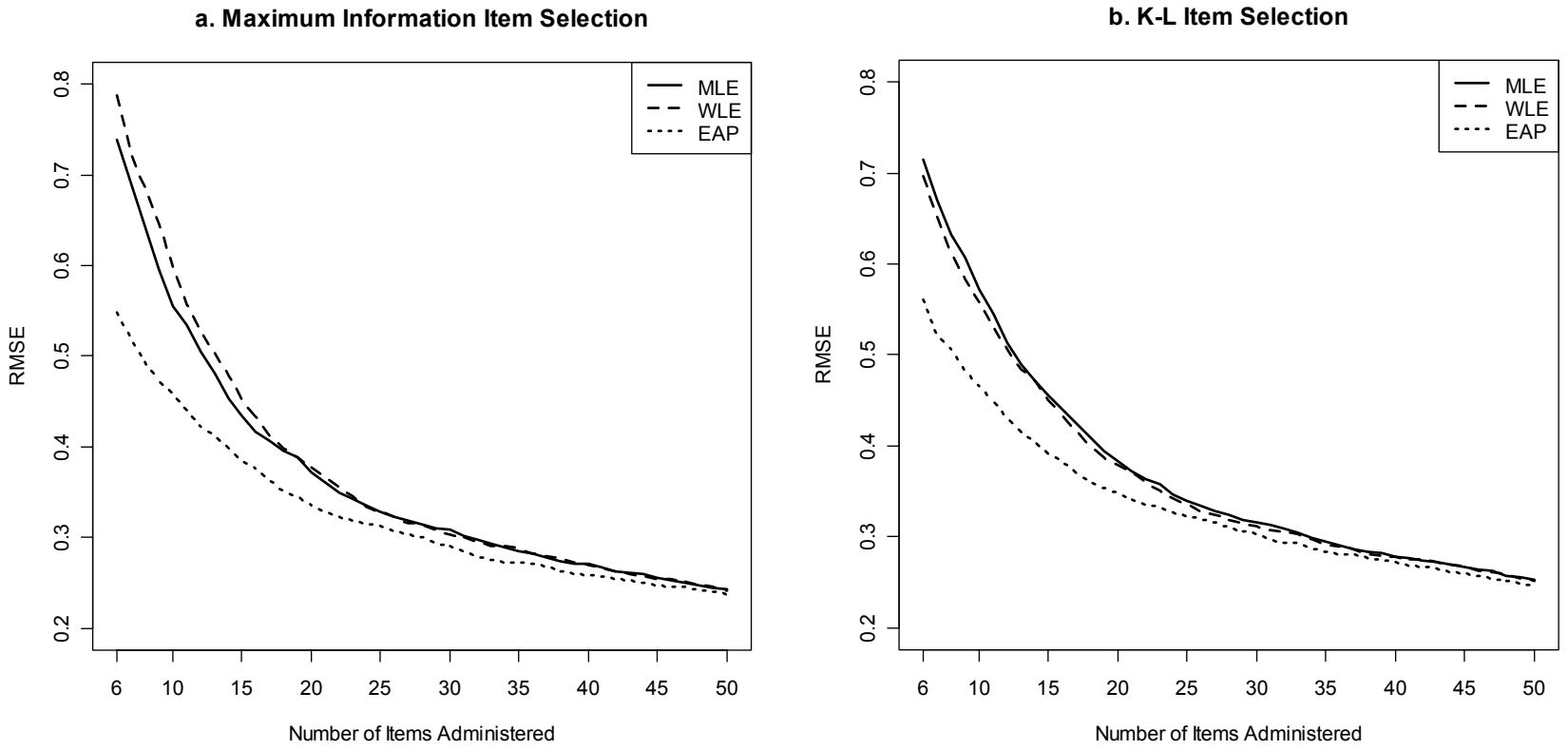


Figure 50
RMSE Across CAT Lengths for the 0-Item Misfit Condition for $\theta = -1$

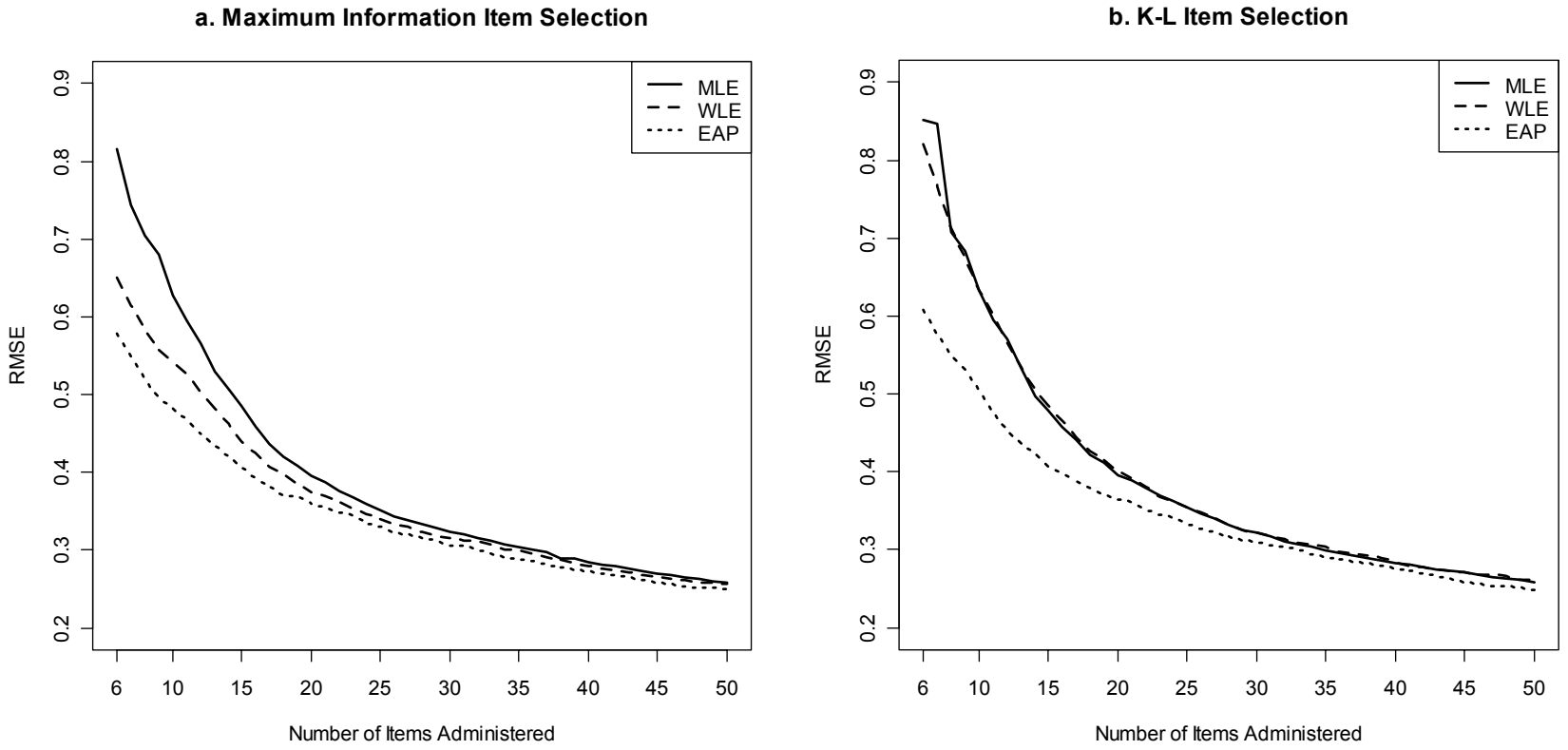
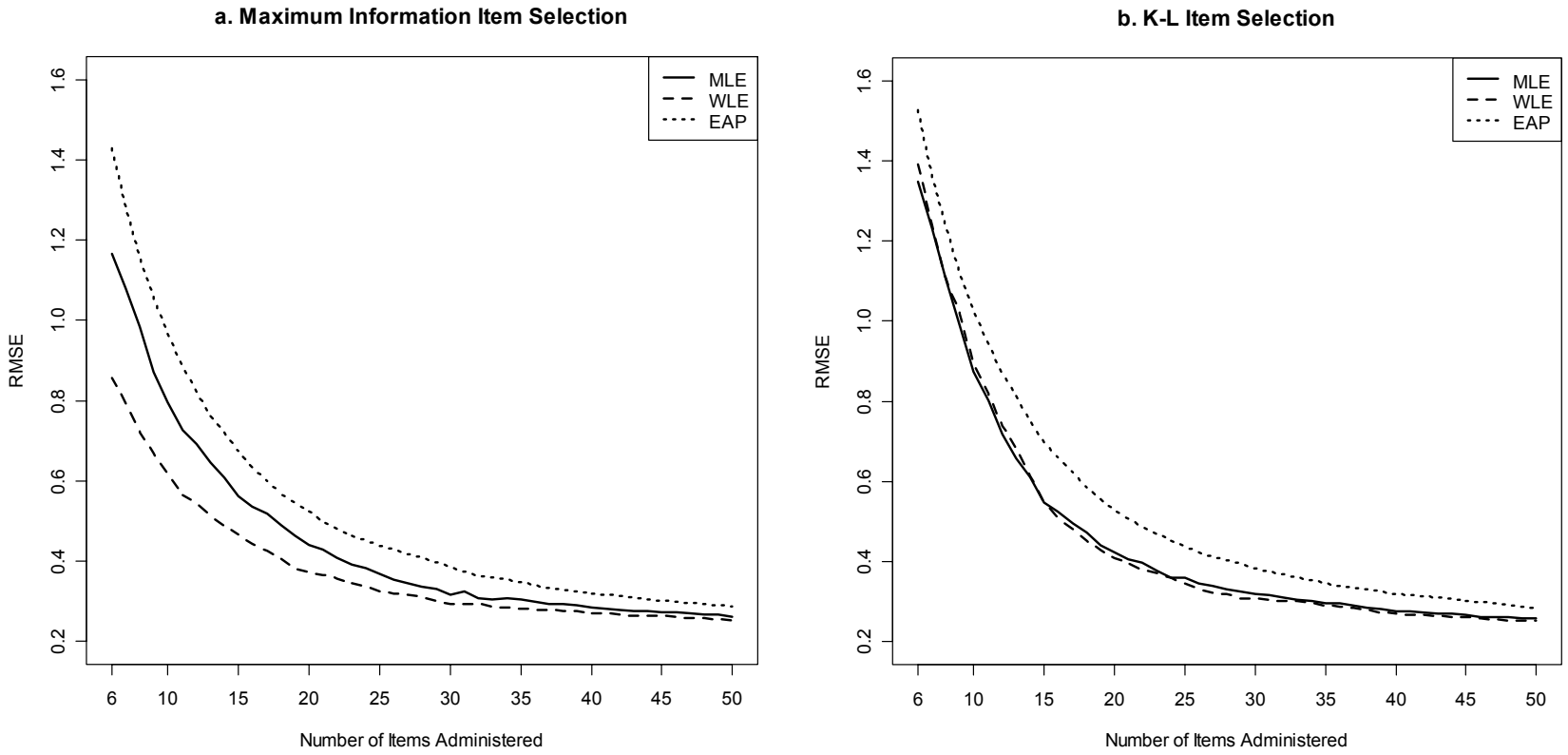


Figure 51
RMSE Across CAT Lengths for the 0-Item Misfit Condition for $\theta = -3$



MIR

As shown Tables A8−A47, the RMSE values increased as the number of misfitting item increased from 1 to 4. In addition, the RMSE values increased as θ increased from 1 to 3. The RMSE values for $\theta = 1$ and 3 for the 1 to 4 misfitting conditions are displayed in Figures 52−59. As shown in Figures 52−59, the RMSE values for WLE were larger than MLE when FI was used to select items. MLE had larger RMSE values than WLE for K-L selection when 2−4 items of misfit were introduced. As observed for the bias and SE, WLE was sensitive to item selection method, performing better when used with K-L information as compared to FI.

1 misfitting item. There was a relatively large difference in the RMSEs early in the CAT across item selection methods, as shown by Figures 52 and 53. MLE and WLE had larger RMSE values when FI was used to select items. It took about 20 items for the RMSE values to be similar across item selection methods when $\theta = 3$. When K-L selection was used for $\theta = 3$ the RMSEs were consistently as follows: EAP > WLE > MLE. For $\theta = 3$ and FI selection, after six items were administered, the RMSEs were ordered as follows: WLE > MLE > EAP. After 17 items the RMSEs were ordered as follows for FI selection: EAP > WLE > MLE. When $\theta = 1$ it can be seen in Figure 53 that EAP estimation had the lowest RMSEs of the three methods independent of item selection method or test length.

2 misfitting items. It was found that WLE was sensitive to item selection method for $\theta =$ 1 or 3. It can be seen in Figures 54 and 55 that WLE had lower RMSEs than MLE when K-L was used to select items, but larger RMSEs when FI was used to select items. As shown by Figure 54 ($\theta = 3$), EAP estimation provided the lowest RMSEs until about 30 items were administered. For the $\theta = 1$ condition, EAP provided the lowest RMSEs for

the entire CAT regardless of item selection method. When FI selection was used to select items, WLE had the largest RMSEs of the three estimation methods.

3 and 4 misfitting items. In general, EAP estimation provided θ estimates with the lowest RMSEs, as shown by Figures 56–59. WLE was sensitive to item selection method, as it had larger RMSEs than MLE when FI was used, but smaller RMSEs when K-L selection was used.

Figure 52
RMSE Across CAT Lengths for the 1-Item Misfit Condition for θ = 3 (MIR)

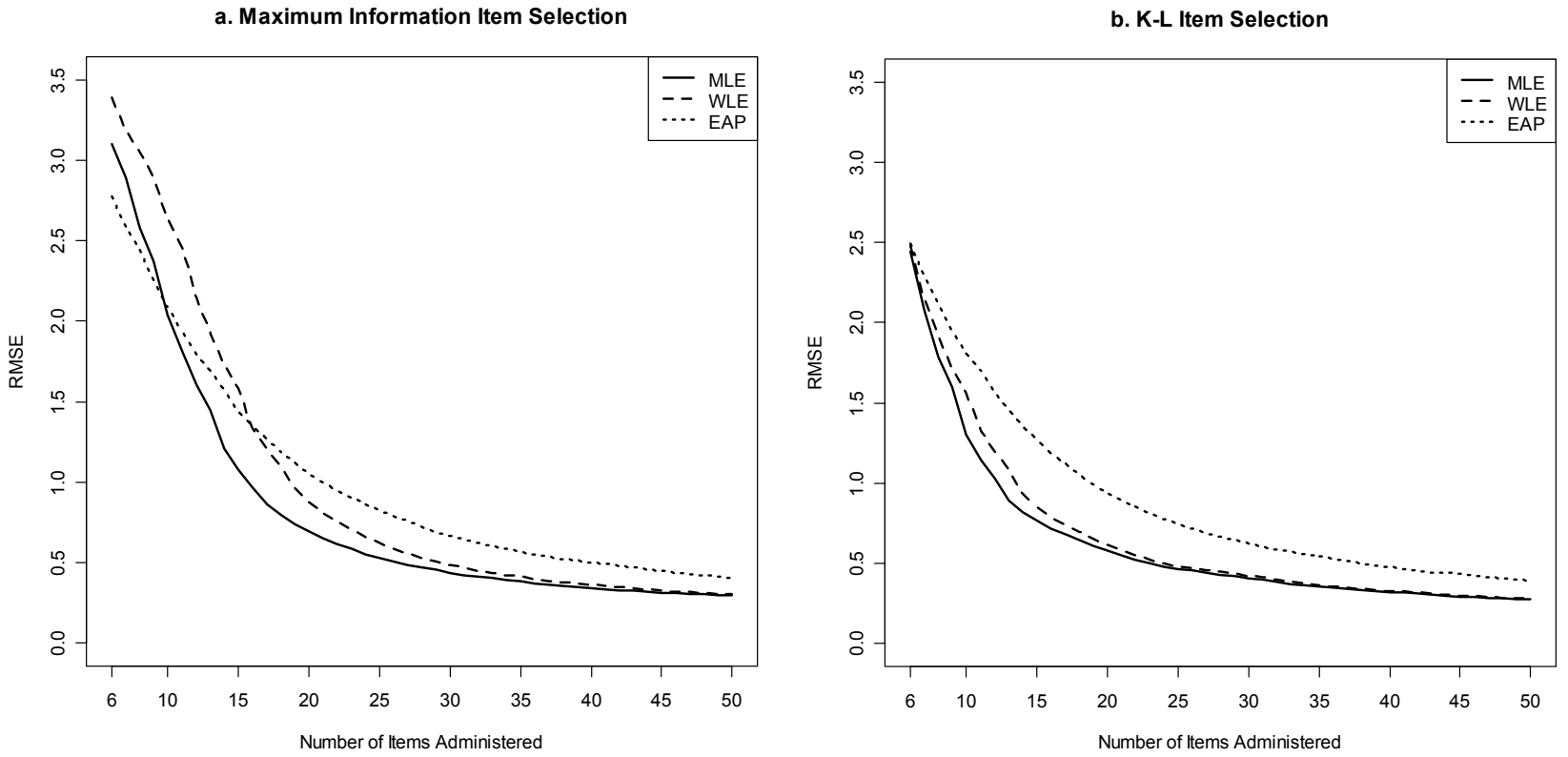


Figure 53
RMSE Across CAT Lengths for the 1-Item Misfit Condition for θ = 1 (MIR)

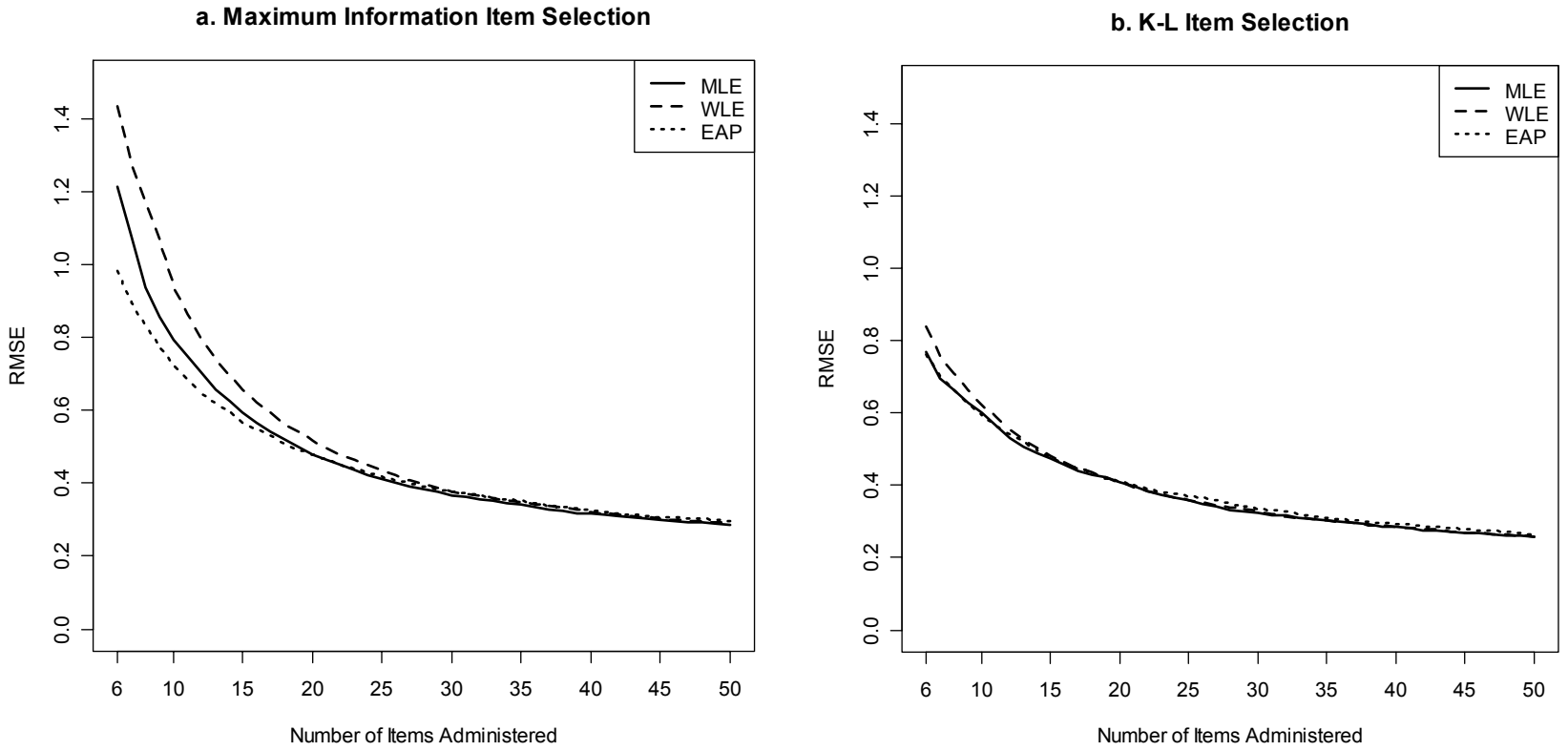


Figure 54
RMSE Across CAT Lengths for the 2-Item Misfit Condition for θ = 3 (MIR)

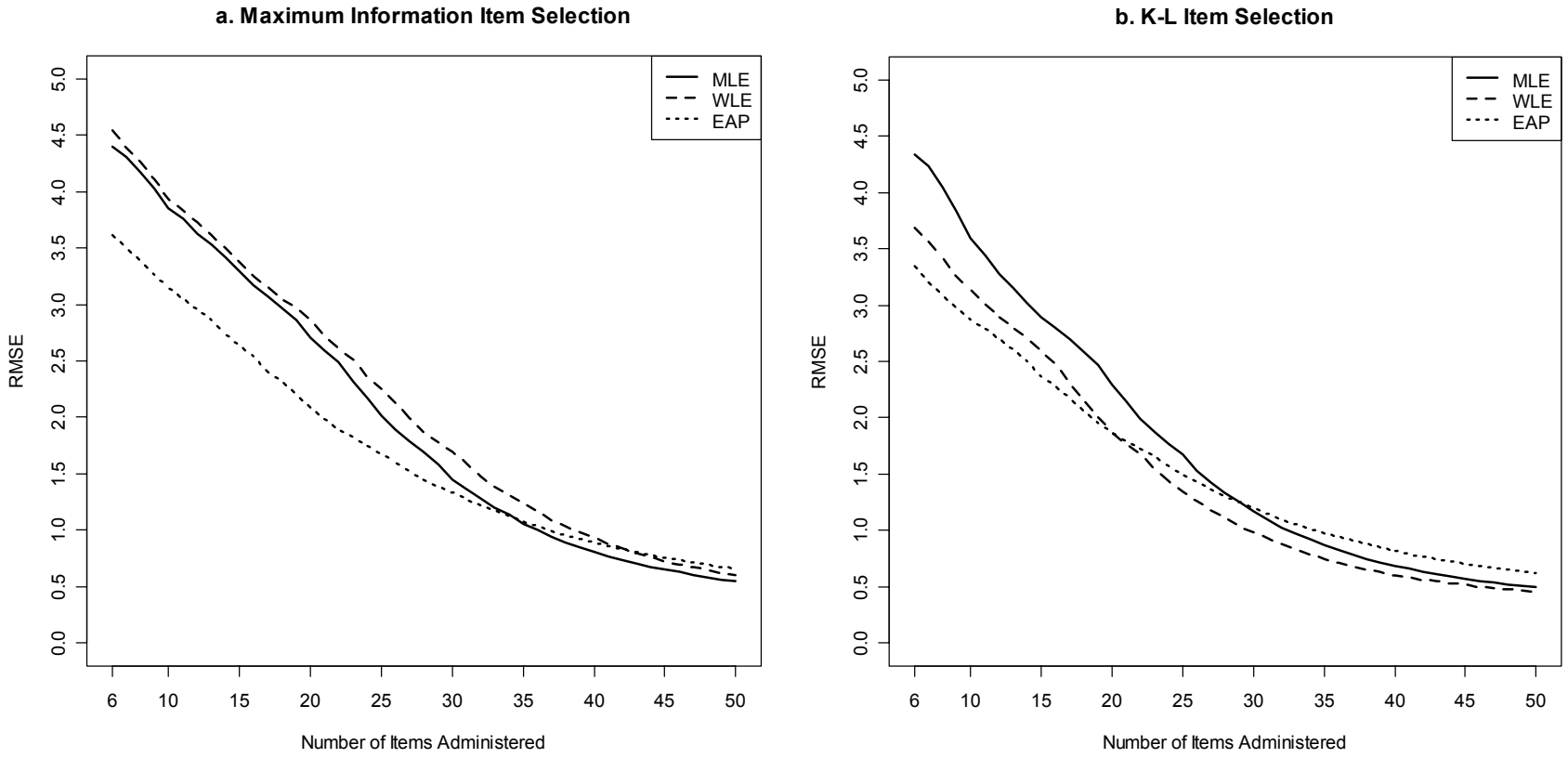


Figure 55
RMSE Across CAT Lengths for the 2-Item Misfit Condition for θ = 1 (MIR)

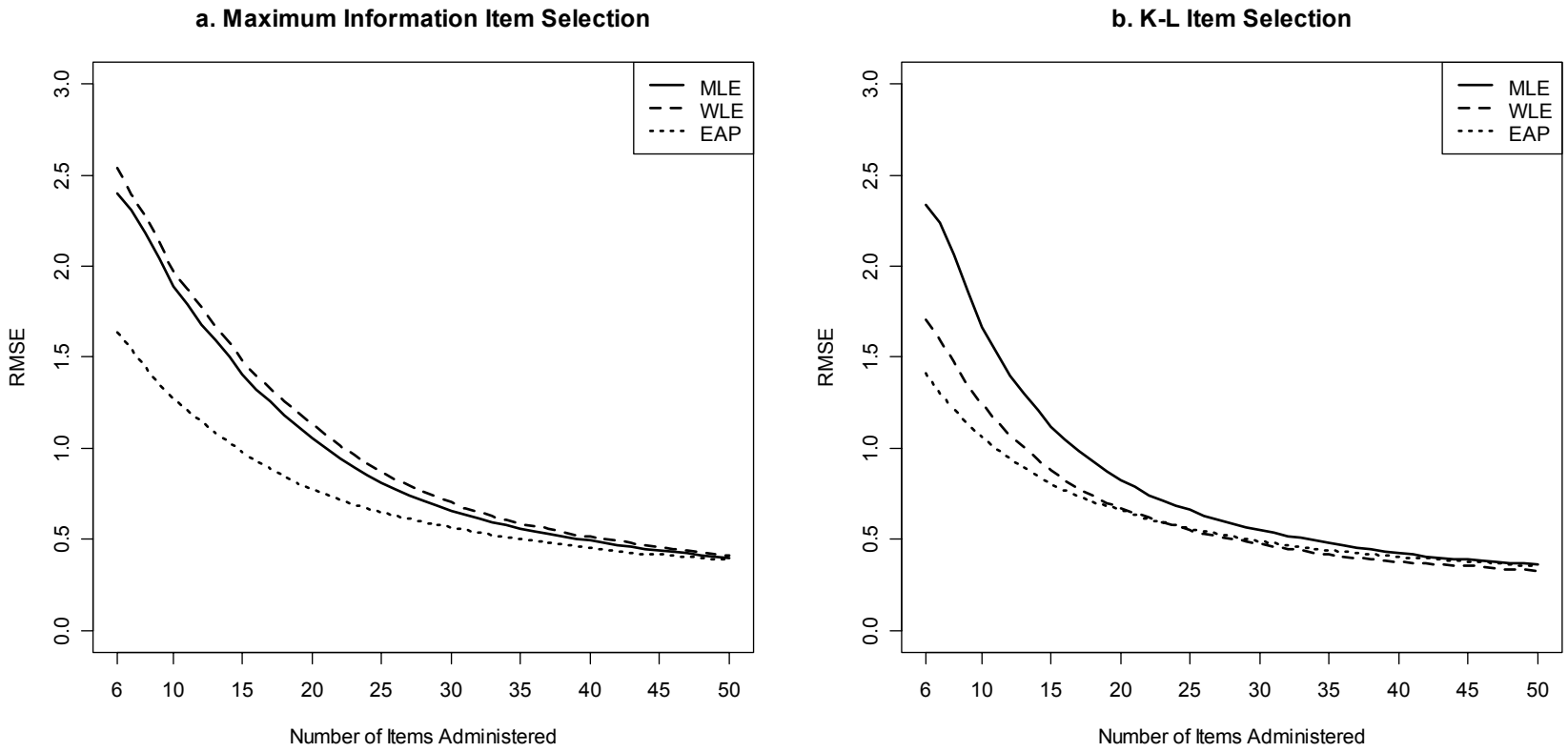


Figure 56
RMSE Across CAT Lengths for the 3-Item Misfit Condition for θ = 3 (MIR)

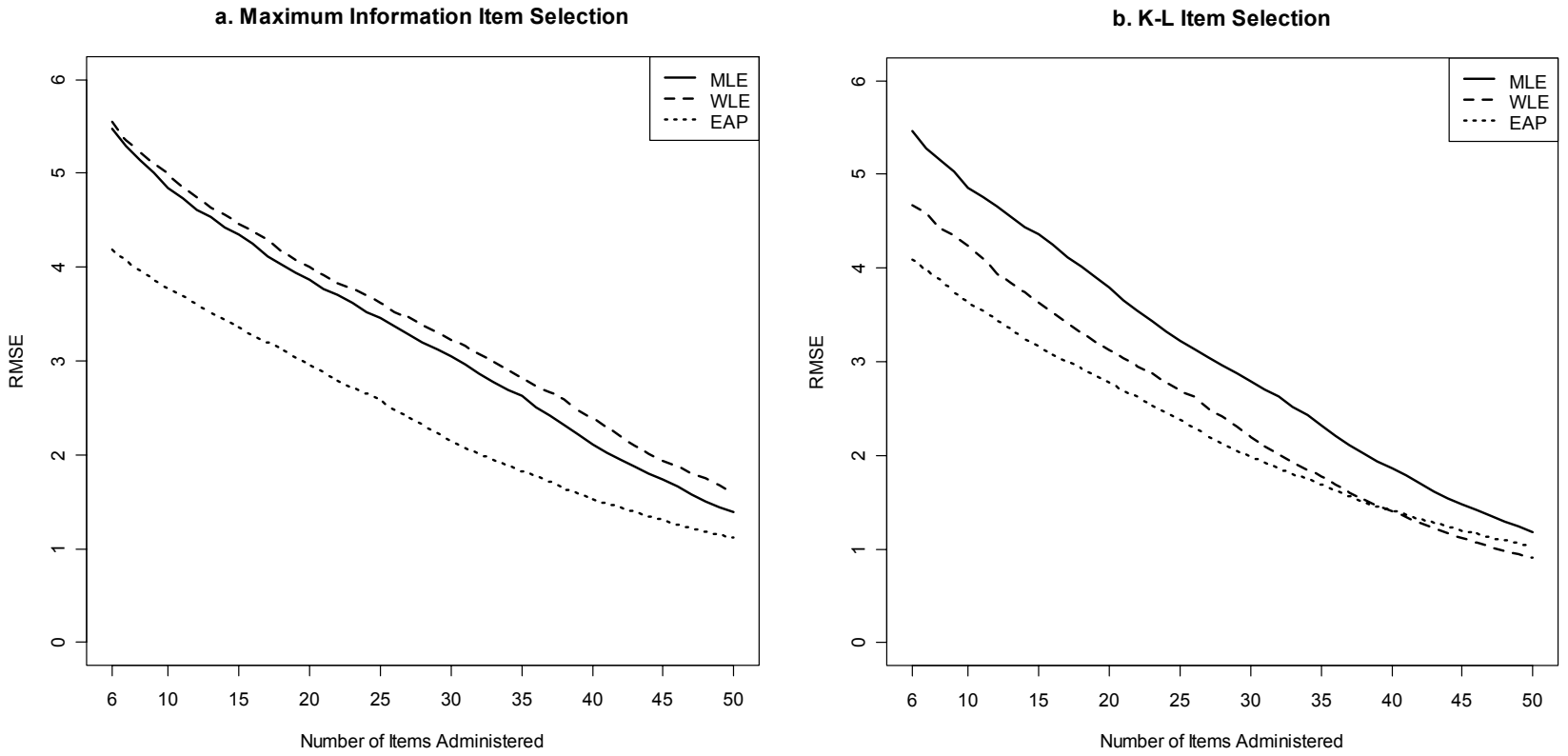


Figure 57
RMSE Across CAT Lengths for the 3-Item Misfit Condition for θ = 1 (MIR)

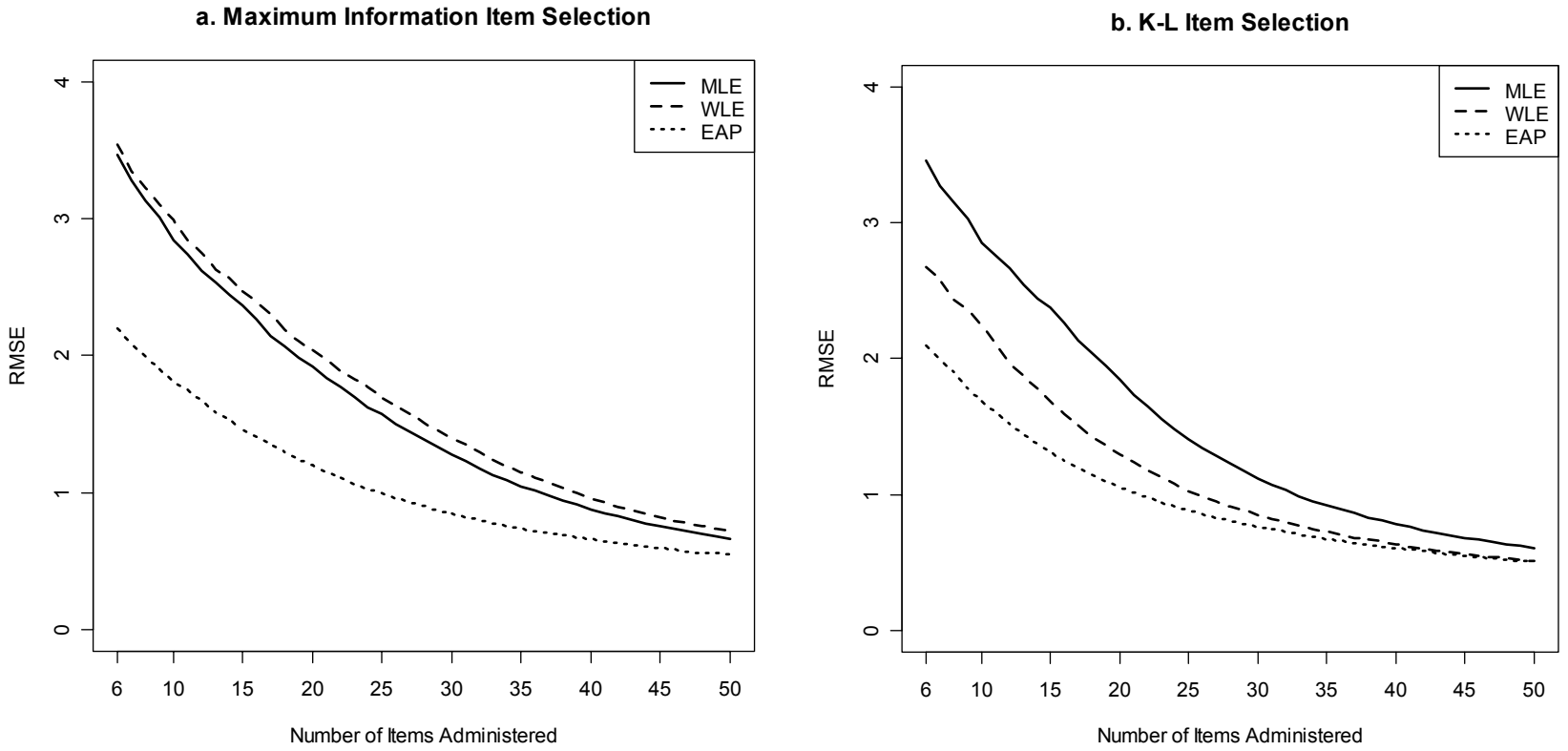


Figure 58
RMSE Across CAT Lengths for the 4-Item Misfit Condition for θ = 3 (MIR)

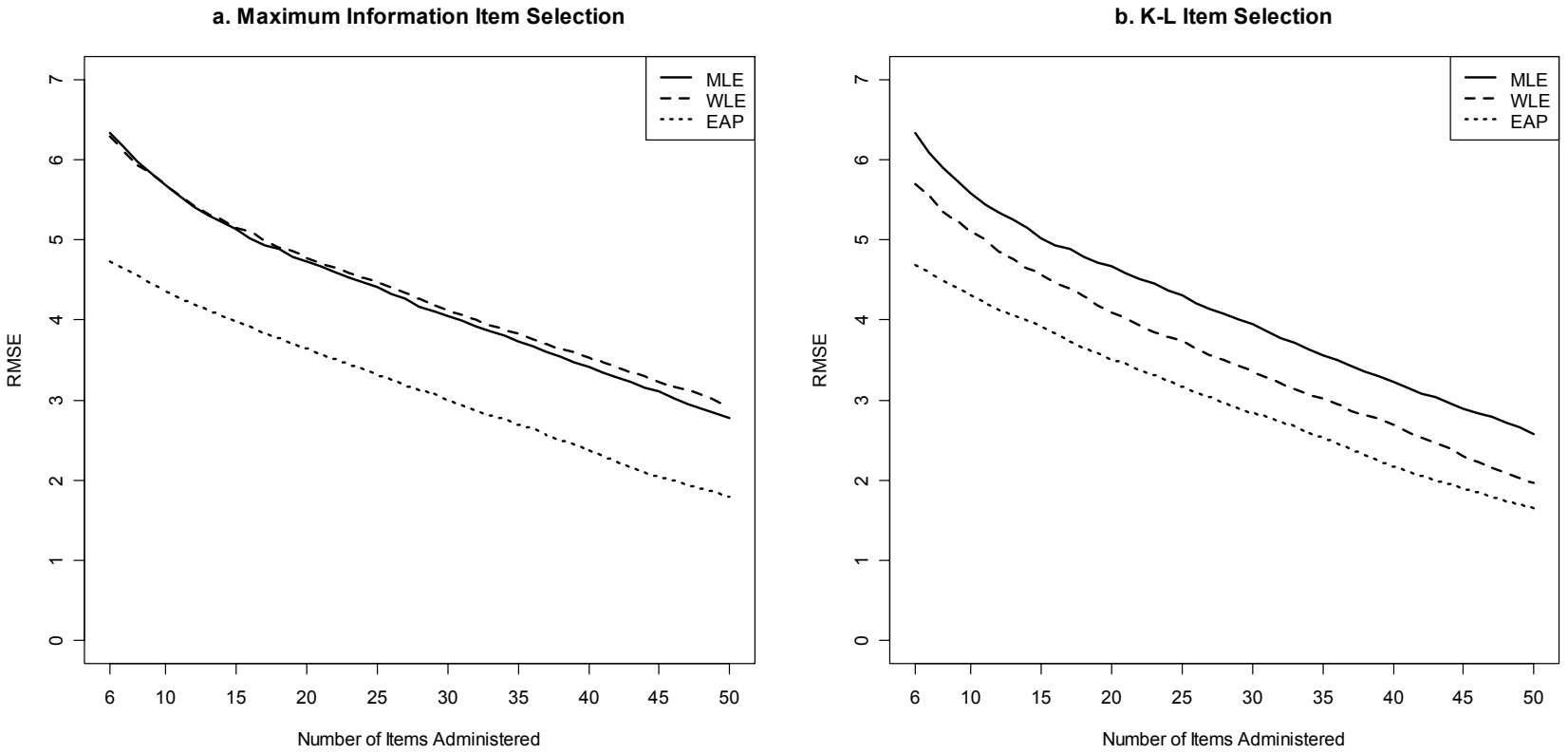
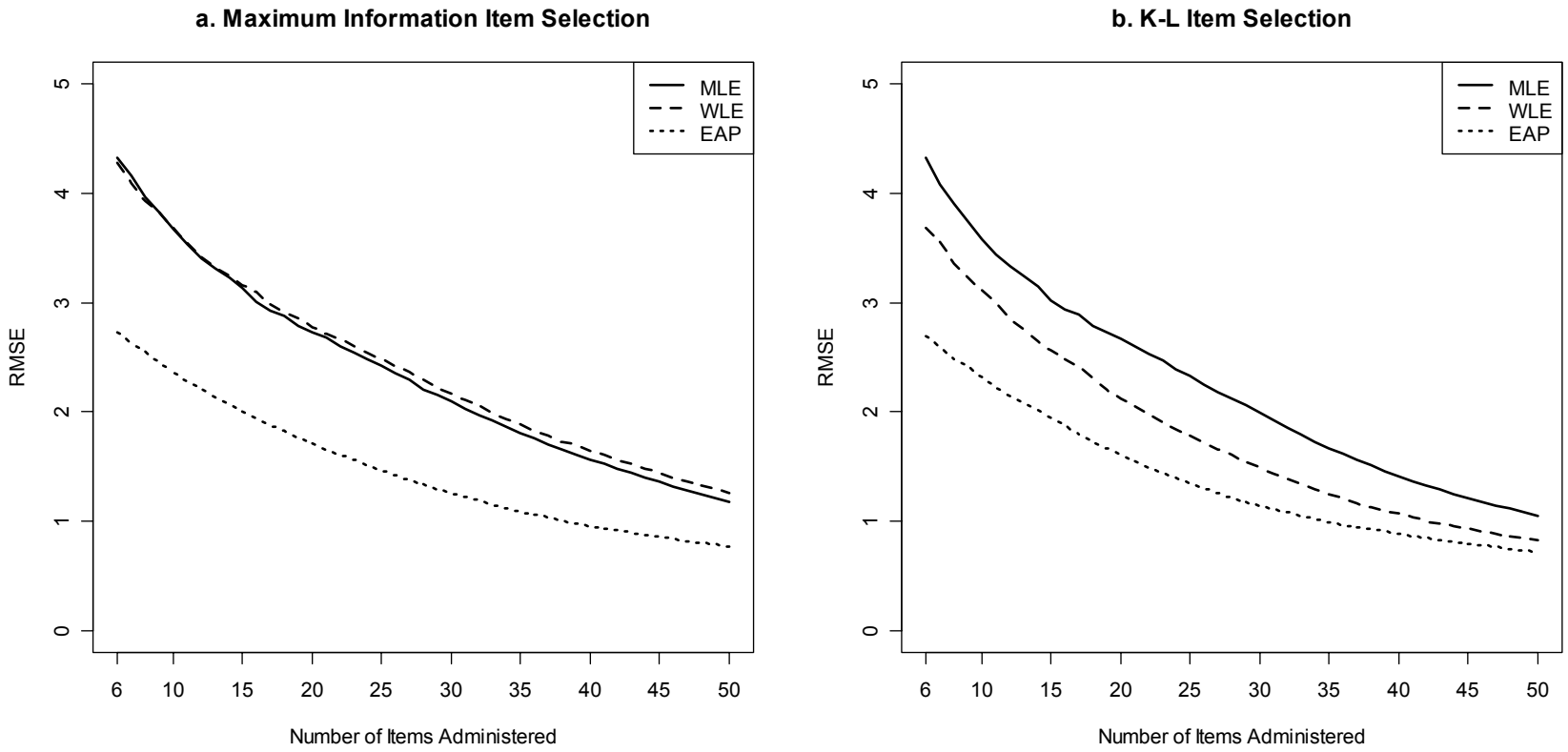


Figure 59
RMSE Across CAT Lengths for the 4-Item Misfit Condition for θ = 1 (MIR)



MCR

Effect of item selection on WLE and MLE. As shown by Tables A8−A47 and Figures 60−67, the RMSEs increased as the number of misfitting items increased. It was found that the RMSEs for $\theta = -3$ were higher than for $\theta = -1$. MLE and EAP were not as sensitive to item selection as was WLE. WLE provided lower RMSEs than MLE when FI was used to select items, but RMSEs similar to MLE when K-L selection was used. For the one- and two-item of misfit conditions, WLE had lower RMSEs than EAP or MLE when FI selection was used. It was evident in Figures 60−67 that as the number of misfitting items increased, the longer it took for the RMSEs to decrease below 1.0. For $\theta = 1$, EAP provided the lowest RMSEs of the three θ estimation methods.

Effect of misfit on EAP for $\theta = -3$. Figures 60, 62, 64, and 66 show that the RMSEs for EAP compared to MLE and WLE changed as number of misfitting items increased from 1 to 4. When there was 1 misfitting item, EAP had slightly lower RMSEs than MLE until about 20 items were administered in the CAT. For the 2-item misfit condition, EAP provided lower RMSEs than MLE until 35 items were administered in the CAT. EAP estimation resulted in the lowest RMSEs throughout the CAT for the three and four item of misfit conditions, as shown by Figures 36 and 37. This result was observed for both FI and K-L.

Figure 60

RMSE Across CAT Lengths for the 1-Item Misfit Condition for θ = −3 (MCR)

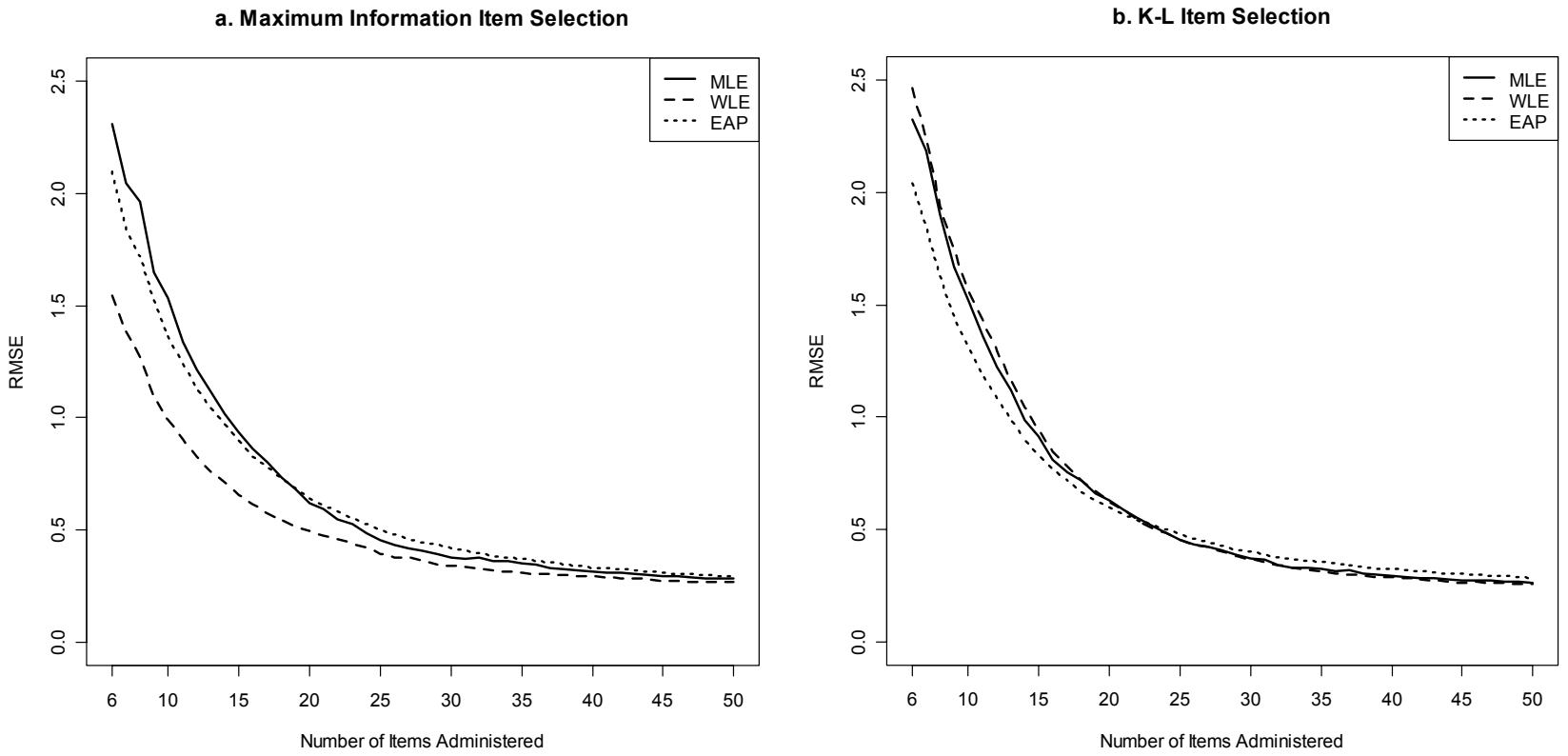


Figure 61
RMSE Across CAT Lengths for the 1-Item Misfit Condition for θ = −1 (MCR)

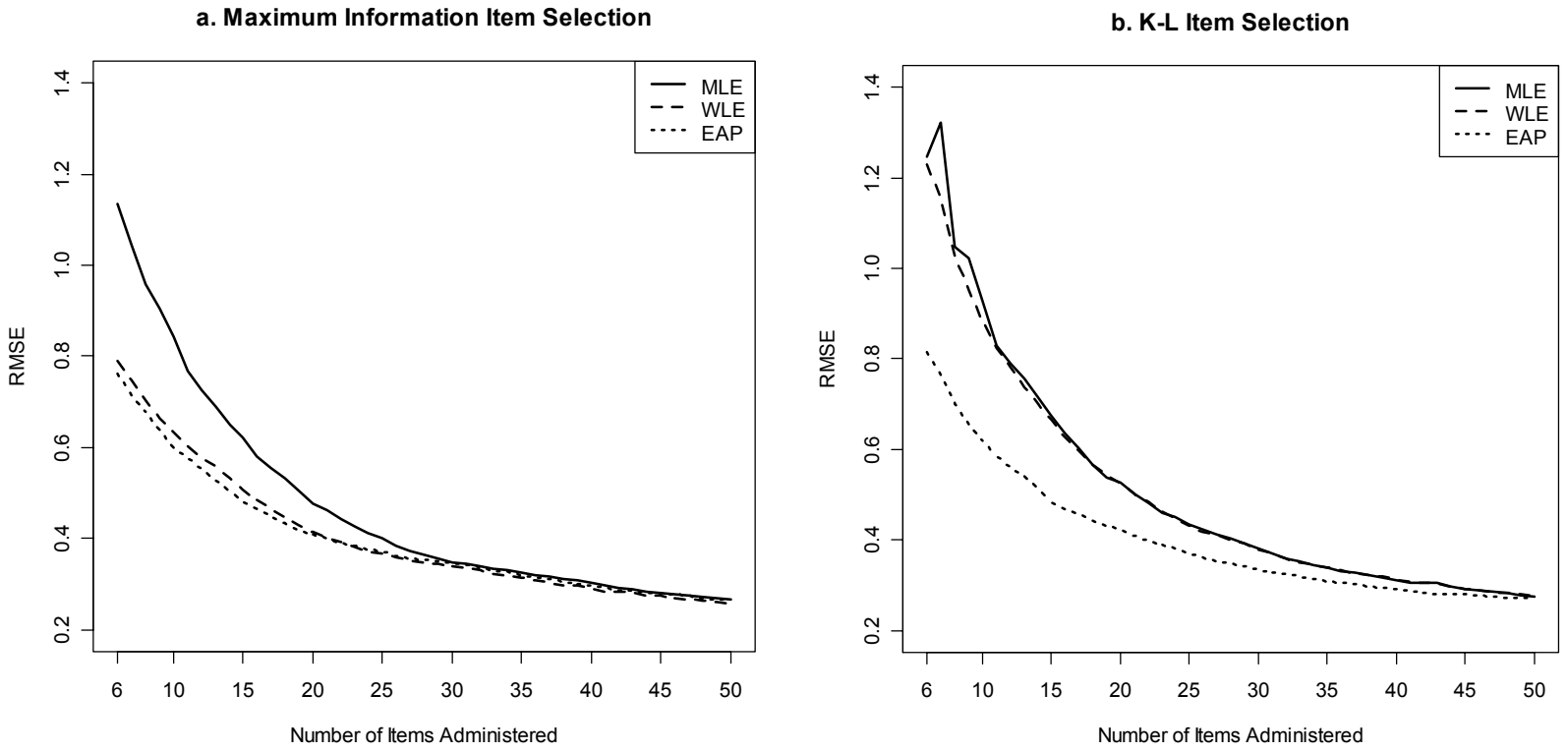


Figure 62

RMSE Across CAT Lengths for the 2-Item Misfit Condition for θ = −3 (MCR)

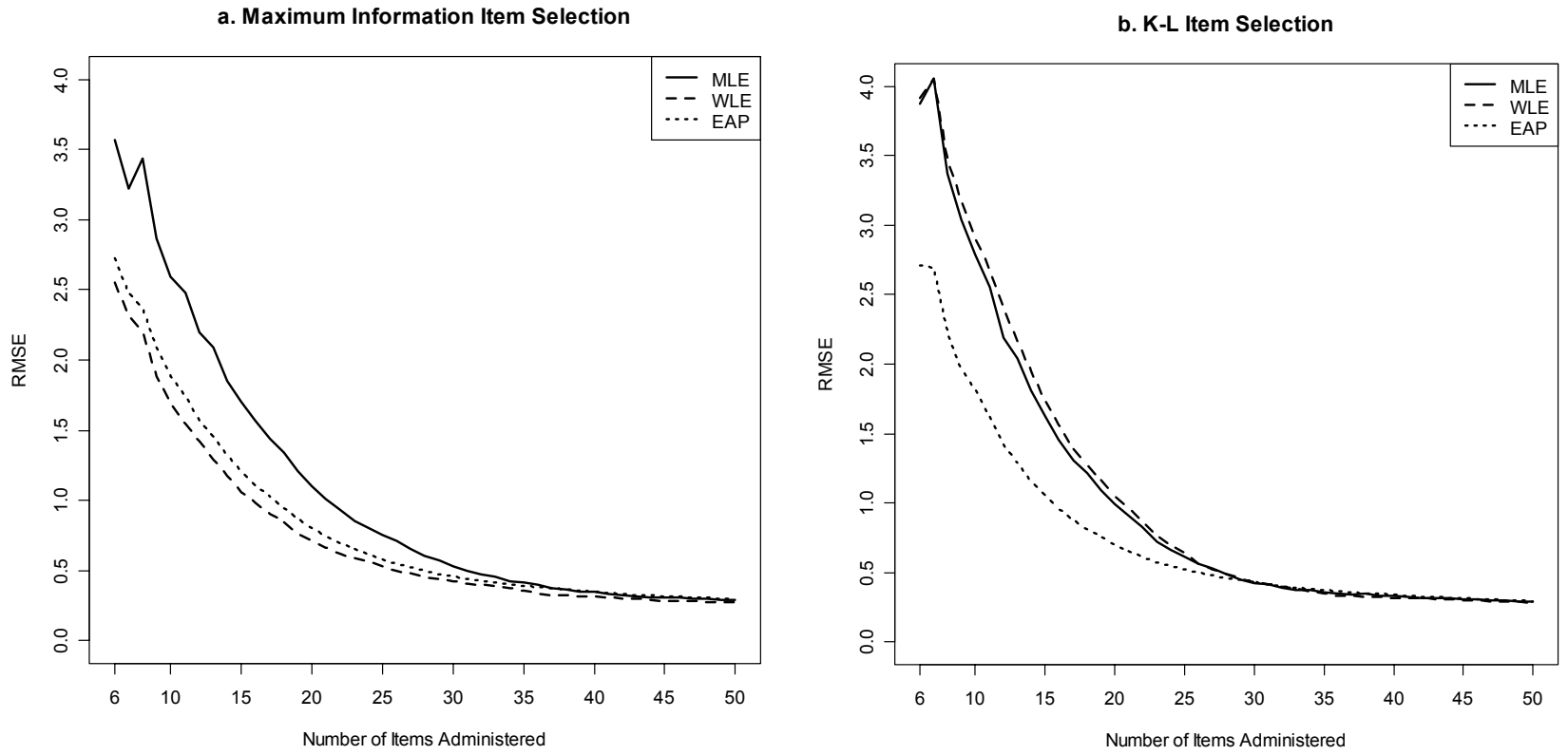


Figure 63
RMSE Across CAT Lengths for the 2-Item Misfit Condition for θ = −1 (MCR)

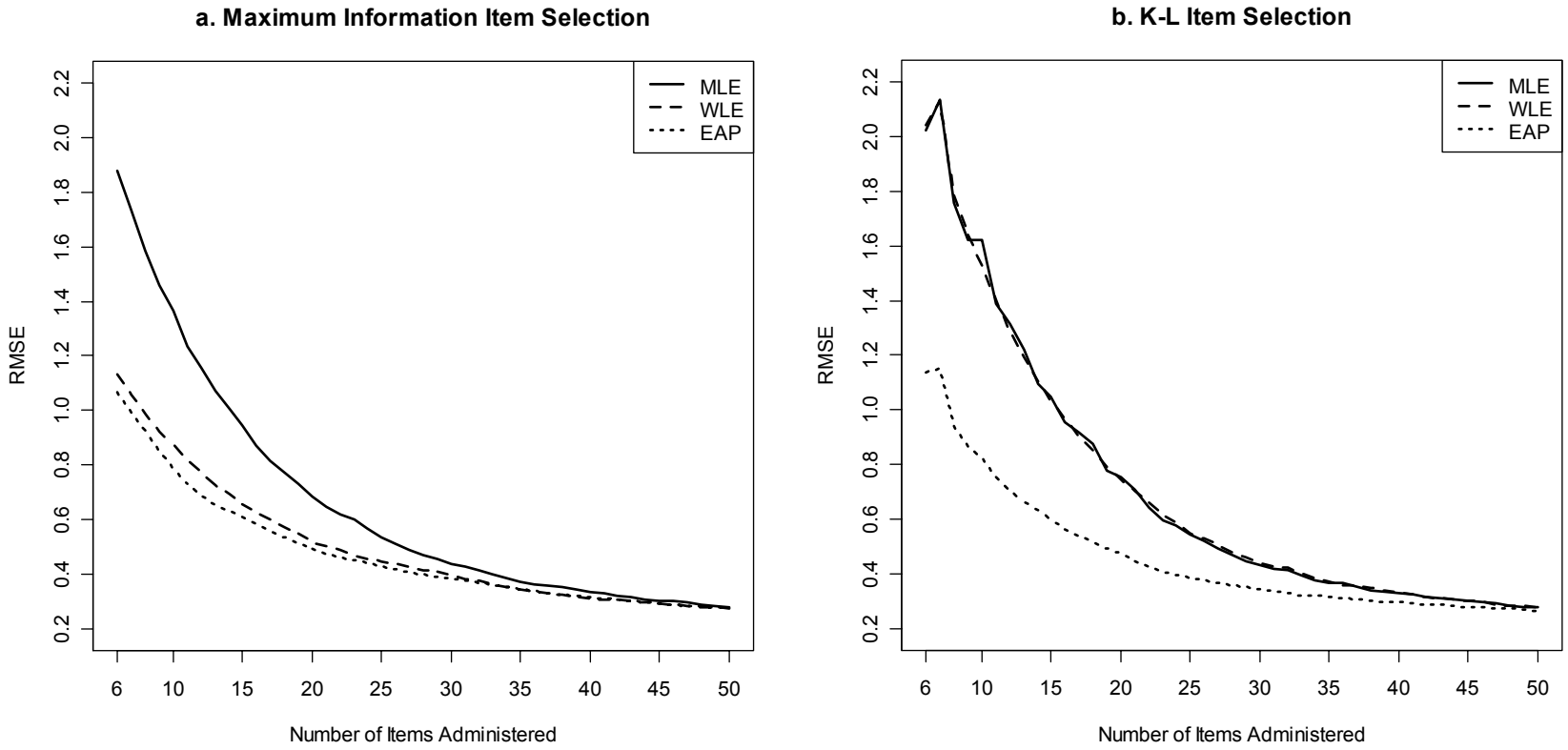


Figure 64

RMSE Across CAT Lengths for the 3-Item Misfit Condition for θ = −3 (MCR)

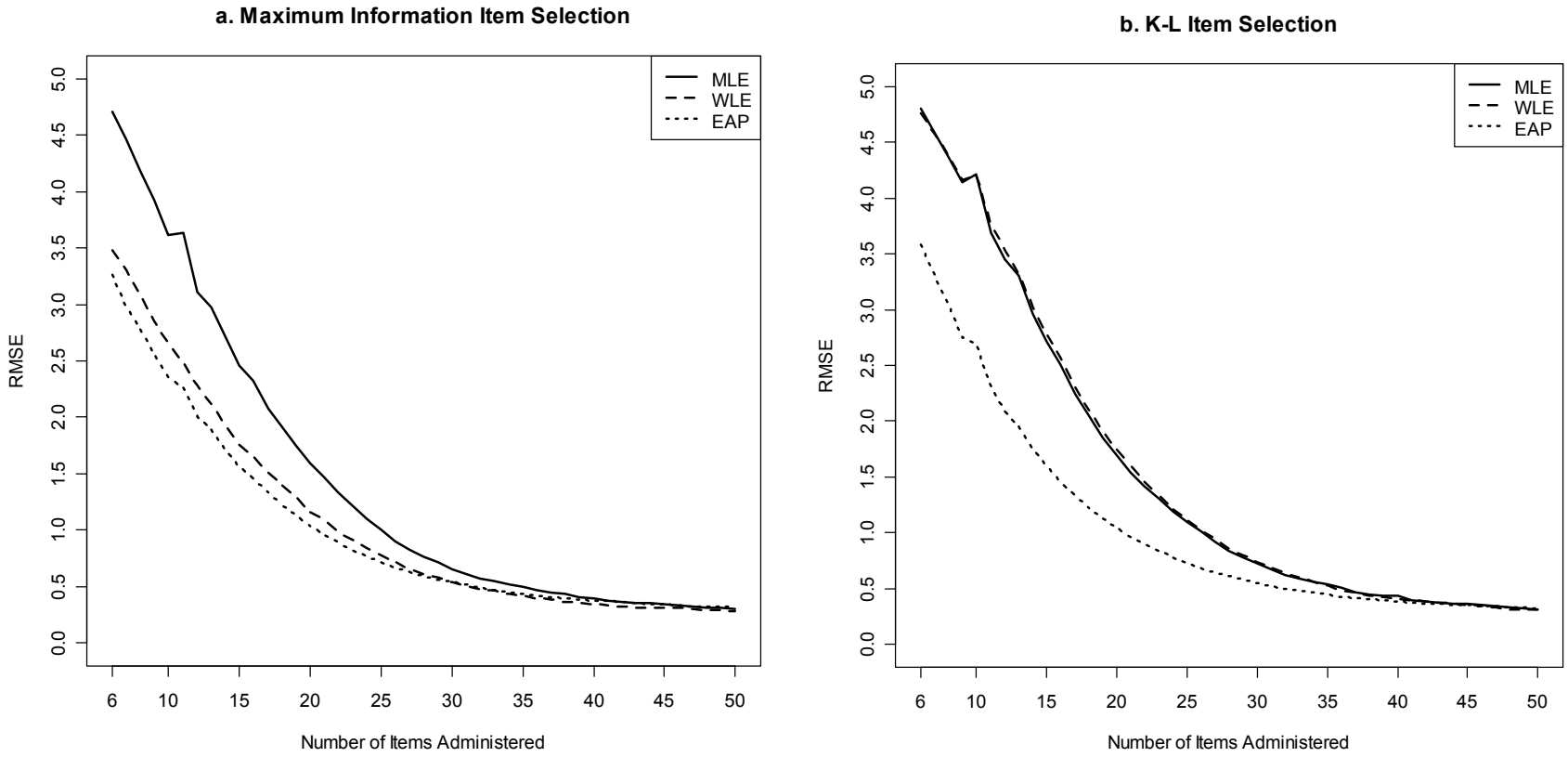


Figure 65
RMSE Across CAT Lengths for the 3-Item Misfit Condition for θ = −1 (MCR)

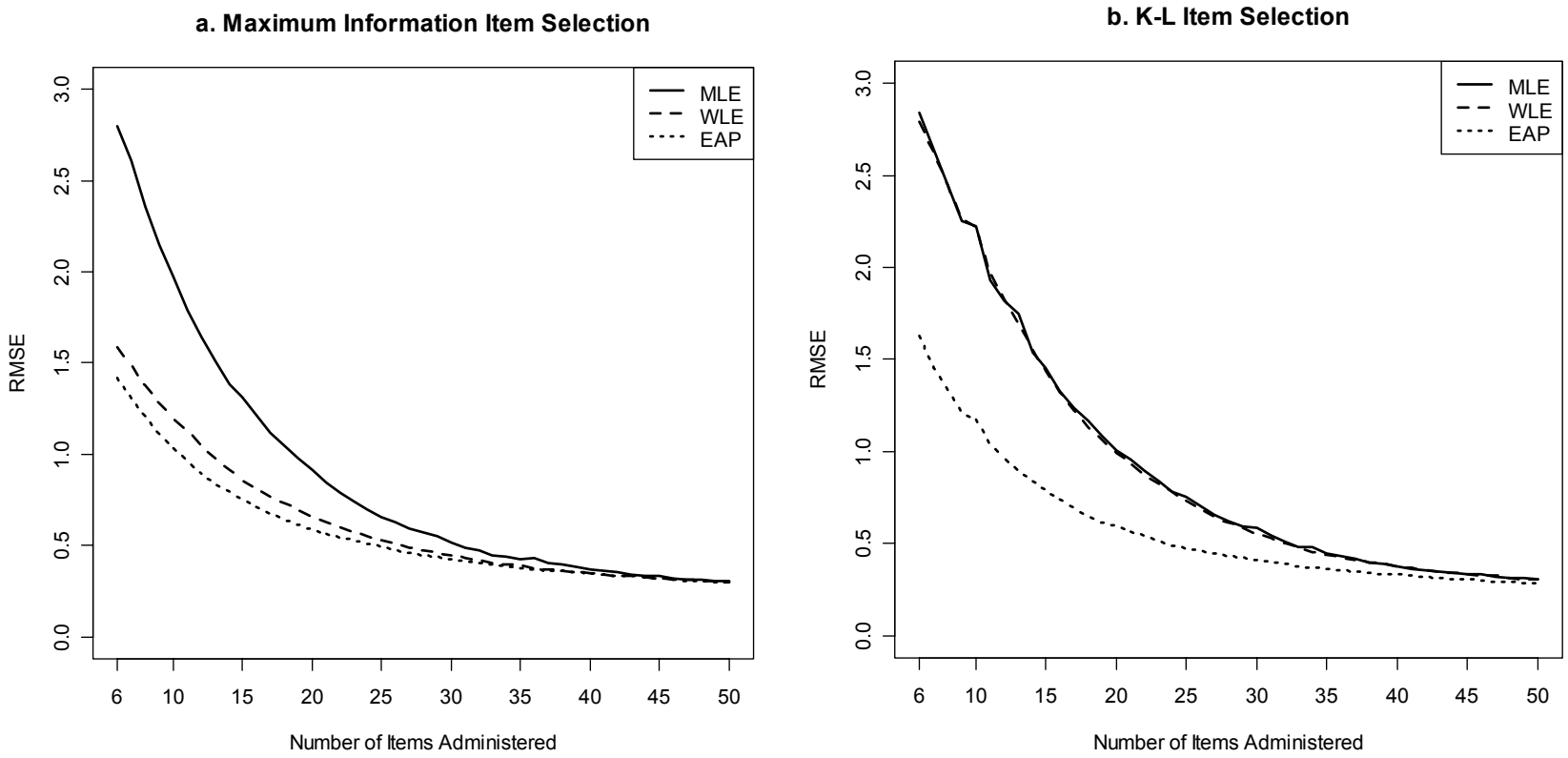


Figure 66
RMSE Across CAT Lengths for the 4-Item Misfit Condition for θ = −3 (MCR)

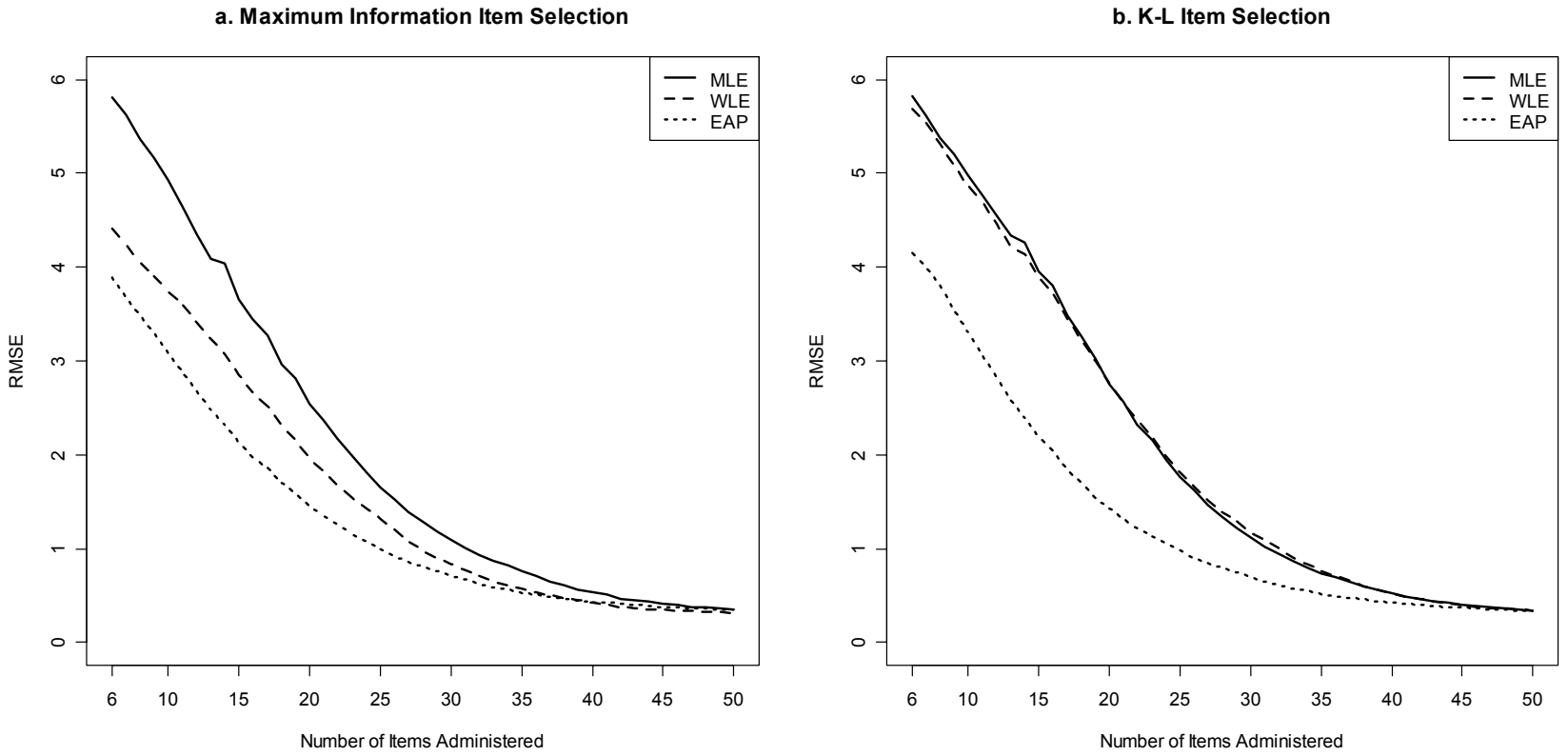
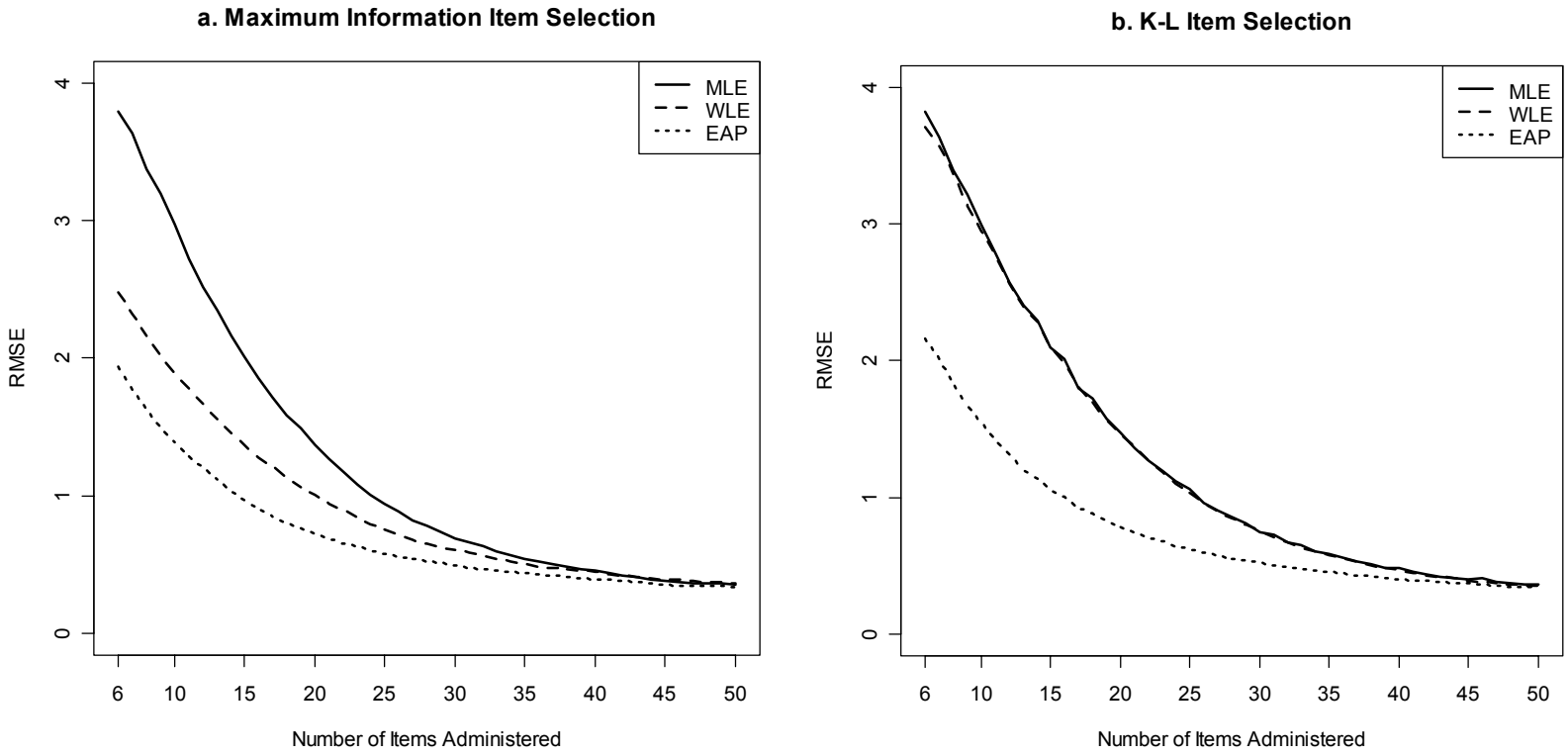


Figure 67
RMSE Across CAT Lengths for the 4-Item Misfit Condition for θ = −1 (MCR)



Chapter 4:

DISCUSSION AND CONCLUSIONS

Convergence Failures

A sizeable number of MLE θ estimates failed to converge for the MCR conditions. Convergence failures resulted from a likelihood function without a maximum, and this occurred when the observed proportion correct was less than the lower asymptote of the test response function. When MCR was introduced, simulees of low ability received items with difficulties greater than θ. This resulted in a low probability of a correct response for the simulee, and typically resulted in a string of incorrect responses. A string of incorrect responses can result in a convergence failure, as shown by Figure A1.

It was observed that K-L selection resulted in more convergence failures than FI selection. As seen in Table 6, K-L selection selected a more difficult initial item in the CAT. This would contribute to a higher θ estimate, thereby increasing the chances of a convergence failure, due to increased probabilities of an incorrect response for the items selected after misfit was introduced.

General Trends From the ANOVA

Generating θ

It was observed that misfit had a more pronounced effect on the bias values as θ was changed from 0 to 3. This result suggested that it was easier for $\hat{\theta}$ to recover for less extreme θ values. The average bias increased as θ decreased from 0 to –3 for a 15-item CAT. However, for a 35-item or 50-item CAT, the effect of θ on the average bias values was negligible, due to recovery of $\hat{\theta}$ to near zero bias for those test lengths. Thus, it was reasonable to conclude that misfit resulted in greater bias for more extreme θ when the

CAT was less than 35 items long. For the MCR conditions it can be concluded that the effect of θ on the average bias depended on test length. This was reflected in the η^2 for θ, as η^2 decreased as test length increased.

Misfit × θ

Direction of Misfit

Figure 2 and Table A2 provided evidence that MIR resulted in greater bias in the θ estimates than MCR. For example, when 15 items were administered in the CAT, the average bias for $\theta = -3$ was 1.205 while the average bias was −2.528 for $\theta = 3$. The effect of misfit on the bias almost dissipated (average bias = 0.063 for $\theta = -3$) after 50 items for MCR, but did not dissipate for MIR (average bias = −0.831 for $\theta = 3$). These results suggested that CAT with the 3PL could not account for MIR as simulated here.

Direction and Degree of Misfit

It was observed that the effect of misfit differed depending on θ and the length of the CAT. When the CAT was terminated after 15 items, an increased number of misfitting item responses resulted in increased bias across the θ continuum. This result did not hold for 50 items, as the MCR conditions recovered to near zero bias, while the MIR conditions still showed increased bias as the number of misfitting item responses was increased. These results implied that an increase in the number of misfitting item responses resulted in more bias for the MIR conditions than the MCR conditions.

These results suggest that CAT based on the 3PL is sensitive to the direction of misfit. A low ability examinee with $\theta = -3$ who guesses the first item correctly (or otherwise obtains a correct answer) will receive an unbiased MLE θ estimate (on average) if they receive at least 25 items (Figure 20). However, if an examinee with $\theta = 3$ responds

incorrectly to the first item, their MLE θ estimate will remain biased even after 50 items (Figure 12).

Conditions Without Misfit

Item Selection Method

Effects for Short CATs

When θ was less than 0, it was found that K-L selection resulted in more bias than FI selection. This difference dissipated after 15 items were administered (Figures 8–11). The SEs for K-L selection were greater than FI when $\theta < 0$, while FI had larger SEs than K-L when $\theta > 0$. WLE was found to be particularly sensitive to item selection method, and the bias and SE differed substantially across item selection methods for the first 10–15 items in the CAT. The RMSEs across item selection methods followed the same pattern as the SEs. These differences in the θ estimates (particularly WLE) after 6–15 items were administered can be attributed to the difference in initial item difficulty across item selection methods. It was apparent in Tables 5 and 6 that WLE was sensitive to initial item difficulty.

Effect for Longer CATs

It was observed that the bias of the θ estimates were similar across item selection methods after at least 15 items were administered. There was also evidence that the empirical SEs became similar across item selection methods after 15 items. As $RMSE^2 = bias^2 + SE^2$ (van der Linden, 1998), it follows that the RMSE values were also similar across item selection methods. It can be seen in Appendix Tables A18−A47 that K-L selection tended to result in negligibly larger SEs and RMSEs across generating θ than FI. Thus, it can be concluded that item selection had a negligible effect on θ recovery for

CATs longer than 15 items.

θ Estimation Method

EAP provided the smallest SEs of the θ estimation methods, as expected. The RMSEs suggested a similar trend, but the RMSEs for EAP were larger than WLE or MLE when $\theta = \pm 3$. When $\theta = \pm 1$, EAP provided the smallest RMSEs of the methods, due to reduced bias in the EAP estimates as θ was closer to the prior mean.

Effects for Short CATs

The effect of estimating MLE θ using just mixed response patterns was evident in Figure 8 for $\theta = 3$, as the θ estimates for MLE did not recover to zero bias until 10 items were administered. An examinee that missed an item early in the CAT would be expected to have a θ estimate below 3 when 6 items were administered. The MLE θ estimates were largely unbiased after 15 items when $\theta = 1$. It took until 23 items were administered for MLE to be unbiased for $\theta = -3$ when there was no misfit. Thus it can be concluded that, MLE and WLE θ estimates were able to recover to near zero bias faster for $\theta = 3$ or 1 than $\theta = -3$ or -1.

MLE generally had the largest SEs of the methods, expect for conditions where $\theta = 3$ or 1 and FI selection was used, where WLE had the largest SEs. Differences in the RMSEs between WLE and MLE were observed when no more than 15 items were administered. In general, WLE had lower RMSEs than MLE for FI selection for $\theta = -3$ or -1 (Figures 50 and 51). When $\theta = 3$ or 1, WLE had larger RMSEs than MLE when FI selection was used (Figures 48 and 49).

Effects for Longer CATs

When more than 15 items were administered in the CAT and $\theta > 0$, the differences

between the bias, SE, and RMSE for MLE and WLE became negligible. The bias of WLE and MLE became similar after 15 items were administered when $\theta = -3$ (Figure 11). When θ was less than 0, it was found that the differences between the SEs and RMSEs for MLE and WLE remained until about 30 items were administered. These results can be largely attributed to the increased SEs found for WLE when $\theta = -3$ or -1 (Figures 30 and 31). It was evident that the effect of initial item difficulty was greater in terms of the SE for WLE when θ was less than 0 than when θ was greater than 0. This result may be attributable to the increased FI (seen in Figure 1) for θ greater than 0, as increased FI would have reduced the SEs more rapidly across estimation methods.

θ

It was found that the bias of the MLE or WLE θ estimates remained near zero after 50 items across the θ continuum. However, due to a reduction in the BIF at the extremes, the SEs were higher for $\theta = \pm 3$ (Figures 28 and 31) than ± 1 (Figures 29 and 30). The increased SEs also contributed to the larger RMSEs observed for more extreme θ values.

Effect of Misfit on the Recovery of *θ*

Misfit and the SE

Number of Misfitting Responses

The empirical SEs of the θ estimates were affected by the direction of misfit. It was observed by comparison of Figures 32–39 to Figures 40–47, that increasing the number of misfitting item responses increased the SEs for the MCR conditions, but decreased the SEs for the MIR conditions.

MIR. The reduction in the SEs as the number of misfitting responses increased can be attributed to two factors. First, as a result of MIR, high ability examinees encountered

items with b parameters much lower than their generating θ. Given this, the probability of a correct response would be near 1.0 for a simulee with $\theta = 3$ that responded to an item with $b = -3.0$. This resulted in a succession of correct responses, and reduced the variation in $\hat{\theta}$. Second, the θ estimates did not recover to zero bias for any of the MIR conditions even after 50 items. As a result, the generating θ tended to be greater than a given item's difficulty, which meant that the probability of a correct response was greater than .50 + (c / 2). These high probabilities resulted in reduced variation in the θ estimates.

MCR. The c parameter ensured that there was variation in $\hat{\theta}$ for the MCR conditions, as a simulee with $\theta = -3$ would have a probability somewhat greater than c of responding correctly to an item with a b parameter of 3.0. Whether or not a correct response was obtained on the difficult items immediately following the misfitting responses had a large effect on the empirical SEs for the subsequent θ estimates. The variability of the θ estimates was increased due to this random guessing for at least the first 20 items in the CAT.

θ

MIR. For the MIR conditions it was observed that the SEs decreased as θ changed from 1 to 3 (Figures 32–39) . A simulee with $\theta = 3$ would have a higher probability of a correct response to a given item than would a simulee with $\theta = 1$. This resulted in a greater incorrect-response rate for simulees with $\theta = 1$, thereby introducing additional variation in $\hat{\theta}$.

MCR. The results showed that the empirical SEs increased as θ changed from –1 to –3 (Figures 40–47). It was evident that the recovery of θ worsened as θ changed from –1 to –3. This finding can be interpreted by consideration of the observed probabilities of a

correct response given θ. Simulees with higher θ would have a greater probability – according to the 3PL model – of a correct response to the items for which misfit was introduced. Large theoretical SEs were shown by Figure A1 to result from a flat likelihood function. Correct responses to difficult items (b = 3) and incorrect responses to easier items (b = –3) resulted in a flattening of the likelihood function for simulees with lower θ values. The lower the θ value, the more likely a simulee was to respond incorrectly to items of low (–3) difficulty, thereby flattening the likelihood function.

Item Selection Method

MIR

FI selection resulted in more biased θ estimates than K-L selection. The SEs for K-L selection were greater than the SEs for FI selection. As the RMSE is a combination of bias and SE, it followed that the RMSEs were higher for FI selection than K-L selection.

These results can be attributed, in part, to the difference in initial item difficulty. As K-L selected a more difficult initial item, it followed that the θ estimates were initially higher for K-L selection when there was MIR. This caused the bias values for K-L selection to be lower than FI selection. The selection of a more difficult initial item increased the variability in the θ estimates (due to lower probabilities of a correct response given θ) compared to FI selection. There was evidence in Figures 19, 39, and 59 that the bias, SE, and RMSE for the WLE θ estimates were most sensitive to item selection method.

MCR

There was evidence that K-L selection resulted in increased bias in the θ estimates compared to FI selection. These differences dissipated as test length increased (Figures 20−27). In addition, the empirical SEs were greater when K-L selection was used. The

increased SEs for K-L selection were most evident for the 1-item misfit condition (Figures 40 and 41). In general the SEs curves peaked due to a large number of convergence failures, then began to decrease as test length increased. These results provided evidence that MLE was especially sensitive to MCR, as the θ estimates failed to converge frequently for extreme θ (−3).

θ Estimation Method

MIR

It was found that the θ estimates differed in average bias when MIR was introduced. The performance of WLE in terms of bias, SE, and RMSE was sensitive to item selection method. When K-L selection was used, WLE resulted in less biased θ estimates with larger SEs than MLE. When there was two misfitting items, EAP estimation resulted in the least biased θ estimates until 30 items were administered. If there were 3 or 4 misfitting responses, it was found that EAP estimation provided consistently less biased θ estimates than MLE or WLE. It can be seen in Figures 12−19 that EAP provided less biased θ estimates than MLE or WLE provided that the test was short enough (generally about 10−30 items). It was also evident that increasing the number of misfitting items also increased the test length for which EAP provided less biased θ estimates than MLE or WLE.

The SEs for EAP were larger than WLE or MLE for $\theta = 3$ when 3 or 4 misfitting items were introduced using FI selection (Figures 36 and 38). This result was contradictory to theory, as the use of prior information about θ should reduce the SE. Due to regression of θ toward the prior mean, EAP θ estimates were initially greater than MLE or WLE. This resulted in lower probabilities of a correct response (given the generating θ), and resulted

in an increase in the variation in the θ estimates for EAP due to more incorrect responses.

As there was substantial bias present in the θ estimates, it was found that the RMSEs for WLE were greater than those for MLE which were, in turn, greater than those for EAP when FI selection was used. When K-L selection was used with 2 or more misfitting items, it was found that the RMSEs were ordered as follows: MLE > WLE > EAP. WLE had the largest RMSEs independent of item selection method when one misfitting item was introduced.

MCR

It was found that the bias of the θ estimates decreased to near zero when MCR was introduced into a CAT (Figures 20−27). The number of items required for MLE and WLE to recover to near zero bias for $\theta = -3$ varied from 25 to 40 items for the 1 to 4 misfitting item conditions, respectively. EAP did not recover to zero bias due to the effect of the prior. It was observed that θ remained positively biased (about 0.08 units) for WLE and MLE when $\theta = -1$ and MCR was introduced (Table A41). It was possible that the examinees did not respond incorrectly to enough items to recover θ without bias.

MCR had a substantial effect on the SEs of the θ estimates, as seen in Figures 40−47. One practical concern was the magnitude of the SEs for MLE estimation. If a simulee with $\theta = -3$ guessed correctly on the first item, then the empirical SE would be larger than 1.0 until 14 items were administered. Four misfitting responses resulted in SEs for MLE that were greater than 1.0 when 14 to 30 items were administered. The SEs for $\theta = -1$ were less extreme than for $\theta = -3$. However, the SEs for $\theta = -1$ were still larger than 0.8 with 10 to 20 items in the CAT when 2 or more misfitting items were introduced (Figures 43, 45, and 47). It can be concluded that MCR was detrimental to the accuracy

of the θ estimates in recovering θ, and researchers should be aware of the large empirical SEs found when a low ability examinee guesses correctly (or otherwise obtains a correct response) early in the CAT.

The results for the MCR conditions revealed that MLE provided the largest RMSE values regardless of the number of misfitting items. The results for WLE were sensitive to item selection method early in the CAT. WLE generally performed better when FI was used to select items, as it selected an easier initial item. When there were 3 or 4 misfitting item responses, it was found that EAP provided the best recovery in terms of the RMSEs.

Effect of the c parameter. As the 3PL was used in this study, there was a non-zero probability of a correct response to a difficult item for low ability simulees due to guessing. However, high ability simulees were expected by the model to get low difficulty items correct nearly 100% of the time. If a simulee responded incorrectly to the initial item in the CAT, it was found that the MLE θ estimates remained biased for $\theta = 3$ or 1 even after 50 items were administered. Alternatively, MLE θ estimates became unbiased after 30 items if a low ability simulee responded correctly to the initial item (Figure 20). This provided evidence that a CAT with the 3PL cannot account for MIR.

Effect of the Prior

The regression of θ toward 0 contributed to the EAP θ estimates being less biased after 15 items than WLE or MLE for both the MCR and MIR conditions. As the EAP estimates were initially less biased it was easier for them to recover. The less extreme difficulty of the first few items selected in the CAT also helped in the recovery for EAP. In the case of MIR, more difficult items were incorrectly answered when using EAP estimation than with MLE or WLE. This resulted in higher θ estimates for those simulees. Alternatively,

less difficult items were correctly answered for the MCR conditions, and resulted in lower θ estimates for those simulees.

Sensitivity of WLE to Item Selection Method

Theoretical Bias Function

As defined by Equation 10, WLE used a *WFD* in the estimation of θ. This *WFD* is a function of the first order bias function, defined by Equation 8. One important property of the bias function was that it crossed zero for a single item when $\theta = b$. As seen in Equation 10, the product of the first order bias function and the TIF was subtracted from the first derivative of the likelihood function. For a mixed response pattern, the location where the bias function crossed zero (given that the items administered were the same) would determine whether $\hat{\theta}_{\text{WLE}}$ was larger or smaller than $\hat{\theta}_{\text{MLE}}$. This location was dependent on the difficulty of the item(s) administered in the CAT.

Differences in the shape of the theoretical bias function (obtained from Equation 8) across item selection methods were observed, as shown in Figure 68. The θ estimate after one item was higher for K-L selection than it was for FI selection. This contributed to $\hat{\theta}_{\text{WLE}}$ with K-L being consistently higher than $\hat{\theta}_{\text{WLE}}$ with FI selection. It was seen in Tables 5 and 6 that K-L selected more difficult items than FI selection. As a result, the bias function after four items crossed zero at a higher θ for K-L than FI selection.

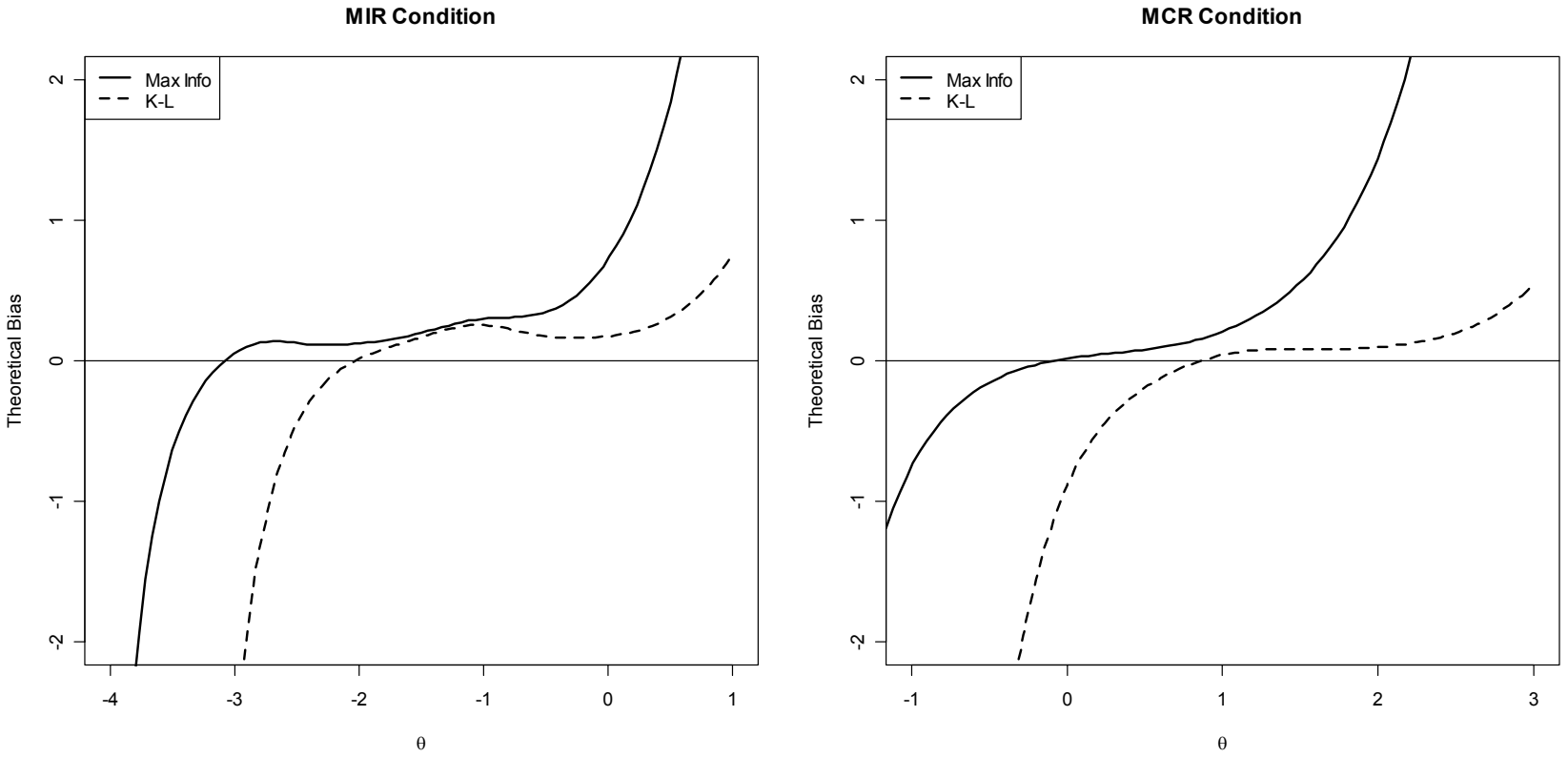
As WLE θ estimates were adjusted by a function of item difficulty, the observed differences in bias for WLE for the MIR conditions between FI and K-L can be meaningfully interpreted: Selection of a more difficult item by K-L meant that the *WFD* crossed zero at a higher θ.

Test Length

It was shown by past research that the difference between WLE and MLE became negligible after 10–15 items were administered in a CAT (Warm, 1989; Yi et al., 2001). This result can be attributed to the selection of highly informative items in CAT, as the bias function became smaller as information increased. For this study, that trend held for the MCR conditions, provided that the bias of MLE became near zero.

For the MIR conditions the discrepancy between WLE with K-L selection versus FI selection did not disappear after 50 items. As seen in Table A47, the bias for WLE differed across item selection methods by 0.15, 0.715, and 0.958 for the two, three, and four misfit conditions, respectively. These results indicated that WLE did not recover from the differences in early item difficulty that were observed across item selection methods.

Figure 68
Theoretical Bias Functions After 4 Items Were Selected in the CAT



Follow-up Study

Rationale

It was observed that WLE was sensitive to item selection method when there was misfit. This sensitivity was most pronounced for the MCR conditions early in the CAT, and was pronounced throughout the CAT for the MIR conditions. An examination of the first five items selected revealed that WLE θ estimates were also sensitive to the difficulty of the initial item selected in the CAT. This follow-up study examined how manipulation of initial θ affected MLE, WLE, and EAP. The goal was to see if the observed sensitivity of WLE to item selection method resulted from differences in the difficulty of the initial item.

Method

MIR and MCR were introduced with the same procedure used for the main study. FI item selection was used for this follow-up study. The most extreme case of misfit examined in the original study, four misfitting item responses, was used for this follow-up study. Likewise the most extreme θ values of ±3 were used in this follow-up study. Recovery of θ was assessed with the average signed bias and the SE.

Initial item difficulty. The θ used to select the initial item was fixed to either –3 or +3. All subsequent MLE θ estimates were fixed to either –4 (all incorrect responses) or 4 (all correct responses) when the response pattern was not mixed. This ensured that just the initial item's difficulty was manipulated in this study. Initial θ used in item selection was crossed with θ (±3) and θ estimation method to result in a 2 × 2 × 3 design.

Results

The average bias was calculated for θ estimates that converged after 6 to 50 items were

administered. As shown by Figures 69 and 70, the bias for each of the three θ estimation methods was lower when initial θ equaled θ. WLE was the least biased θ estimation method when initial θ equaled θ, but WLE was the most biased method when initial θ differed from θ by 6 units.

MIR. As shown by Table 7, WLE was most sensitive to initial item difficulty. When 25 items were administered, the difference in $\hat{\theta}$ for WLE across initial difficulty conditions was 3.751. This discrepancy was apparent in Figure 69, as WLE was heavily biased when initial θ equaled –3, but less biased when initial θ was 3.

MCR. WLE remained sensitive to initial item difficulty for the MCR conditions. When 15 items were administered, $\hat{\theta}_{\text{WLE}}$ differed across initial item difficulty by 3.163 units. The discrepancy in $\hat{\theta}$ was the largest for WLE across initial item difficulty conditions, regardless of test length. It was observed that the bias for WLE was 0.179 units after 50 items were administered and initial θ was –3. The bias for WLE after 50 items was 0.006 when initial θ was 3.

Initial items administered. The θ estimates used to select the items and the item parameters are provided for the first five items selected in the CAT in Tables 8 and 9. The θ estimate after the first item was administered was similar to b for WLE, regardless of whether the item response was correct or incorrect. It was found that the θ estimates for WLE differed across initial θ by 4.949 units in the MCR conditions and by 3.038 units in the MIR conditions after four items were administered. By comparison, after four items EAP θ estimates differed across initial θ by 1.513 and 0.238 units for the MIR and MCR conditions, respectively.

Tables 8 and 9 showed that EAP θ estimates were negative for the MIR conditions and

positive for the MCR conditions. In contrast, WLE θ estimates after the first item was administered were negative when initial θ was –3 and were positive when initial θ was 3, regardless of whether the examinee responded correctly or incorrectly.

Figure 69

Average Bias Across CAT Lengths for the Different Initial θ Conditions for θ = 3 (MIR)

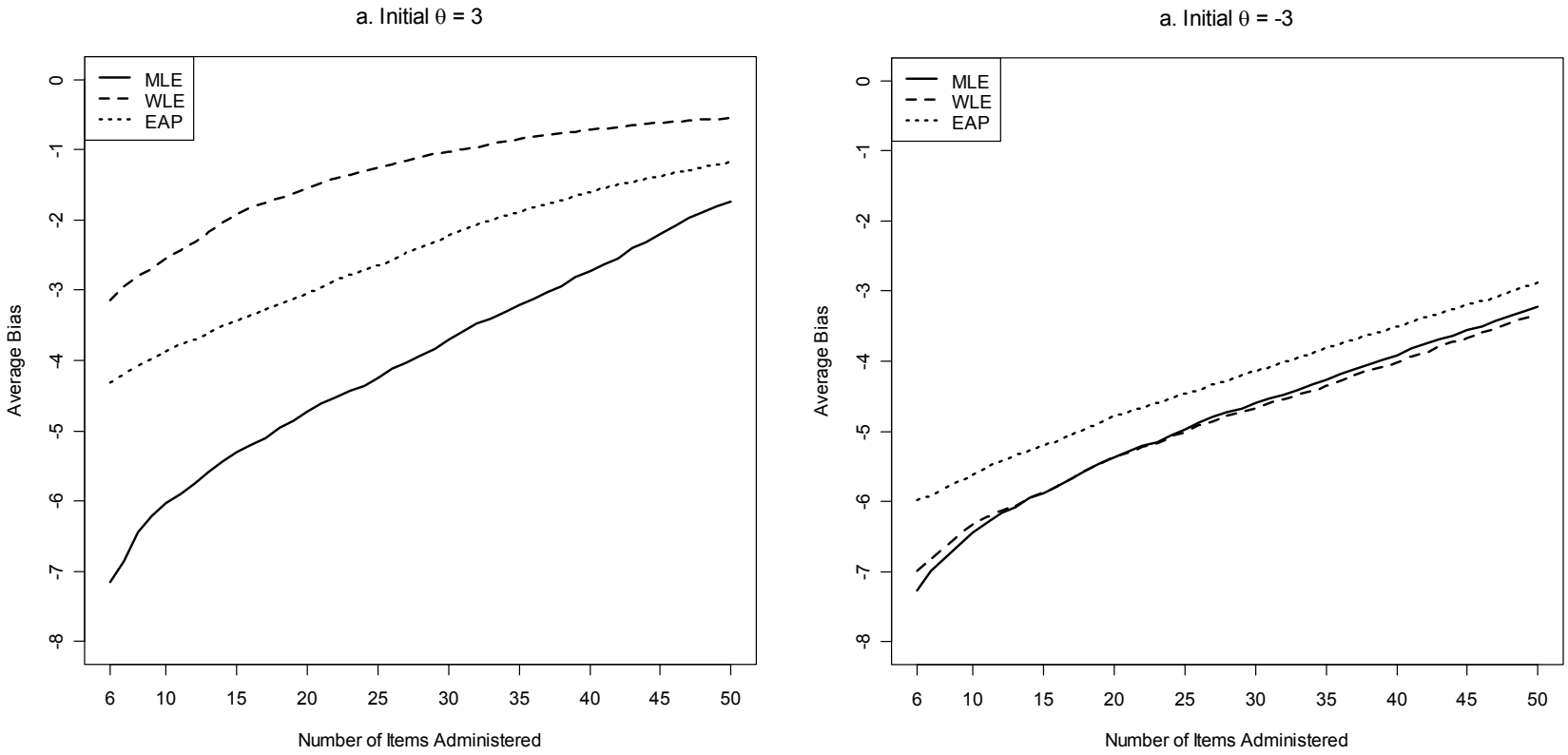


Figure 70

Average Bias Across CAT Lengths for the Different Initial θ Conditions for θ = −3 (MCR)

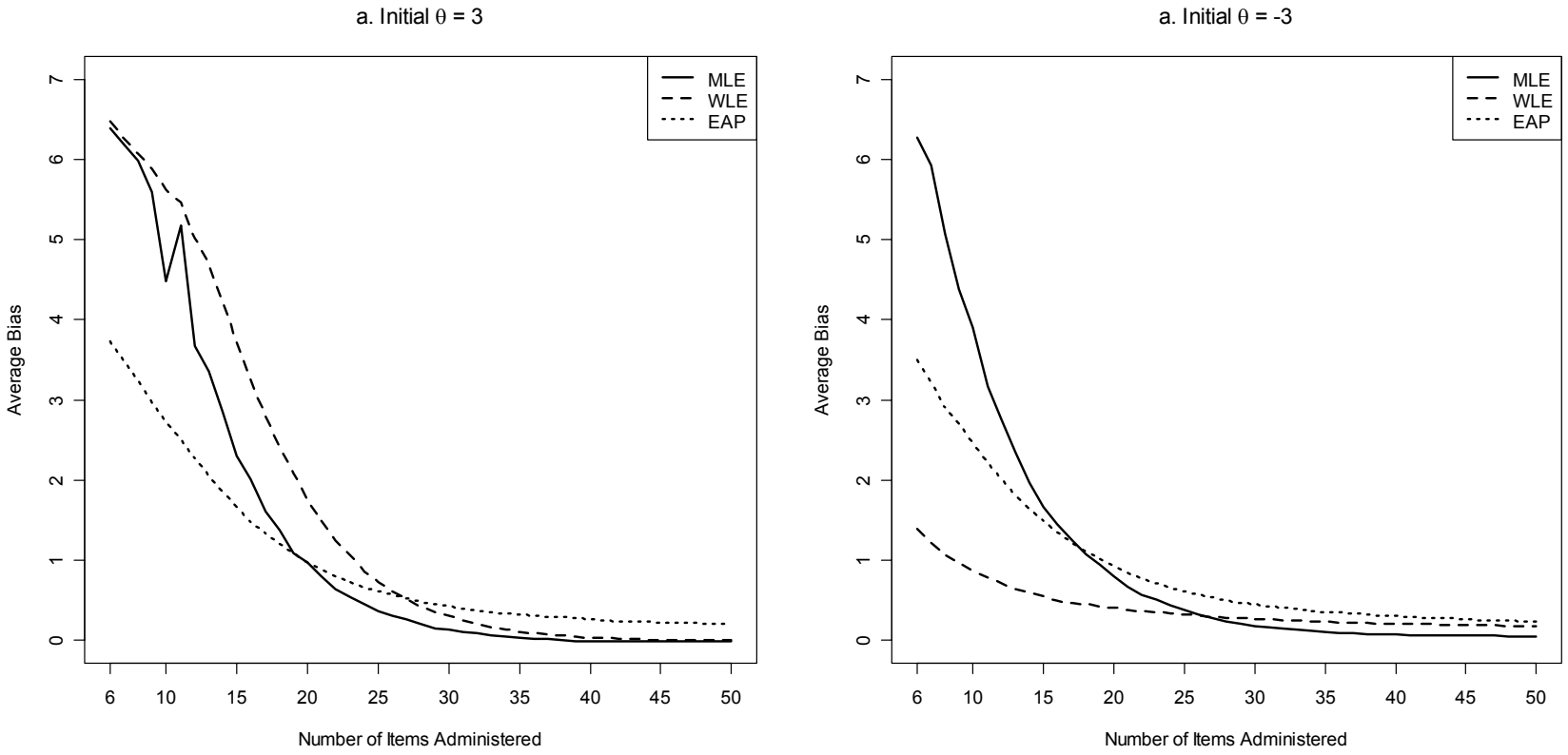


Table 7
Results From the Follow-Up Study for Different Test Lengths

Initial θ and	15 Items		25 Items		35 Items		50 Items	
θ Estimation	*Bias*	*SE*	*Bias*	*SE*	*Bias*	*SE*	*Bias*	*SE*
MIR								
−3								
MLE	−5.882	0.014	−4.974	0.017	−4.267	0.025	−3.217	0.056
WLE	−5.871	0.010	−5.004	0.016	−4.349	0.021	−3.325	0.051
EAP	−5.189	0.000	−4.468	0.014	−3.808	0.029	−2.881	0.076
3								
MLE	−5.308	0.012	−4.252	0.018	−3.212	0.049	−1.735	0.144
WLE	−1.919	0.133	−1.253	0.193	−0.853	0.238	−0.550	0.233
EAP	−3.443	0.027	−2.646	0.038	−1.826	0.118	−1.180	0.198
MCR								
−3								
MLE	1.672	2.050	0.385	1.031	0.104	0.491	0.047	0.277
WLE	0.549	0.420	0.323	0.385	0.235	0.286	0.179	0.249
EAP	1.490	0.864	0.616	0.777	0.357	0.341	0.233	0.248
3								
MLE	2.305	2.521	0.367	1.197	0.033	0.524	−0.007	0.295
WLE	3.712	1.559	0.728	1.399	0.108	0.641	0.006	0.297
EAP	3.490	0.860	0.611	0.565	0.323	0.359	0.201	0.250

Table 8

Item Parameters for the First Five Items Selected and θ Used to Select the Item for the MIR Conditions

Selection	MLE				WLE				EAP			
and Item No.	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$
Initial $\theta = -3$												
1	1.527	−3.179	.194	−3	1.527	−3.179	.194	−3	1.527	−3.179	.194	−3
2	1.005	−3.405	.193	−4	1.126	−3.176	.170	−3.602	1.338	−2.274	.206	−2.159
3	0.980	−3.352	.177	−4	1.005	−3.405	.193	−3.884	1.299	−2.863	.223	−2.641
4	1.126	−3.176	.170	−4	0.980	−3.352	.177	−4.240	1.126	−3.176	.170	−2.965
5	0.792	−3.478	.189	−4	0.792	−3.478	.189	−4.469	1.144	−2.936	.179	−3.177
Initial $\theta = 3$												
1	1.250	2.600	.197	3	1.250	2.600	.197	3	1.250	2.600	.197	3
2	1.005	−3.405	.193	−4	1.243	1.878	.213	2.084	1.118	−0.440	.196	−0.042
3	0.980	−3.352	.177	−4	1.151	0.969	.175	1.307	1.167	−0.831	.198	−0.758
4	1.126	−3.176	.170	−4	1.086	0.484	.192	0.356	1.301	−1.535	.224	−1.203
5	0.792	−3.478	.189	−4	1.118	−0.440	.196	−0.219	1.221	−1.691	.166	−1.664

Table 9

Item Parameters for the First Five Items Selected and θ Used to Select the Item for the MCR Conditions

Selection	MLE				WLE				EAP			
and Item No.	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$	*a*	*b*	*c*	$\hat{\theta}$
Initial $\theta = -3$												
1	1.527	−3.179	.194	−3	1.527	−3.179	.194	−3	1.527	−3.179	.194	−3
2	1.115	3.183	.222	4	1.299	−2.863	.223	−2.828	1.118	−0.440	.196	0.001
3	1.147	3.036	.178	4	1.338	−2.274	.206	−2.439	1.086	0.484	.192	0.325
4	0.837	3.414	.208	4	1.221	−1.691	.166	−1.857	1.151	0.969	.175	0.683
5	1.096	2.950	.206	4	1.301	−1.535	.224	−1.238	1.031	1.156	.161	1.043
Initial $\theta = 3$												
1	1.250	2.600	.197	3	1.250	2.600	.197	3	1.250	2.600	.197	3
2	1.115	3.183	.222	4	1.147	3.036	.178	3.029	1.118	−0.440	.196	0.152
3	1.147	3.036	.178	4	1.115	3.183	.222	3.554	1.086	0.484	.192	0.490
4	0.837	3.414	.208	4	1.096	2.950	.206	3.866	1.151	0.969	.175	0.886
5	1.096	2.950	.206	4	0.837	3.414	.208	3.999	1.031	1.156	.161	1.281

Discussion

Sensitivity of WLE to initial item difficulty. It was found that WLE was highly sensitive to initial item difficulty. The results provided evidence that the WLE θ estimates for the first item were a weighted function of item difficulty. This result was consistent with the observation that WLE adjusted the first derivative by a function of item difficulty.

These results produced a rather counterintuitive scenario – simulees could get four items correct and still have $\hat{\theta} = -1.238$. Likewise for the MIR conditions, a simulee could get the first three items incorrect but have a θ estimate of 0.356. All that was required was an initial item that was extreme on *b*.

Differences in item difficulty not item selection. The use of a global information index (K-L) resulted in a more difficult initial item being selected. As seen above, the WLE θ estimate after one item depended on the difficulty of that item. Through selection of a more difficult initial item, K-L selection provided less biased θ estimates for the MIR conditions, and more biased estimates for the MCR conditions. These results provided empirical evidence that the differences in $\hat{\theta}_{\text{WLE}}$ observed across item selection procedures resulted from the difference in initial item difficulty.

Conclusions

Recovery of θ When There Was No Misfit

Item Selection Method

The results provided evidence that K-L selection and FI selection resulted in similar recovery of θ in terms of bias, SE, and RMSE. Any consistent differences between these methods dissipated after 15 items were administered. Previous research (e.g., Cheng & Liou, 2000; Chen & Ankenmann, 2004) found that K-L selection reduced the bias, SE,

and RMSE for test lengths of about 10 items. Those results did not replicate in this study, as K-L selection did not reduce the bias, SE, or RMSE across the θ continuum. Differences in item bank structure might have contributed to this disparity.

θ Estimation Method

There was evidence that the differences in bias, SE, and RMSE between MLE and WLE dissipated as test length increased. The SEs and RMSEs for WLE were consistently lower than MLE, particularly when less than 20 items were administered. When θ was not at the mean of the prior it was found, in agreement with other research (Bock & Mislevy, 1982) and theoretical expectations, that EAP was more biased than MLE or WLE. However, EAP estimation had lower SEs than MLE or WLE. When θ was near the prior (less than 2 in absolute value) it was found that EAP had lower RMSEs than MLE or WLE due to the prior reducing the SEs of the θ estimates.

Recovery of θ When There Was Misfit

Direction of Misfit

It was found that CAT with the 3PL could recover from MCR due to the fact that the 3PL modeled guessing. A 50-item CAT resulted in highly biased θ estimates when there was MIR. It was evident that the adaptive testing procedure could not adequately recover from incorrect-response-based misfit for high θ examinees. Even a 50-item test and one initial incorrect response resulted in negatively biased θ estimates. When a low ability examinee responded correctly to the initial item(s) in the CAT, the CAT was able to recover to near-zero bias – given a sufficient (e.g., 25 items) test length. The SEs of the CAT were increased as a consequence of MCR.

Item Selection Method

For the MIR conditions, it was observed that item selection method affected the bias, SE, and RMSE throughout the 50-item CAT. It was observed for the MCR conditions that the differences in bias, SE, and RMSE across item selection methods dissipated as CAT length increased. However, it was shown in the follow-up study that the differences between item selection methods largely resulted from a difference in initial item difficulty. After the effect of initial item difficulty was considered, it can be concluded that item selection method did not affect recovery of θ when there was misfit.

θ Estimation Method

There was evidence that EAP provided the lowest bias and RMSEs of the θ estimation methods when there was misfit – for conditions with at least 2 misfitting items and a CAT length less than 15 items Though MLE and WLE provided similar results in terms of bias, it was evident that MLE θ estimates had extreme SEs when there was MCR. Thus, MLE resulted in θ estimates with the highest RMSEs across both MIR and MCR conditions. For this reason, it can be concluded that EAP generally provided the best recovery of θ and MLE the worst recovery of θ when there was misfit.

Sensitivity of WLE to Initial Item Difficulty

It was found that WLE was highly sensitive to initial item difficulty. This posed a practical problem for applied researchers who use WLE with CAT. If the initial item selected for one individual differed in difficulty by two units from that selected for another individual, and they had the same response pattern, then the θ estimates for the two individuals would differ solely due to the difference in initial item difficulty. This effect would disappear if there was no misfit because there was sufficient psychometric

information to reduce the effect of the bias correction on the θ estimates. For instance, Figures 8 and 9 show smaller differences across item selection methods than Figures 10 and 11, as there was more FI for $\theta = 1$ than there was FI for $\theta = 3$. EAP and MLE θ estimates also would differ in that circumstance, but WLE was shown to be much more sensitive to this difference.

The sensitivity of WLE to initial item difficulty made it possible to make erroneous conclusions about an individual solely because of responses that they made to the initial items when they were not consistent with the IRT model being used in the CAT. In light of the large differences in recovery for WLE across item selection procedures early in the CAT (which differed in initial item difficulty), the author does not recommend WLE for use in CATs shorter than 20 items.

Implications for Future Research

The results of this study showed serious mis-estimation of θ for high θ examinees when there were responses not predicted by the 3PL in the first four items in a CAT. The effect was evident even when only the first item response did not fit the IRT model and became more severe as additional items (up to four) had responses that did not meet model expectations. In addition, examinees with θ less than 0 had unstable (high SEs) θ estimates if they were to respond correctly to the initial item(s) in the CAT.

These results provided evidence that early misfit has severe implications for applied CATs. A CAT cannot recover θ for high ability examinees even when they responded incorrectly to just the first item. It is evident that additional research is necessary to help correct for the problem of misfit, as it was found that the recovery of θ (in terms of bias, SE, and RMSE) was adversely affected.

Use of EAP for Non-Mixed Response Patterns

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.

Robust Item Selection

Fixed Number of Items

This study found that the θ estimates used to select items when MIR was introduced were often well below the generating value. One possible method to improve recovery of θ would be to select items based on the recent responses of the examinee. One method would be to specify a fixed number of items for use in for the robust estimation of θ. An example of the fixed number of items method would be to estimate θ using the 10 most recent item responses and use that θ for item selection. If an examinee incorrectly answered the first item in a CAT, then items 2−11 would be used for selection of the twelfth and succeeding items in the CAT. Examinees with early MIR would have a chance to response correctly to difficult items with this method, and raise their θ estimate.

The robust selection method described above also could have benefits for MCR. Basing the item selection procedure on the recent string of item responses would likely introduce easier items to the examinee more quickly, and might reduce the rate of convergence

failures.

Target Information Criterion

It is known that FI selection uses the most informative items early in the CAT, and FI decreases as test length increases due to items with less information being available in the bank. This presents a problem for longer CATs, as the robust θ values used in the item selection would be increasingly imprecise as test length increases. One method to eliminate this problem would be to define a minimum amount of FI necessary for estimation of the robust θ for item selection. The most recently administered items, with a combined FI greater than the criterion, would be used in estimation of θ. This target information criterion (TIC) must be realistic given the information structure of the test bank. For example, the researcher could specify that the items used in the robust procedure have total FI of no less than 4.

The TIC would ensure stability in the precision of the robust θ estimate used in the item selection routine. In addition, it would be expected that the TIC would require fewer items early in the CAT and more items later in the CAT. Additional research is needed to examine the effects of a TIC on the recovery of θ when there is misfit present.

Modeling MIR in CAT

It was observed that modeling guessing resulted in unbiased θ estimates for the MCR conditions. However, the 3PL model does model incorrect responses for high ability individuals due to psychological factors or "carelessness". As a consequence, the θ estimates for simulees with MIR did not recover to zero bias after 50 items were administered. As MIR would be expected due to psychological causes in applied testing applications (nervousness, unfamiliarity with the computer station), it followed that the

introduction of an upper-asymptote parameter (*d*) might help with recovery in those circumstances. Barton and Lord (1981) proposed such a four-parameter dichotomous IRT model (4PM). Following Barton and Lord's paper there was little interest in the 4PM until Reise and Waller (2003) proposed an application to psychopathology data. Waller and Reise (in press) described their research in which the 4PM was fit to psychopathology data. They indicated that the upper asymptote for many pathology items was not 1.0, and they speculated that this was due to examinees with a high level of the trait not always endorsing the item.

Future research is needed that examines how the introduction of an upper-asymptote parameter affects recovery of θ for high θ examinees. It would be necessary to assess the recovery of θ for different values of *d,* to provide recommendations for different CAT lengths. One practical issue with the introduction of a *d* parameter is convergence failures. A convergence failure would occur when an examinee responds correctly to a proportion of items that was greater than the upper asymptote of the TRF. For this reason, a small *d* parameter of .01 would be recommendable over a larger *d* (e.g., .05 or .10).

In addition, the effect of *d* on the bias, SE, and RMSE for high θ examinees that do not misfit the model must be investigated. The effect of a small *d* parameter on the empirical and theoretical SEs of θ must also be investigated. As θ failed to recover without modeling carelessness, it was evident that MIR posed a practical problem that should not be ignored by high-stakes testing.

Limitations of the Current Study

The current study used an item bank that produced an essentially flat BIF for much of the θ range. The rationale for this decision was to minimize the effect of 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.

Chang, H-H., & Ying, Z. L. (2008). To weight or not to weight? Balancing influence of initial items in adaptive testing. *Psychometrika, 73,* 441-450.

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.

Choi, S. W. (2007). *FIRESTAR: Computerized adaptive testing (CAT) simulation program for polytomous IRT models.* [Computer software and manual.] Evanston, IL: Evanston Northwestern Healthcare Research Institute.

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

Flaugher, R. (2000). Item pools. In Wainer, H. (2000). Computerized adaptive testing: a primer. Mahwah, NJ: Erlbaum.

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.

Owen, Roger 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.

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

Waller, N. G., & Reise, S. P. (in press). 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.

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1,* 233-247.

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

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.

van der Linden, W. J. & Pashley, P. J. (2000). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (eds.), *Computerized adaptive testing: Theory and practice* (pp.1-25). Norwell MA: Kluwer.

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

Wang, T. (1997, March). *Essentially unbiased EAP estimates in computerized adaptive testing.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of 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.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37,* 70-84.

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.

APPENDIX

Table A1
Number of MLE θ Convergence Failures Per 1,000 for θ Conditions When at Least One Estimate Did Not Converge

Length of the CAT, Item Selection, and θ Condition	Number of Misfitting Items 0	1	2	3	4
15 Items					
Max. Info.					
−3	5	30	50	57	27
−2.5	2	11	21	27	15
−2	3	6	9	17	6
−1	0	0	1	2	0
3	1	0	0	0	0
K-L					
−3	15	36	71	116	53
−2.5	6	16	68	104	69
−2	2	15	44	86	39
−1	1	3	16	56	26
0 (correct)	0	0	4	11	2
25 Items					
Max. Info.					
−3	2	4	5	30	50
−2.5	0	2	2	14	19
−2	0	0	4	8	19
−1	0	0	0	0	1
K-L					
−3	1	5	16	27	74
−2.5	0	3	5	18	44
−2	0	2	3	8	47
−1	0	0	0	4	12
0 (correct)	0	0	1	0	3

Table A1, cont.
Number of MLE θ Convergence Failures Per 1,000 for θ Conditions When at Least One Estimate Did Not Converge

Length of the CAT, Item Selection, and θ Condition	Number of Misfitting Items 0	1	2	3	4
35 Items					
Max. Info.					
−3	1	1	2	4	12
−2.5	0	1	1	2	4
−2	0	0	0	1	1
K-L					
−3	0	1	2	6	13
−2.5	0	1	0	3	6
−2	0	0	2	0	2
−1	0	0	0	0	3
50 Items					
Max. Info.					
−3	0	0	0	0	1
K-L					
−3	0	0	0	0	1

Table A2
Summary Statistics for the θ Main Effect From the ANOVA

	15 Items		25 Items		35 Items		50 Items	
θ	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
−3	1.205	1.483	0.341	0.790	0.125	0.432	0.063	0.290
−2.5	0.986	1.304	0.281	0.690	0.115	0.399	0.060	0.271
−2	0.830	1.152	0.255	0.627	0.112	0.379	0.064	0.267
−1	0.592	0.847	0.233	0.519	0.128	0.365	0.081	0.279
0 (MCR)	0.352	0.599	0.177	0.409	0.119	0.331	0.082	0.274
0 (MIR)	−0.686	0.758	−0.447	0.543	−0.321	0.417	−0.229	0.324
1	−1.216	1.059	−0.812	0.804	−0.568	0.607	−0.373	0.426
2	−1.831	1.404	−1.277	1.140	−0.906	0.894	−0.560	0.600
2.5	−2.170	1.573	−1.548	1.320	−1.111	1.070	−0.686	0.732
3	−2.528	1.737	−1.835	1.513	−1.337	1.260	−0.831	0.894

Table A3
Summary Statistics for the Misfit Main Effect From the ANOVA

	15 Items		25 Items		35 Items		50 Items	
Item Misfit	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
0	0.020	0.474	0.008	0.349	0.006	0.295	0.004	0.253
1	−0.173	0.737	−0.092	0.439	−0.064	0.342	−0.044	0.278
2	−0.584	1.480	−0.406	0.815	−0.241	0.505	−0.138	0.350
3	−0.787	2.186	−0.788	1.369	−0.587	0.934	−0.330	0.546
4	−0.710	2.914	−1.037	1.918	−0.935	1.400	−0.656	0.951

Table A4
Summary Statistics for the θ ×Misfit Interaction From the ANOVA After 15 Items Were Administered

	Zero Items		One Item		Two Items		Three Items		Four Items	
θ	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
−3	0.226	0.543	0.403	0.769	0.832	1.162	1.653	1.462	2.910	1.334
−2.5	0.163	0.479	0.312	0.678	0.707	1.037	1.326	1.318	2.421	1.274
−2	0.134	0.445	0.270	0.624	0.585	0.913	1.094	1.166	2.069	1.180
−1	0.097	0.441	0.230	0.531	0.466	0.696	0.782	0.833	1.384	0.936
0 (MCR)	0.029	0.424	0.122	0.453	0.296	0.487	0.499	0.577	0.812	0.666
0 (MIR)	0.030	0.423	−0.142	0.371	−0.520	0.359	−1.109	0.417	−1.699	0.479
1	−0.059	0.425	−0.363	0.407	−1.081	0.363	−1.940	0.492	−2.637	0.510
2	−0.100	0.436	−0.635	0.448	−1.900	0.370	−2.895	0.518	−3.626	0.518
2.5	−0.140	0.451	−0.827	0.461	−2.370	0.373	−3.389	0.523	−4.124	0.519
3	−0.176	0.484	−1.095	0.495	−2.858	0.375	−3.885	0.524	−4.624	0.519

Table A5
Summary Statistics for the θ ×Misfit Interaction From the ANOVA After 50 Items Were Administered

	Zero Items		One Item		Two Items		Three Items		Four Items	
θ	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
−3	0.059	0.260	0.054	0.271	0.054	0.284	0.074	0.301	0.075	0.329
−2.5	0.041	0.238	0.049	0.253	0.066	0.269	0.059	0.279	0.083	0.309
−2	0.045	0.242	0.053	0.248	0.059	0.269	0.076	0.274	0.088	0.298
−1	0.035	0.253	0.065	0.260	0.084	0.262	0.090	0.285	0.132	0.320
0 (MCR)	0.010	0.250	0.039	0.255	0.075	0.267	0.125	0.278	0.162	0.292
0 (MIR)	0.006	0.247	−0.054	0.242	−0.180	0.245	−0.349	0.241	−0.569	0.263
1	−0.020	0.245	−0.114	0.252	−0.273	0.256	−0.522	0.285	−0.934	0.312
2	−0.042	0.252	−0.169	0.254	−0.371	0.267	−0.745	0.298	−1.476	0.428
2.5	−0.049	0.249	−0.187	0.265	−0.417	0.296	−0.930	0.309	−1.847	0.476
3	−0.049	0.260	−0.181	0.277	−0.474	0.312	−1.174	0.339	−2.228	0.515

Table A6

Summary Statistics for the θ × Estimation × Item Selection Interaction From the ANOVA After 15 Items

	MLE		WLE		EAP	
θ	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
FI Selection						
−3	1.266	1.758	0.915	1.325	1.086	0.870
−2.5	1.072	1.551	0.757	1.143	0.842	0.777
−2	0.921	1.368	0.650	0.967	0.673	0.667
−1	0.679	0.998	0.479	0.684	0.441	0.510
0 (MCR)	0.423	0.688	0.248	0.488	0.190	0.404
0 (MIR)	−0.891	0.861	−0.945	0.855	−0.485	0.488
1	−1.456	1.209	−1.520	1.210	−1.007	0.710
2	−2.066	1.617	−2.171	1.579	−1.630	1.009
2.5	−2.403	1.796	−2.537	1.734	−1.990	1.147
3	−2.755	1.975	−2.915	1.897	−2.385	1.267
K-L Selection						
−3	1.398	1.844	1.502	1.773	1.061	0.899
−2.5	1.186	1.604	1.238	1.570	0.819	0.805
−2	1.027	1.434	1.052	1.404	0.658	0.708
−1	0.749	1.046	0.769	1.026	0.434	0.555
0 (MCR)	0.500	0.709	0.505	0.693	0.245	0.440
0 (MIR)	−0.804	0.875	−0.581	0.688	−0.409	0.497
1	−1.338	1.225	−1.071	0.992	−0.904	0.700
2	−1.904	1.604	−1.664	1.346	−1.514	0.980
2.5	−2.225	1.814	−1.978	1.543	−1.863	1.124
3	−2.561	2.027	−2.308	1.729	−2.241	1.250

Table A7

Summary Statistics for the θ × Estimation × Item Selection Interaction From the ANOVA After 50 Items

	MLE		WLE		EAP	
θ	*Mean*	*SD*	*Mean*	*SD*	*Mean*	*SD*
FI Selection						
−3	0.001	0.301	0.011	0.280	0.181	0.251
−2.5	0.020	0.274	0.020	0.264	0.152	0.245
−2	0.040	0.274	0.042	0.268	0.145	0.250
−1	0.070	0.285	0.089	0.278	0.129	0.254
0 (MCR)	0.071	0.276	0.073	0.264	0.072	0.249
0 (MIR)	−0.287	0.353	−0.304	0.359	−0.205	0.276
1	−0.432	0.487	−0.466	0.510	−0.351	0.336
2	−0.637	0.721	−0.697	0.764	−0.518	0.422
2.5	−0.777	0.877	−0.852	0.927	−0.642	0.502
3	−0.938	1.067	−1.027	1.122	−0.795	0.632
K-L Selection						
−3	0.001	0.296	0.008	0.291	0.177	0.248
−2.5	0.014	0.277	0.008	0.276	0.143	0.244
−2	0.020	0.269	0.012	0.267	0.128	0.244
−1	0.052	0.292	0.043	0.292	0.106	0.259
0 (MCR)	0.097	0.295	0.092	0.295	0.090	0.265
0 (MIR)	−0.236	0.345	−0.181	0.301	−0.161	0.273
1	−0.373	0.455	−0.298	0.380	−0.315	0.324
2	−0.578	0.659	−0.447	0.498	−0.485	0.392
2.5	−0.701	0.817	−0.543	0.614	−0.601	0.472
3	−0.839	0.992	−0.655	0.760	−0.732	0.584

Table A8

Summary Statistics After 15 Items for the θ Estimates for θ = −3 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.057	.560	.563	.083	.541	.547
1	.267	.896	.657	.295	.866	.915
2	.859	1.474	1.706	.803	1.421	1.632
3	1.759	1.710	2.454	2.010	1.832	2.721
4	3.387	1.368	3.655	3.798	1.068	3.947
WLE						
0	.048	.466	.468	.129	.531	.546
1	.166	.636	.657	.344	.879	.944
2	.477	.943	1.057	1.043	1.385	1.734
3	1.326	1.264	1.761	2.265	1.600	2.775
4	2.659	1.015	2.848	3.731	1.096	3.890
EAP						
0	.496	.456	.674	.544	.438	.699
1	.703	.557	.897	.642	.528	.831
2	.970	.712	1.204	.839	.641	1.056
3	1.326	.825	1.563	1.329	.893	1.601
4	1.935	.893	2.132	1.951	1.005	2.195

Table A9

Summary Statistics After 15 Items for the θ Estimates for θ = −2.5 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.073	.517	.522	.059	.487	.491
1	.263	.813	.855	.223	.746	.778
2	.722	1.228	1.425	.754	1.316	1.516
3	1.457	1.579	2.149	1.692	1.596	2.327
4	2.485	1.423	3.183	3.204	1.039	3.370
WLE						
0	.045	.444	.447	.059	.475	.479
1	.122	.566	.579	.263	.781	.824
2	.417	.815	.916	.914	1.278	1.572
3	1.016	1.107	1.503	1.804	1.497	2.345
4	2.182	1.001	2.402	3.150	1.093	3.336
EAP						
0	.375	.401	.549	.366	.400	.543
1	.522	.523	.739	.478	.466	.668
2	.774	.637	1.002	.658	.611	.899
3	.976	.754	1.234	1.010	.806	1.292
4	1.562	.863	1.786	1.583	.935	1.839

Table A10

Summary Statistics After 15 Items for the θ Estimates for θ = −2 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.074	.469	.475	.059	.473	.477
1	.225	.676	.713	.206	.762	.789
2	.569	1.040	1.185	.670	1.175	1.353
3	1.237	1.410	1.876	1.379	1.458	2.007
4	2.499	1.303	2.819	2.821	.946	2.977
WLE						
0	.040	.404	.406	.044	.468	.470
1	.162	.522	.547	.219	.749	.780
2	.393	.733	.832	.762	1.148	1.378
3	.826	.914	1.232	1.483	1.345	2.003
4	1.827	.881	2.029	2.752	1.038	2.943
EAP						
0	.298	.384	.487	.289	.369	.469
1	.432	.458	.630	.373	.446	.581
2	.581	.551	.801	.536	.555	.772
3	.816	.658	1.048	.820	.708	1.084
4	1.239	.764	1.457	1.273	.852	1.532

Table A11

Summary Statistics After 15 Items for the θ Estimates for θ = −1 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.060	.481	.485	.093	.469	.478
1	.254	.566	.621	.204	.645	.676
2	.534	.779	.944	.535	.900	1.047
3	.875	.980	1.314	1.002	1.052	1.453
4	1.671	1.120	2.013	1.909	.860	2.095
WLE						
0	.023	.439	.440	.075	.479	.485
1	.178	.475	.508	.203	.635	.667
2	.376	.538	.656	.586	.852	1.034
3	.628	.582	.857	1.066	.966	1.439
4	1.191	.676	1.370	1.912	.863	2.099
EAP						
0	.155	.376	.407	.177	.367	.407
1	.280	.394	.483	.261	.409	.486
2	.416	.444	.609	.347	.486	.597
3	.565	.500	.754	.555	.563	.790
4	.789	.561	.969	.829	.641	1.048

Table A12

Summary Statistics After 15 Items for the θ Estimates for θ = 0 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.055	.467	.470	.053	.446	.449
1	.106	.454	.466	.185	.543	.574
2	.365	.526	.640	.405	.570	.700
3	.569	.669	.879	.705	.666	.970
4	1.020	.776	1.282	1.151	.691	1.343
WLE						
0	.017	.444	.445	.023	.454	.455
1	.077	.431	.438	.201	.505	.544
2	.188	.411	.452	.432	.537	.690
3	.356	.443	.568	.716	.635	.958
4	.600	.471	.763	1.154	.670	1.335
EAP						
0	.012	.359	.360	.013	.357	.358
1	.053	.353	.357	.112	.385	.401
2	.176	.363	.404	.211	.404	.455
3	.291	.388	.485	.359	.424	.555
4	.419	.409	.585	.529	.431	.683

Table A13

Summary Statistics After 15 Items for the θ Estimates for θ = 0 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.046	.462	.465	.072	.450	.456
1	−.213	.370	.427	−.059	.386	.391
2	−.708	.371	.764	−.560	.372	.672
3	−1.436	.148	1.445	−1.433	.140	1.440
4	−2.146	.065	2.148	−2.041	.086	2.043
WLE						
0	.000	.449	.449	.039	.452	.453
1	−.262	.371	.455	−.087	.387	.397
2	−.770	.279	.820	−.371	.345	.507
3	−1.525	.126	1.531	−.878	.251	.913
4	−2.167	.068	2.169	−1.609	.115	1.614
EAP						
0	.008	.354	.354	.017	.353	.354
1	−.171	.319	.361	−.060	.337	.342
2	−.409	.269	.489	−.302	.319	.440
3	−.742	.241	.780	−.642	.274	.698
4	−1.114	.163	1.126	−1.061	.168	1.075

Table A14

Summary Statistics After 15 Items for the θ Estimates for θ = 1 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.008	.434	.435	.007	.456	.456
1	−.401	.437	.593	−.216	.423	.475
2	−1.392	.187	1.405	−1.087	.273	1.121
3	−2.363	.065	2.365	−2.374	.058	2.376
4	−3.132	.023	3.134	−3.020	.040	3.022
WLE						
0	−.006	.453	.453	−.008	.451	.451
1	−.504	.420	.656	−.253	.409	.481
2	−1.469	.183	1.481	−.841	.266	.883
3	−2.469	.053	2.470	−1.682	.134	1.688
4	−3.153	.028	3.155	−2.567	.047	2.569
EAP						
0	−.175	.343	.385	−.180	.347	.391
1	−.460	.331	.567	−.346	.330	.478
2	−.943	.264	.979	−.754	.277	.804
3	−1.457	.165	1.467	−1.297	.194	1.312
4	−2.003	.080	2.005	−1.945	.090	1.949

Table A15

Summary Statistics After 15 Items for the θ Estimates for θ = 2 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.046	.440	.442	−.005	.414	.414
1	−.590	.510	.780	−.386	.419	.570
2	−2.305	.081	2.307	−1.929	.136	1.935
3	−3.351	.027	3.353	−3.365	.022	3.367
4	−4.131	.010	4.133	−4.016	.010	4.018
WLE						
0	.009	.464	.464	−.033	.416	.417
1	−.871	.412	.964	−.449	.418	.614
2	−2.385	.077	2.387	−1.637	.150	1.645
3	−3.459	.021	3.461	−2.641	.056	2.643
4	−4.150	.010	4.153	−3.561	.013	3.562
EAP						
0	−.293	.355	.461	−.324	.360	.484
1	−.824	.356	.898	−.688	.324	.761
2	−1.689	.152	1.696	−1.453	.182	1.465
3	−2.367	.071	2.369	−2.187	.102	2.191
4	−2.980	.038	2.981	−2.917	.036	2.918

Table A16

Summary Statistics After 15 Items for the θ Estimates for θ = 2.5 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	−.007	.428	.428	.021	.425	.426
1	−.730	.443	.854	−.485	.473	.677
2	−2.795	.056	2.797	−2.404	.082	2.406
3	−3.850	.023	3.852	−3.863	.008	3.865
4	−4.630	.000	4.632	−4.516	.000	4.518
WLE						
0	−.042	.451	.453	−.005	.418	.418
1	−1.158	.319	1.201	−.589	.452	.743
2	−2.876	.055	2.878	−2.101	.095	2.104
3	−3.959	.016	3.961	−3.134	.028	3.136
4	−4.650	.000	4.652	−4.060	.006	4.062
EAP						
0	−.408	.361	.545	−.398	.371	.544
1	−1.062	.297	1.103	−.938	.319	.991
2	−2.147	.108	2.151	−1.897	.131	1.902
3	−2.858	.052	2.860	−2.666	.060	2.668
4	−3.475	.015	3.476	−3.414	.022	3.415

Table A17

Summary Statistics After 15 Items for the θ Estimates for θ = 3 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.024	.448	.449	.034	.431	.433
1	−1.026	.332	1.079	−.567	.517	.767
2	−3.292	.044	3.294	−2.985	.050	2.897
3	−4.348	.000	4.351	−4.363	.008	4.365
4	−5.130	.000	5.133	−5.016	.000	5.018
WLE						
0	−.030	.460	.461	−.020	.410	.411
1	−1.566	.214	1.581	−.741	.418	.851
2	−3.373	.041	3.375	−2.586	.060	2.588
3	−4.458	.000	4.460	−3.632	.007	3.634
4	−5.150	.000	5.153	−4.560	.000	4.562
EAP						
0	−.543	.354	.649	−.522	.364	.637
1	−1.424	.221	1.442	−1.245	.252	1.271
2	−2.632	.084	2.634	−2.367	.088	2.370
3	−3.353	.035	3.355	−3.158	.031	3.160
4	−3.974	.015	3.976	−3.913	.020	3.915

Table A18

Summary Statistics After 25 Items for the θ Estimates for θ = −3 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	−.001	.367	.367	.022	.358	.359
1	.039	.456	.458	.044	.452	.454
2	.149	.739	.754	.120	.605	.617
3	.315	.952	1.003	.359	1.039	1.099
4	.843	1.417	1.649	.923	1.495	1.757
WLE						
0	.007	.326	.327	.035	.345	.346
1	.035	.393	.394	.066	.450	.455
2	.091	.522	.530	.158	.617	.637
3	.246	.732	.772	.426	1.028	1.113
4	.526	.482	.714	1.097	1.443	1.813
EAP						
0	.296	.323	.438	.307	.312	.438
1	.350	.358	.501	.325	.354	.481
2	.399	.417	.577	.373	.362	.520
3	.526	.482	.714	.518	.506	.725
4	.738	.663	.993	.706	.674	.976

Table A19

Summary Statistics After 25 Items for the θ Estimates for θ = −2.5 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.029	.349	.350	.021	.338	.339
1	.031	.446	.449	.043	.390	.393
2	.166	.600	.623	.125	.625	.637
3	.320	.858	.916	.295	.896	.943
4	.732	1.266	1.463	.705	1.261	1.445
WLE						
0	.022	.334	.334	.012	.340	.340
1	.031	.354	.355	.046	.399	.401
2	.136	.460	.479	.142	.634	.649
3	.242	.661	.704	.329	.898	.957
4	.604	.940	1.117	.777	1.256	1.477
EAP						
0	.227	.309	.384	.218	.304	.374
1	.277	.357	.452	.260	.335	.425
2	.368	.411	.552	.294	.381	.482
3	.418	.469	.629	.389	.456	.599
4	.597	.590	.840	.558	.589	.811

Table A20

Summary Statistics After 25 Items for the θ Estimates for θ = −2 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.038	.327	.329	.024	.341	.341
1	.093	.399	.410	.042	.393	.395
2	.158	.544	.567	.114	.596	.607
3	.293	.792	.844	.265	.805	.847
4	.660	1.102	1.285	.663	1.132	1.312
WLE						
0	.018	.308	.309	.010	.331	.331
1	.068	.349	.356	.033	.402	.403
2	.134	.465	.484	.126	.592	.606
3	.252	.557	.612	.300	.792	.847
4	.576	.828	1.009	.715	1.138	1.344
EAP						
0	.192	.299	.355	.186	.306	.358
1	.250	.333	.416	.207	.325	.386
2	.288	.397	.491	.251	.374	.451
3	.363	.413	.550	.333	.417	.534
4	.519	.551	.757	.488	.535	.724

Table A21

Summary Statistics After 25 Items for the θ Estimates for θ = −1 with Misfit as Correct Responses

θ Estimation and Item Misfit	FI			K-L Information		
	Bias	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.032	.349	.351	.046	.352	.355
1	.120	.381	.400	.101	.422	.434
2	.222	.487	.535	.167	.519	.545
3	.315	.576	.656	.291	.698	.756
4	.513	.789	.941	.549	.905	1.058
WLE						
0	.014	.339	.340	.033	.353	.354
1	.093	.356	.368	.090	.421	.431
2	.206	.397	.447	.159	.525	.549
3	.311	.425	.527	.296	.674	.736
4	.510	.550	.750	.581	.855	1.034
EAP						
0	.104	.313	.330	.123	.310	.333
1	.181	.324	.372	.162	.333	.371
2	.249	.347	.427	.181	.340	.386
3	.313	.380	.492	.246	.402	.471
4	.387	.426	.576	.397	.476	.620

Table A22

Summary Statistics After 25 Items for the θ Estimates for θ = 0 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.029	.342	.343	.036	.339	.341
1	.055	.342	.346	.115	.380	.397
2	.195	.386	.432	.209	.419	.468
3	.244	.450	.512	.341	.483	.591
4	.349	.542	.645	.463	.545	.715
WLE						
0	.010	.340	.340	.021	.347	.348
1	.046	.331	.334	.111	.377	.393
2	.124	.333	.355	.207	.425	.473
3	.208	.345	.402	.336	.487	.592
4	.339	.374	.505	.469	.540	.715
EAP						
0	.010	.299	.299	.015	.297	.297
1	.031	.292	.294	.074	.318	.327
2	.120	.314	.336	.127	.339	.362
3	.189	.319	.371	.237	.355	.427
4	.271	.343	.438	.327	.376	.498

Table A23

Summary Statistics After 25 Items for the θ Estimates for θ = 0 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.019	.341	.341	.034	.342	.344
1	−.132	.317	.343	−.035	.316	.318
2	−.424	.298	.518	−.321	.325	.457
3	−.861	.243	.895	−.793	.288	.845
4	−1.507	.135	1.514	−1.429	.153	1.438
WLE						
0	−.003	.340	.340	.021	.342	.343
1	−.159	.315	.352	−.055	.312	.316
2	−.461	.293	.546	−.222	.306	.378
3	−.941	.227	.969	−.525	.283	.597
4	−1.552	.120	1.558	−1.001	.221	1.026
EAP						
0	.001	.293	.293	.016	.294	.294
1	−.121	.285	.309	−.041	.275	.278
2	−.282	.263	.386	−.209	.284	.353
3	−.516	.249	.573	−.424	.261	.497
4	−.775	.209	.803	−.701	.228	.738

Table A24

Summary Statistics After 25 Items for the θ Estimates for θ = 1 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.012	.328	.328	.005	.340	.340
1	−.217	.352	.414	−.127	.336	.359
2	−.742	.316	.807	−.574	.329	.661
3	−1.561	.165	1.570	−1.391	.222	1.409
4	−2.422	.062	2.424	−2.330	.080	2.333
WLE						
0	.000	.328	.328	−.004	.336	.336
1	−.251	.356	.436	−.140	.331	.359
2	−.811	.306	.867	−.448	.314	.547
3	−1.693	.138	1.700	−.994	.252	1.026
4	−2.480	.056	2.482	−1.782	.120	1.787
EAP						
0	−.115	.290	.312	−.115	.302	.323
1	−.290	.303	.420	−.227	.290	.369
2	−.582	.285	.649	−.483	.280	.558
3	−.961	.245	.992	−.844	.257	.883
4	−1.453	.157	1.462	−1.333	.175	1.345

Table A25

Summary Statistics After 25 Items for the θ Estimates for θ = 2 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.024	.342	.343	−.006	.338	.338
1	−.276	.381	.471	−.225	.327	.397
2	−1.231	.295	1.267	−.950	.317	1.002
3	−2.469	.072	2.472	−2.259	.114	2.263
4	−3.407	.029	3.409	−3.306	.028	3.307
WLE						
0	.013	.350	.350	−.020	.335	.335
1	−.342	.387	.517	−.252	.326	.413
2	−1.396	.229	1.416	−.766	.323	.832
3	−2.626	.060	2.628	−1.763	.151	1.770
4	−3.466	.024	3.468	−2.733	.052	2.734
EAP						
0	−.200	.295	.356	−.219	.303	.374
1	−.484	.316	.578	−.434	.290	.522
2	−.974	.267	1.010	−.843	.269	.885
3	−1.659	.150	1.667	−1.508	.179	1.519
4	−2.340	.085	2.343	−2.193	.089	2.196

Table A26

Summary Statistics After 25 Items for the θ Estimates for θ = 2.5 with Misfit as Incorrect Responses

θ Estimation		FI			K-L Information		
and Item Misfit		*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE							
	0	−.009	.325	.326	.012	.342	.343
	1	−.299	.400	.500	−.257	.365	.446
	2	−1.579	.232	1.597	−1.262	.238	1.285
	3	−2.964	.057	2.966	−2.735	.068	2.737
	4	−3.903	.012	3.905	−3.803	.016	3.805
WLE							
	0	−.015	.329	.329	−.001	.336	.336
	1	−.392	.412	.569	−.297	.370	.475
	2	−1.791	.172	1.800	−.989	.295	1.032
	3	−3.121	.044	3.123	−2.213	.100	2.217
	4	−3.963	.009	3.965	−3.226	.033	3.228
EAP							
	0	−.273	.306	.410	−.257	.307	.400
	1	−.605	.310	.680	−.567	.306	.644
	2	−1.278	.220	1.298	−1.112	.233	1.137
	3	−2.101	.112	2.105	−1.929	.125	1.934
	4	−2.821	.052	2.822	−2.678	.064	2.680

Table A27

Summary Statistics After 25 Items for the θ Estimates for θ = 3 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.016	.352	.352	.030	.331	.332
1	−.334	.411	.530	−.237	.402	.466
2	−2.011	.174	2.019	−1.664	.166	1.673
3	−3.456	.031	3.458	−3.226	.037	3.228
4	−4.403	.011	4.405	−4.303	.014	4.305
WLE						
0	−.006	.352	.352	.003	.324	.324
1	−.452	.428	.623	−.286	.388	.482
2	−2.250	.132	2.255	−1.318	.235	1.340
3	−3.616	.024	3.618	−2.689	.061	2.691
4	−4.464	.012	4.466	−3.723	.022	3.725
EAP						
0	−.320	.310	.446	−.312	.306	.438
1	−.764	.307	.824	−.669	.326	.744
2	−1.668	.163	1.677	−1.482	.181	1.494
3	−2.575	.078	2.577	−2.382	.079	2.384
4	−3.317	.046	3.319	−3.170	.053	3.172

Table A28

Summary Statistics After 35 Items for the θ Estimates for θ = −3 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	−.002	.304	.304	.010	.297	.298
1	−.001	.351	.351	.007	.326	.326
2	.022	.411	.411	.005	.356	.356
3	.052	.492	.495	.073	.533	.538
4	.160	.743	.760	.156	.721	.738
WLE						
0	.007	.283	.283	.020	.289	.290
1	.008	.312	.312	.012	.315	.315
2	.023	.355	.356	.016	.347	.348
3	.069	.411	.417	.096	.520	.529
4	.126	.563	.577	.204	.731	.759
EAP						
0	.215	.272	.347	.222	.264	.345
1	.236	.286	.371	.224	.277	.356
2	.244	.303	.390	.235	.290	.373
3	.293	.323	.436	.299	.335	.449
4	.365	.382	.529	.339	.384	.512

Table A29

Summary Statistics After 35 Items for the θ Estimates for θ = −2.5 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.013	.289	.289	.014	.269	.269
1	.023	.299	.300	.025	.298	.299
2	.056	.341	.345	.018	.350	.351
3	.084	.448	.456	.059	.461	.465
4	.172	.642	.665	.182	.700	.723
WLE						
0	.006	.275	.276	.008	.265	.265
1	.008	.282	.282	.021	.304	.305
2	.065	.325	.332	.017	.346	.347
3	.081	.402	.410	.061	.474	.478
4	.145	.521	.541	.196	.702	.728
EAP						
0	.163	.263	.309	.172	.246	.300
1	.187	.280	.337	.189	.278	.336
2	.244	.300	.386	.186	.285	.341
3	.246	.315	.400	.225	.312	.385
4	.306	.377	.485	.291	.374	.474

Table A30

Summary Statistics After 35 Items for the θ Estimates for θ = −2 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.028	.278	.279	.011	.289	.289
1	.051	.302	.306	.017	.295	.296
2	.063	.376	.381	.032	.374	.375
3	.096	.431	.442	.060	.439	.443
4	.182	.575	.603	.174	.594	.619
WLE						
0	.020	.269	.270	.000	.285	.285
1	.034	.290	.292	.008	.301	.301
2	.068	.336	.343	.018	.361	.362
3	.093	.371	.383	.059	.414	.419
4	.197	.514	.550	.168	.584	.608
EAP						
0	.150	.264	.303	.136	.268	.301
1	.179	.276	.329	.155	.274	.315
2	.205	.315	.376	.161	.292	.333
3	.216	.311	.379	.213	.312	.377
4	.291	.378	.477	.263	.351	.439

Table A31

Summary Statistics After 35 Items for the θ Estimates for θ = −1 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.020	.303	.304	.035	.297	.299
1	.084	.314	.325	.058	.334	.339
2	.129	.349	.372	.069	.358	.364
3	.154	.397	.426	.101	.433	.445
4	.220	.490	.537	.234	.534	.583
WLE						
0	.006	.298	.298	.021	.303	.303
1	.067	.306	.313	.048	.337	.340
2	.137	.313	.342	.057	.365	.369
3	.186	.339	.386	.093	.429	.439
4	.274	.422	.503	.216	.531	.573
EAP						
0	.083	.276	.288	.096	.274	.290
1	.145	.286	.321	.119	.284	.309
2	.184	.292	.345	.127	.286	.313
3	.211	.314	.378	.159	.322	.359
4	.250	.356	.435	.255	.373	.452

Table A32

Summary Statistics After 35 Items for the θ Estimates for θ = 0 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.023	.295	.296	.031	.295	.296
1	.038	.291	.294	.085	.320	.331
2	.132	.323	.349	.131	.342	.366
3	.154	.353	.386	.221	.379	.439
4	.211	.393	.446	.265	.430	.505
WLE						
0	.010	.296	.296	.017	.299	.299
1	.030	.286	.288	.076	.321	.330
2	.086	.287	.299	.124	.343	.365
3	.145	.302	.335	.212	.379	.434
4	.240	.310	.392	.264	.429	.504
EAP						
0	.011	.264	.264	.011	.263	.263
1	.026	.259	.261	.060	.279	.286
2	.091	.272	.287	.084	.296	.308
3	.140	.285	.317	.187	.305	.358
4	.208	.289	.356	.256	.328	.417

Table A33

Summary Statistics After 35 Items for the θ Estimates for θ = 0 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.010	.300	.300	.027	.288	.289
1	−.098	.283	.300	−.030	.276	.277
2	−.311	.277	.417	−.243	.288	.377
3	−.606	.252	.656	−.520	.280	.591
4	−1.053	.206	1.074	−.965	.237	.994
WLE						
0	−.007	.301	.301	.017	.291	.291
1	−.115	.285	.307	−.043	.277	.281
2	−.332	.273	.430	−.171	.270	.320
3	−.651	.245	.696	−.374	.266	.459
4	−1.098	.194	1.116	−.695	.247	.738
EAP						
0	−.004	.275	.275	.015	.259	.259
1	−.098	.262	.280	−.031	.251	.253
2	−.225	.243	.331	−.166	.253	.303
3	−.405	.230	.466	−.333	.242	.412
4	−.597	.213	.634	−.534	.237	.585

Table A34

Summary Statistics After 35 Items for the θ Estimates for θ = 1 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.010	.284	.284	.008	.295	.295
1	−.154	.306	.342	−.089	.289	.302
2	−.465	.308	.558	−.363	.311	.478
3	−1.010	.266	1.045	−.872	.290	.919
4	−1.804	.120	1.809	−1.659	.150	1.666
WLE						
0	.003	.287	.287	.000	.292	.292
1	−.169	.304	.348	−.099	.286	.302
2	−.502	.308	.589	−.295	.296	.418
3	−1.124	.238	1.150	−.664	.284	.722
4	−1.883	.105	1.887	−1.232	.198	1.248
EAP						
0	−.079	.260	.272	−.088	.270	.284
1	−.221	.273	.351	−.172	.258	.311
2	−.424	.268	.502	−.351	.264	.439
3	−.684	.270	.736	−.620	.258	.672
4	−1.065	.211	1.086	−.966	.221	.991

Table A35

Summary Statistics After 35 Items for the θ Estimates for θ = 2 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.016	.292	.292	−.003	.285	.285
1	−.186	.316	.367	−.164	.270	.316
2	−.650	.336	.732	−.534	.316	.621
3	−1.701	.156	1.709	−1.477	.223	1.495
4	−2.744	.058	2.746	−2.567	.067	2.570
WLE						
0	.005	.296	.296	−.013	.286	.286
1	−.221	.325	.393	−.183	.274	.330
2	−.724	.317	.791	−.461	.293	.546
3	−1.872	.139	1.878	−1.070	.269	1.104
4	−2.833	.049	2.834	−2.054	.101	2.057
EAP						
0	−.159	.265	.309	−.166	.270	.317
1	−.351	.276	.446	−.328	.249	.412
2	−.645	.274	.701	−.583	.261	.639
3	−1.088	.240	1.115	−1.005	.241	1.034
4	−1.777	.141	1.783	−1.636	.164	1.645

Table A36

Summary Statistics After 35 Items for the θ Estimates for θ = 2.5 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.002	.279	.279	.008	.290	.290
1	−.195	.324	.378	−.178	.303	.352
2	−.800	.342	.870	−.627	.340	.713
3	−2.150	.119	2.154	−1.875	.164	1.883
4	−3.230	.027	3.232	−3.057	.044	3.059
WLE						
0	−.006	.280	.280	.000	.287	.287
1	−.235	.330	.406	−.200	.303	.364
2	−.917	.301	.966	−.528	.336	.626
3	−2.333	.103	2.336	−1.384	.203	1.400
4	−3.322	.022	3.324	−2.531	.070	2.533
EAP						
0	−.206	.263	.334	−.193	.270	.332
1	−.421	.284	.507	−.408	.278	.494
2	−.817	.267	.861	−.713	.269	.763
3	−1.408	.193	1.422	−1.312	.191	1.327
4	−2.215	.100	2.219	−2.067	.131	2.072

Table A37

Summary Statistics After 35 Items for the θ Estimates for θ = 3 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.016	.301	.302	.027	.283	.284
1	−.182	.337	.383	−.149	.325	.357
2	−1.004	.338	1.060	−.804	.342	.874
3	−2.621	.080	2.623	−2.315	.105	2.319
4	−3.731	.031	3.733	−3.553	.030	3.555
WLE						
0	−.004	.297	.297	.011	.277	.277
1	−.228	.341	.410	−.180	.320	.367
2	−1.202	.264	1.231	−.669	.330	.746
3	−2.810	.071	2.812	−1.773	.140	1.780
4	−3.823	.027	3.825	−3.017	.049	3.019
EAP						
0	−.232	.278	.362	−.220	.264	.344
1	−.487	.293	.568	−.452	.302	.544
2	−1.046	.261	1.079	−.941	.260	.977
3	−1.819	.138	1.825	−1.683	.138	1.689
4	−2.693	.078	2.696	−2.524	.101	2.527

Table A38

Summary Statistics After 50 Items for the θ Estimates for θ = −3 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.005	.262	.262	.001	.258	.258
1	−.012	.284	.284	−.002	.267	.267
2	−.009	.290	.290	−.006	.291	.291
3	.010	.307	.307	.012	.316	.317
4	.008	.352	.353	.002	.339	.339
WLE						
0	.010	.253	.254	.007	.253	.253
1	.000	.268	.268	.005	.258	.259
2	.004	.272	.272	−.003	.287	.287
3	.019	.285	.286	.015	.307	.308
4	.023	.317	.317	.016	.339	.340
EAP						
0	.167	.234	.288	.165	.232	.285
1	.166	.244	.296	.169	.232	.287
2	.169	.243	.296	.166	.244	.295
3	.190	.252	.316	.195	.260	.325
4	.213	.276	.349	.190	.267	.328

Table A39

Summary Statistics After 50 Items for the θ Estimates for θ = −2.5 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.001	.239	.239	.003	.230	.230
1	.002	.250	.250	.013	.253	.253
2	.039	.270	.273	.006	.266	.266
3	.020	.283	.284	.005	.283	.283
4	.040	.320	.322	.041	.337	.340
WLE						
0	.000	.236	.236	.000	.229	.229
1	−.002	.242	.243	.011	.251	.251
2	.043	.266	.269	.000	.267	.267
3	.024	.278	.279	.001	.280	.280
4	.036	.290	.293	.028	.340	.341
EAP						
0	.123	.230	.260	.121	.221	.252
1	.131	.237	.271	.138	.238	.275
2	.171	.250	.303	.136	.247	.282
3	.156	.251	.295	.145	.252	.291
4	.179	.252	.309	.174	.259	.312

Table A40

Summary Statistics After 50 Items for the θ Estimates for θ = −2 with Misfit as Correct Responses

θ Estimation and Item Misfit	FI			K-L Information		
	Bias	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.026	.237	.239	.008	.244	.244
1	.027	.249	.251	.013	.248	.248
2	.043	.282	.285	.005	.263	.263
3	.043	.274	.278	.033	.278	.280
4	.050	.318	.322	.041	.305	.308
WLE						
0	.016	.235	.236	.001	.243	.243
1	.022	.244	.245	.005	.249	.249
2	.048	.272	.276	−.002	.264	.264
3	.056	.278	.283	.022	.276	.277
4	.067	.302	.309	.034	.298	.300
EAP						
0	.118	.232	.260	.101	.231	.252
1	.133	.235	.270	.116	.233	.260
2	.149	.254	.297	.111	.239	.264
3	.156	.254	.298	.146	.253	.292
4	.168	.270	.318	.164	.259	.307

Table A41

Summary Statistics After 50 Items for the θ Estimates for θ = −1 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.017	.257	.258	.026	.257	.259
1	.064	.258	.266	.037	.273	.275
2	.082	.268	.280	.050	.274	.279
3	.082	.291	.302	.048	.301	.304
4	.103	.335	.351	.100	.342	.356
WLE						
0	.009	.256	.256	.018	.261	.261
1	.054	.251	.257	.032	.276	.278
2	.097	.257	.275	.040	.276	.279
3	.119	.276	.301	.036	.302	.305
4	.165	.319	.359	.091	.336	.348
EAP						
0	.067	.240	.249	.074	.238	.249
1	.113	.239	.264	.090	.253	.268
2	.137	.238	.274	.099	.244	.263
3	.150	.259	.299	.108	.259	.281
4	.178	.279	.332	.157	.292	.332

Table A42

Summary Statistics After 50 Items for the θ Estimates for θ = 0 with Misfit as Correct Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.017	.256	.256	.016	.253	.254
1	.029	.250	.252	.060	.272	.279
2	.090	.271	.286	.085	.283	.295
3	.093	.292	.307	.153	.301	.338
4	.125	.291	.317	.171	.329	.371
WLE						
0	.007	.257	.257	.008	.254	.254
1	.023	.247	.249	.056	.274	.280
2	.063	.255	.262	.079	.286	.297
3	.105	.261	.282	.147	.301	.335
4	.168	.265	.313	.167	.325	.365
EAP						
0	.006	.240	.240	.007	.238	.238
1	.020	.229	.230	.047	.249	.254
2	.069	.241	.251	.064	.262	.270
3	.105	.251	.272	.143	.252	.290
4	.157	.250	.295	.186	.283	.338

Table A43

Summary Statistics After 50 Items for the θ Estimates for θ = 0 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.004	.259	.259	.018	.247	.248
1	−.079	.247	.259	−.019	.247	.247
2	−.225	.251	.337	−.184	.250	.310
3	−.425	.234	.486	−.364	.247	.440
4	−.711	.221	.745	−.633	.251	.681
WLE						
0	−.005	.256	.256	.012	.247	.247
1	−.088	.249	.264	−.029	.244	.246
2	−.237	.251	.346	−.131	.239	.273
3	−.449	.235	.507	−.280	.236	.367
4	−.740	.213	.770	−.477	.243	.536
EAP						
0	−.005	.242	.242	.013	.230	.231
1	−.081	.230	.244	−.024	.226	.227
2	−.170	.225	.282	−.133	.231	.266
3	−.312	.210	.376	−.260	.222	.342
4	−.455	.211	.502	−.398	.226	.458

Table A44

Summary Statistics After 50 Items for the θ Estimates for θ = 1 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.004	.243	.243	.006	.253	.253
1	−.118	.261	.287	−.064	.251	.259
2	−.298	.265	.399	−.247	.263	.361
3	−.593	.296	.663	−.534	.284	.605
4	−1.153	.218	1.174	−1.025	.228	1.051
WLE						
0	−.004	.242	.242	.002	.251	.259
1	−.127	.263	.292	−.071	.249	.258
2	−.314	.265	.411	−.208	.252	.327
3	−.651	.294	.715	−.431	.268	.508
4	−1.235	.200	1.252	−.782	.245	.820
EAP						
0	−.063	.229	.238	−.063	.237	.245
1	−.174	.242	.298	−.131	.231	.266
2	−.305	.238	.387	−.262	.235	.352
3	−.481	.255	.544	−.444	.240	.505
4	−.733	.228	.768	−.677	.234	.716

Table A45

Summary Statistics After 50 Items for the θ Estimates for θ = 2 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.011	.249	.249	−.006	.249	.249
1	−.128	.263	.293	−.116	.243	.269
2	−.376	.288	.474	−.324	.259	.415
3	−.835	.289	.884	−.751	.292	.807
4	−1.858	.152	1.865	−1.690	.163	1.699
WLE						
0	.003	.252	.252	−.012	.250	.250
1	−.147	.267	.305	−.127	.244	.275
2	−.403	.286	.494	−.293	.253	.387
3	−.954	.273	.992	−.595	.287	.661
4	−1.986	.135	1.992	−1.208	.231	1.231
EAP						
0	−.118	.234	.262	−.130	.237	.270
1	−.252	.242	.350	−.241	.227	.332
2	−.433	.249	.500	−.396	.234	.460
3	−.687	.244	.729	−.646	.250	.693
4	−1.102	.217	1.124	−1.010	.221	1.035

Table A46

Summary Statistics After 50 Items for the θ Estimates for θ = 2.5 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.006	.241	.241	.004	.245	.245
1	−.129	.263	.293	−.123	.254	.282
2	−.415	.307	.517	−.341	.297	.453
3	−1.049	.270	1.083	−.924	.283	.967
4	−2.295	.104	2.299	−2.121	.129	2.126
WLE						
0	−.002	.240	.240	−.003	.244	.244
1	−.146	.267	.304	−.135	.254	.288
2	−.456	.305	.549	−.305	.286	.418
3	−1.223	.237	1.246	−.718	.283	.772
4	−2.432	.091	2.435	−1.553	.177	1.564
EAP						
0	−.148	.229	.272	−.151	.236	.280
1	−.294	.248	.384	−.135	.239	.378
2	−.519	.261	.581	−.305	.258	.531
3	−.861	.246	.896	−.804	.240	.839
4	−1.388	.184	1.401	−1.293	.195	1.308

Table A47

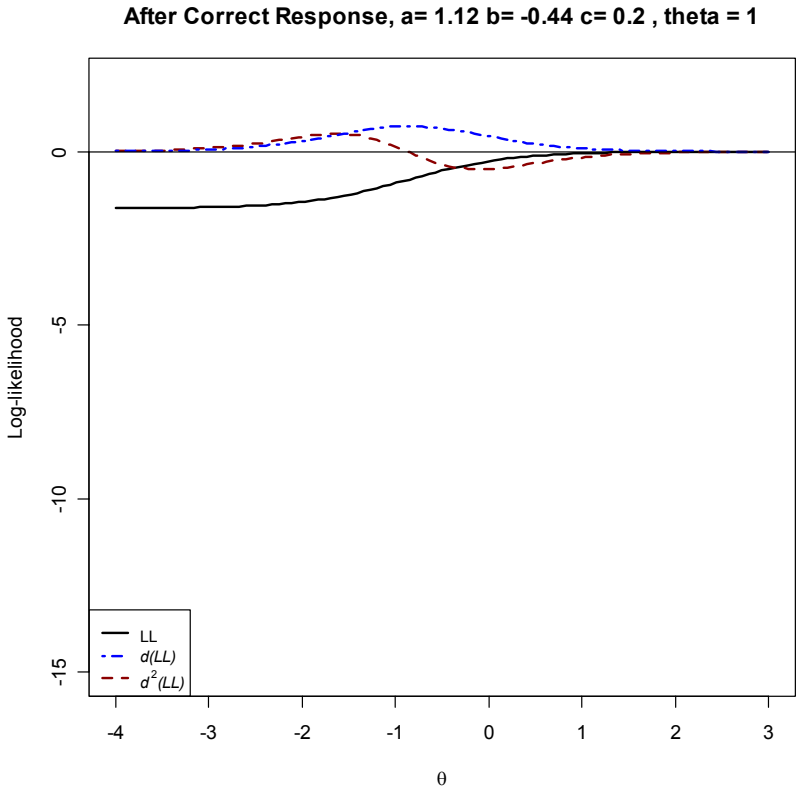
Summary Statistics After 50 Items for the θ Estimates for θ = 3 with Misfit as Incorrect Responses

θ Estimation	FI			K-L Information		
and Item Misfit	*Bias*	*SE*	*RMSE*	*Bias*	*SE*	*RMSE*
MLE						
0	.011	.260	.260	.020	.248	.249
1	−.117	.272	.296	−.096	.258	.275
2	−.440	.325	.547	−.389	.307	.495
3	−1.366	.224	1.385	−1.154	.267	1.185
4	−2.777	.086	2.780	−2.578	.095	2.581
WLE						
0	−.001	.258	.258	.006	.243	.243
1	−.138	.271	.304	−.111	.256	.279
2	−.499	.330	.599	−.347	.294	.455
3	−1.577	.195	1.590	−.862	.288	.909
4	−2.919	.077	2.922	−1.961	.129	1.966
EAP						
0	−.170	.241	.295	−.160	.232	.282
1	−.321	.252	.408	−.301	.256	.395
2	−.603	.264	.659	−.565	.263	.624
3	−1.089	.237	1.115	−.996	.237	1.025
4	−1.790	.145	1.797	−1.634	.178	1.646

Figure A1

LL and First and Second Derivatives of the LL for the First 12 Items in the CAT for the 3-Item MCR Condition

a.



b.

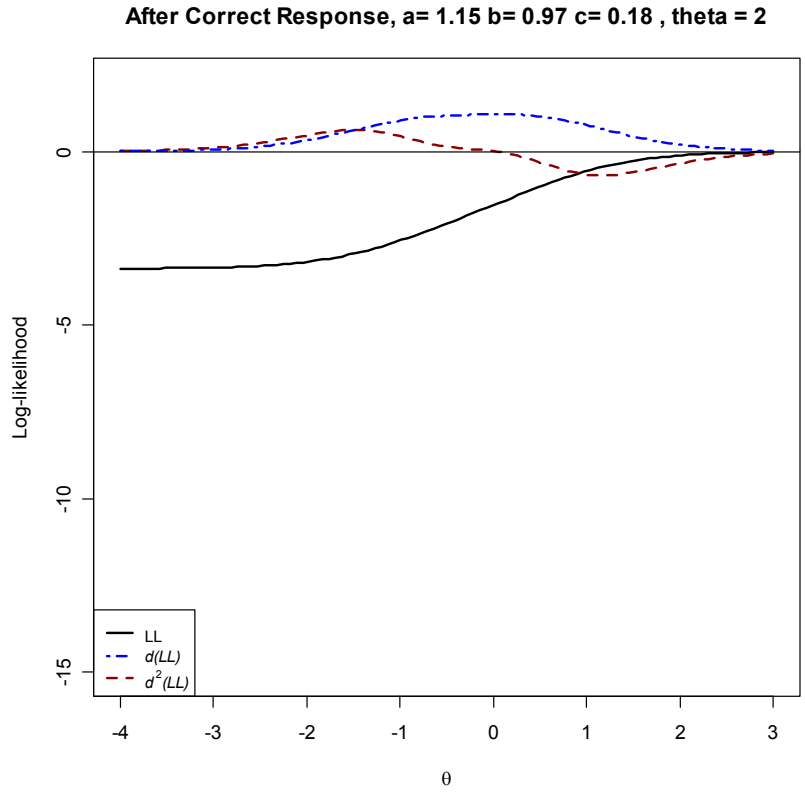
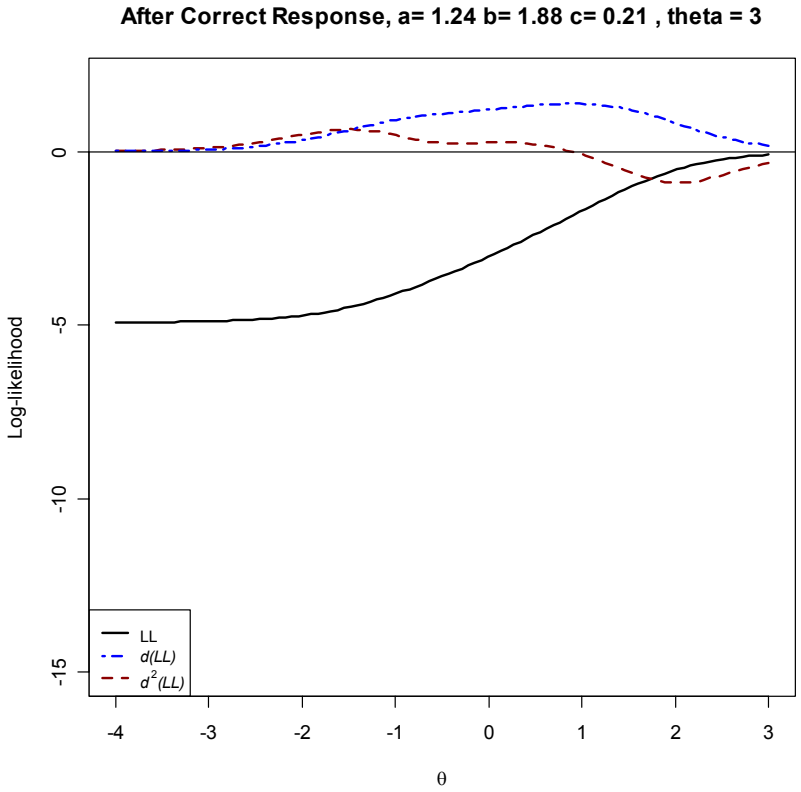


Figure A1, cont.

LL and First and Second Derivatives of the LL for the First 12 Items in the CAT for the 3 Misfit-as-Correct-Responses Condition

c.



d.

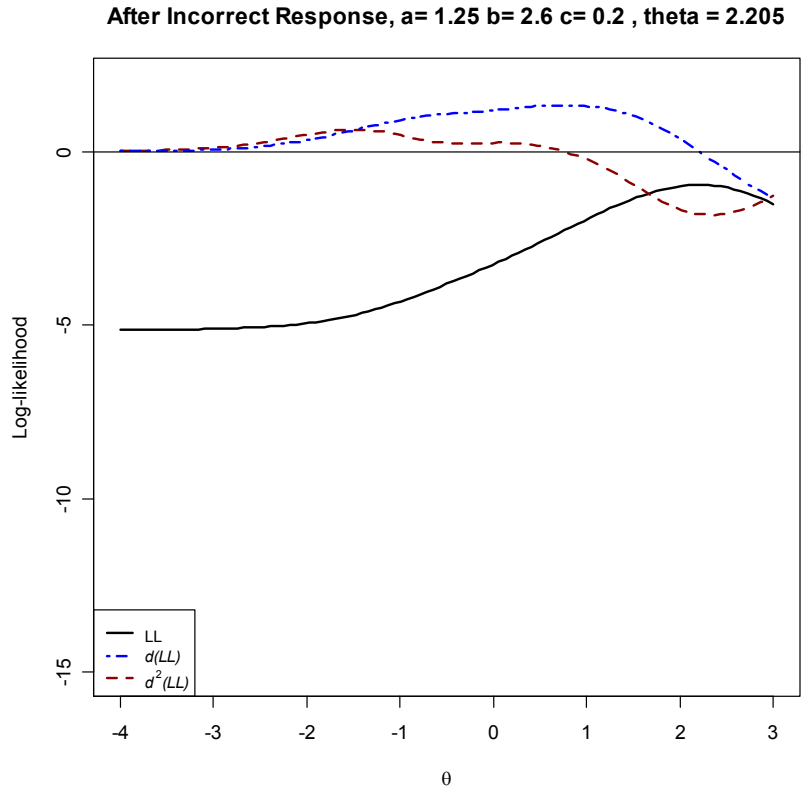
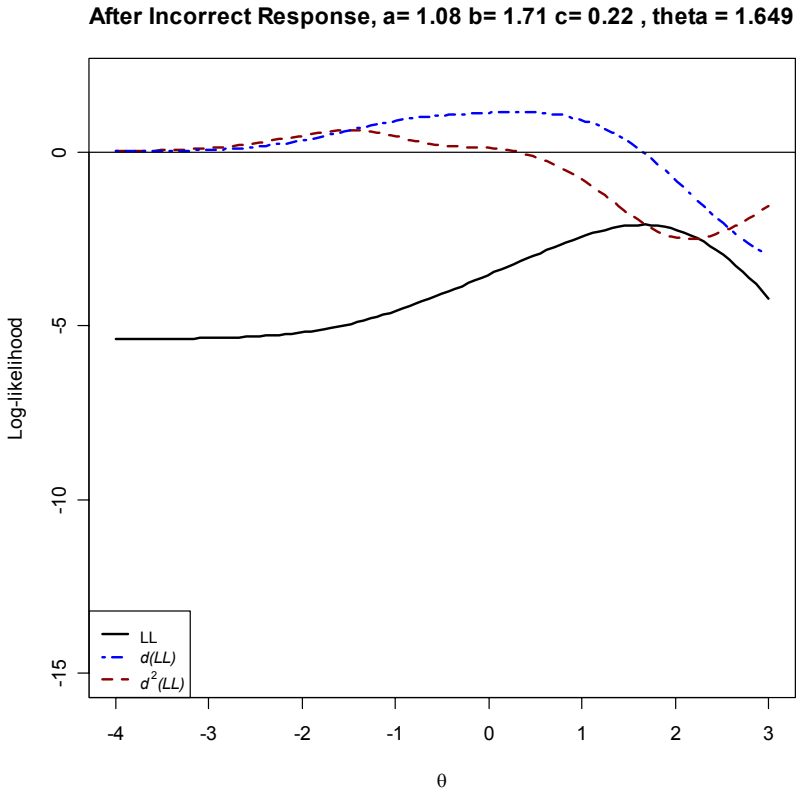


Figure A1, cont.

LL and First and Second Derivatives of the LL for the First 12 Items in the CAT for the 3 Misfit-as-Correct-Responses Condition

e.



f.

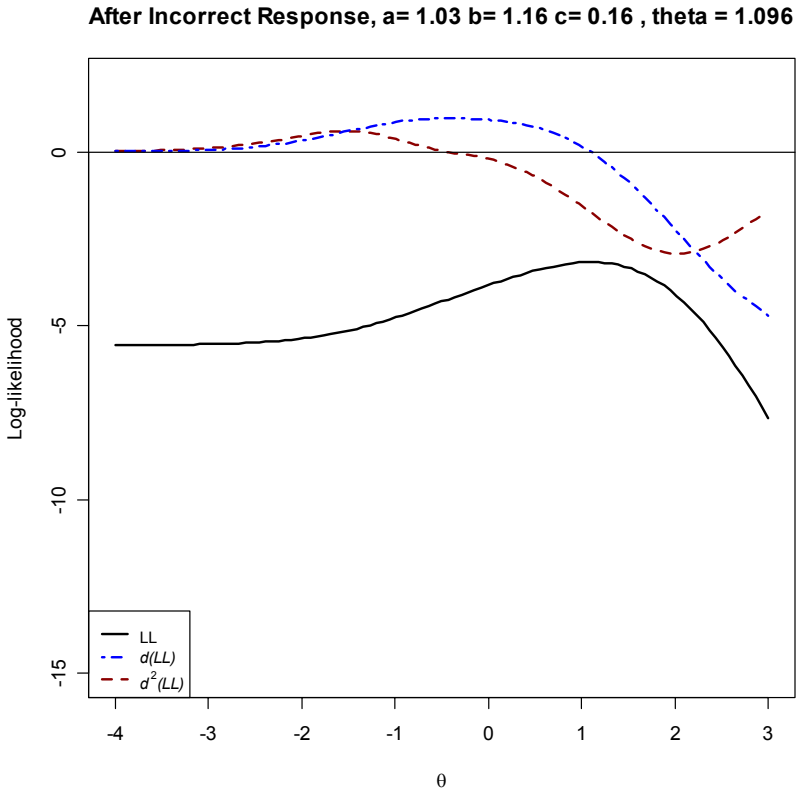
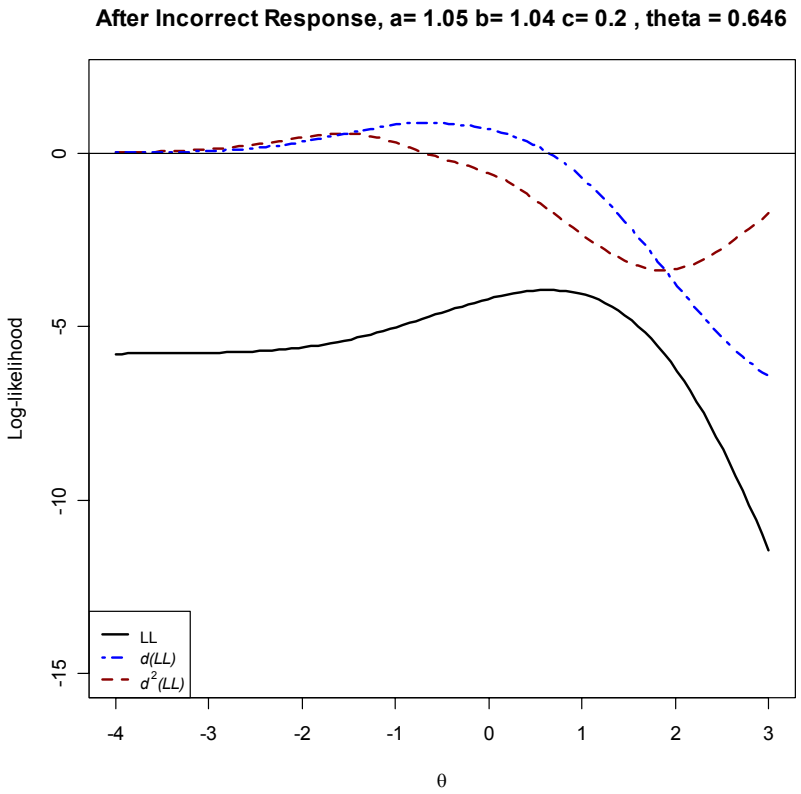


Figure A1, cont.

LL and First and Second Derivatives of the LL for the First 12 Items in the CAT for the 3 Misfit-as-Correct-Responses Condition

g.



h.

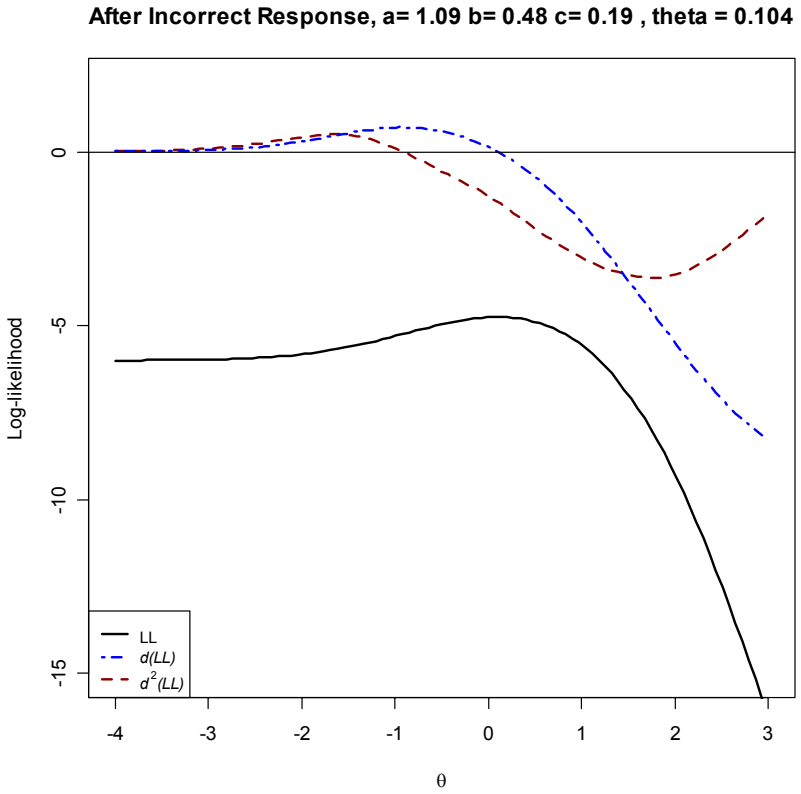
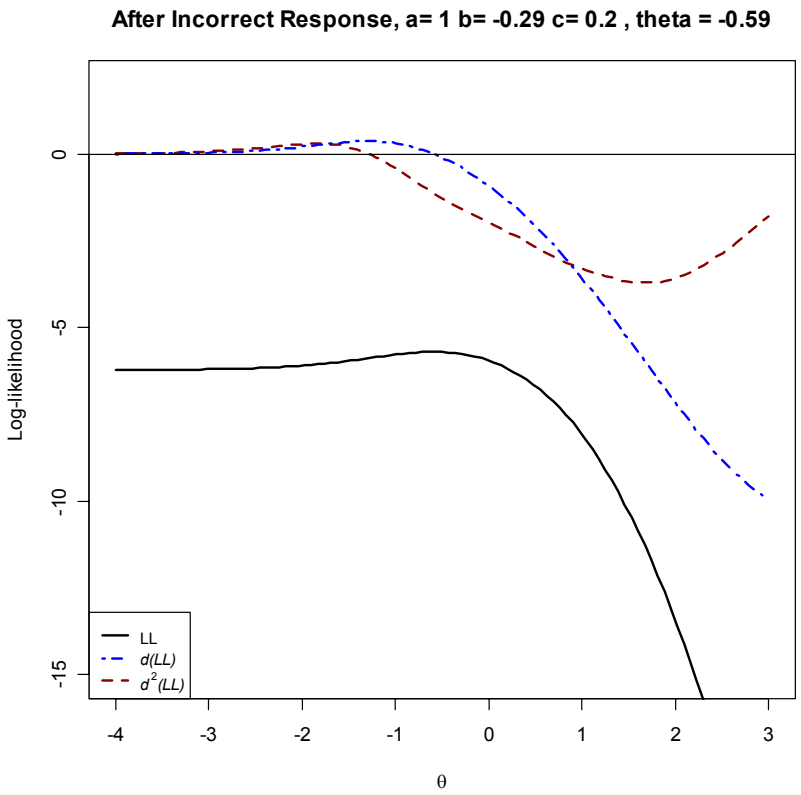


Figure A1, cont.

LL and First and Second Derivatives of the LL for the First 12 Items in the CAT for the 3 Misfit-as-Correct-Responses Condition

i.



j.

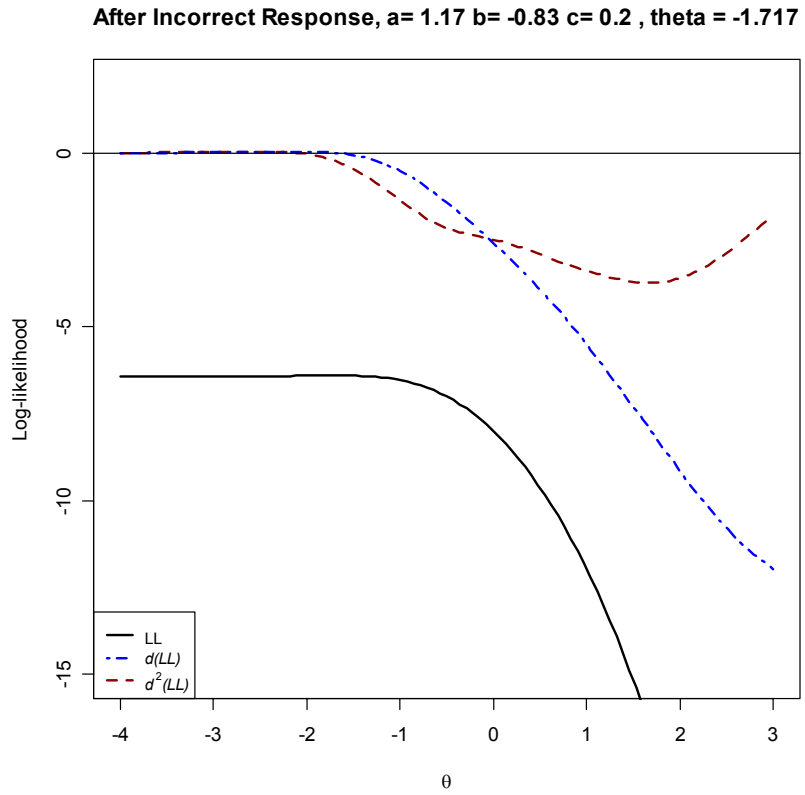
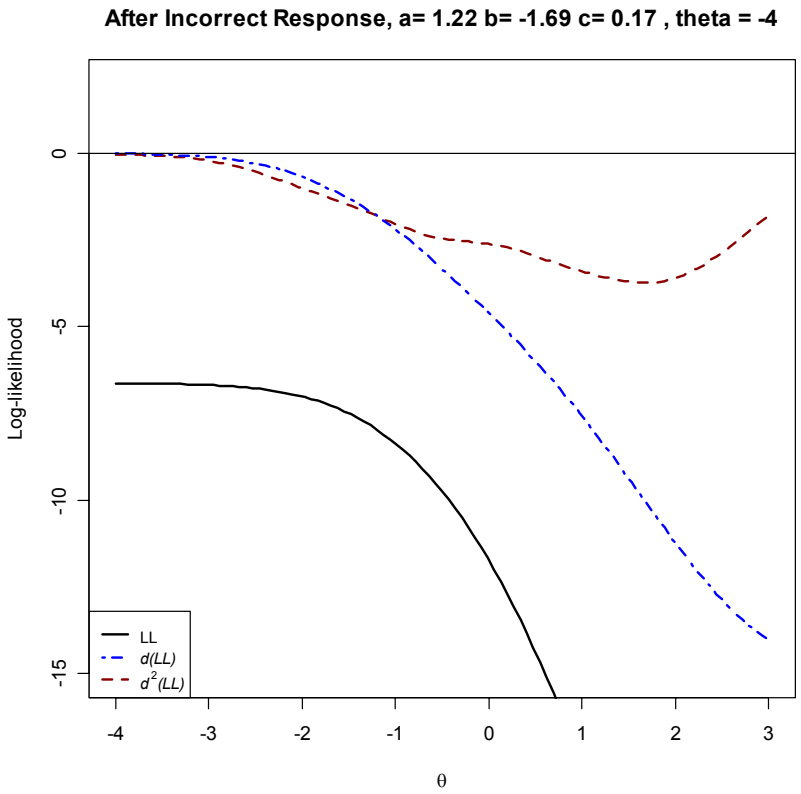


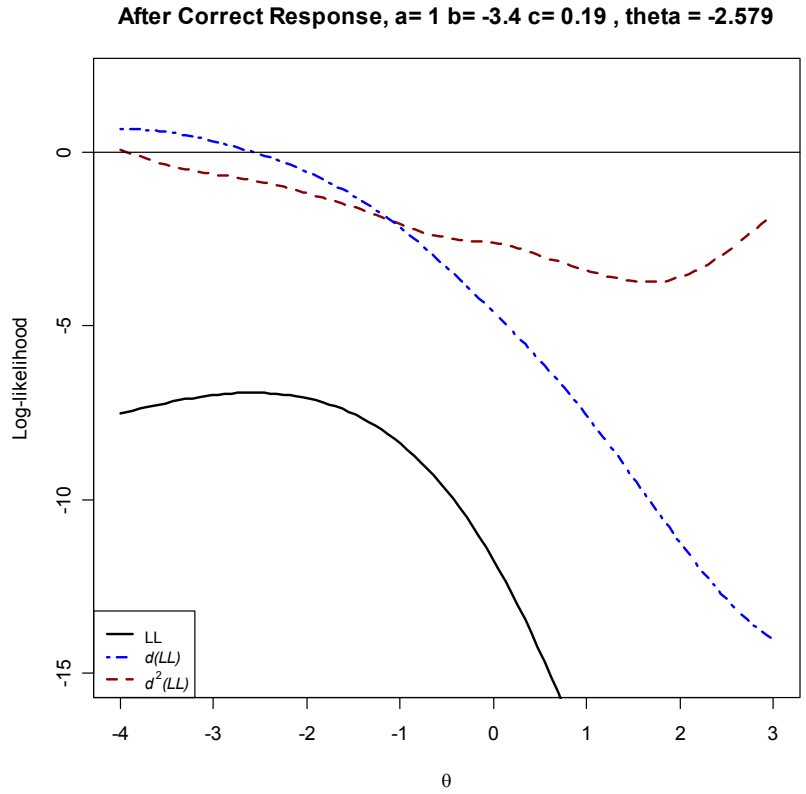
Figure A1, cont.

LL and First and Second Derivatives of the LL for the First 12 Items in the CAT for the 3 Misfit-as-Correct-Responses Condition

k.



l.



Formulas for the ANOVA Sums of Squares and Degrees of Freedom

To reduce the number of formulas required for definition of all of the interaction *SS* terms in the model, the procedure defined by Howell (2007) was used. Howell recommended calculation of the total cell variance for a given effect, then subtracting the lower order *SS* from the cell variance. As the misfitting items condition was within-subjects, the model assumed that the same simulee participated in all five conditions. Thus, as there were 300,000 replications, a total of 60,000 unique simulees (*S*) were modeled by the mixed-design ANOVA. For purposes of the formulas to be presented below, the following notation was used:

T = a given θ condition,

E = a given θ estimation method,

I = a given item selection method,

M = a given misfitting item condition,

Bet = total between subjects variability in the model,

WS = total within subjects variability in the model,

$\overline{X}$ = mean of all observations.

Sums of Squares

$$SS_{total} = \sum_{j=1}^{300,000} (X_j - \overline{X})^2 \qquad \text{(A1)}$$

$$SS_{Bet} = \sum_{s=1}^{60,000}(\overline{X}_S - \overline{X})^2 \quad \text{(A2)}$$

$$SS_{WS} = SS_{total} - SS_{Bet} \quad \text{(A3)}$$

$$SS_T = 30,000\sum_{c=1}^{10}(\overline{X}_T - \overline{X})^2 \quad \text{(A4)}$$

$$SS_E = 100,000\sum_{c=1}^{3}(\overline{X}_E - \overline{X})^2 \quad \text{(A5)}$$

$$SS_I = 150,000\sum_{c=1}^{2}(\overline{X}_I - \overline{X})^2 \quad \text{(A6)}$$

$$SS_M = 60,000\sum_{c=1}^{5}(\overline{X}_M - \overline{X})^2 \quad \text{(A7)}$$

$$SS_{Cells(TE)} = 10,000\sum_{c=1}^{30}(\overline{X}_{TE} - \overline{X})^2 \quad \text{(A8)}$$

$$SS_{TE} = SS_{Cells(TE)} - SS_T - SS_E \quad \text{(A9)}$$

$$SS_{Cells(TI)} = 15,000\sum_{c=1}^{20}(\overline{X}_{TI} - \overline{X})^2 \quad \text{(A10)}$$

$$SS_{TI} = SS_{Cells(TI)} - SS_T - SS_I \quad \text{(A11)}$$

$$SS_{Cells(TM)} = 6,000\sum_{c=1}^{50}(\overline{X}_{TM} - \overline{X})^2 \quad \text{(A11)}$$

$$SS_{TM} = SS_{Cells(TM)} - SS_T - SS_M \quad \text{(A12)}$$

$$SS_{Cells(EI)} = 50,000\sum_{c=1}^{6}(\overline{X}_{EI} - \overline{X})^2 \quad \text{(A13)}$$

$$SS_{EI} = SS_{Cells(EI)} - SS_E - SS_I \quad \text{(A14)}$$

$$SS_{Cells(EM)} = 20{,}000\sum_{c=1}^{15}(\overline{X}_{EM} - \overline{X})^2 \quad \text{(A15)}$$

$$SS_{EM} = SS_{Cells(EM)} - SS_E - SS_M \quad \text{(A16)}$$

$$SS_{Cells(IM)} = 30{,}000\sum_{c=1}^{10}(\overline{X}_{IM} - \overline{X})^2 \quad \text{(A17)}$$

$$SS_{IM} = SS_{Cells(IM)} - SS_I - SS_M \quad \text{(A18)}$$

$$SS_{Cells(TEI)} = 5{,}000\sum_{c=1}^{60}(\overline{X}_{TEI} - \overline{X})^2 \quad \text{(A19)}$$

$$SS_{TEI} = SS_{Cells(TEI)} - SS_T - SS_E - SS_I - SS_{TE} - SS_{TI} - SS_{EI} \quad \text{(A20)}$$

$$SS_{Cells(TEM)} = 2{,}000\sum_{c=1}^{150}(\overline{X}_{TEM} - \overline{X})^2 \quad \text{(A21)}$$

$$SS_{TEM} = SS_{Cells(TEM)} - SS_T - SS_E - SS_M - SS_{TE} - SS_{TM} - SS_{EM} \quad \text{(A22)}$$

$$SS_{Cells(TIM)} = 3{,}000\sum_{c=1}^{100}(\overline{X}_{TIM} - \overline{X})^2 \quad \text{(A23)}$$

$$SS_{TIM} = SS_{Cells(TIM)} - SS_T - SS_I - SS_M - SS_{TI} - SS_{TM} - SS_{IM} \quad \text{(A24)}$$

$$SS_{Cells(EIM)} = 10{,}000\sum_{c=1}^{30}(\overline{X}_{EIM} - \overline{X})^2 \quad \text{(A25)}$$

$$SS_{EIM} = SS_{Cells(EIM)} - SS_E - SS_I - SS_M - SS_{EI} - SS_{EM} - SS_{IM} \quad \text{(A26)}$$

$$SS_{Cells(TIME)} = 1{,}000\sum_{c=1}^{300}(\overline{X}_{TIME} - \overline{X})^2 \quad \text{(A27)}$$

$$SS_{TIME} = SS_{Cells(TIME)} - SS_T - SS_I - SS_M - SS_E - SS_{TE} - SS_{TI} - SS_{TM} - SS_{EI} - SS_{EM} - SS_{IM} - SS_{TEI} - SS_{TEM} - SS_{TIM} - SS_{EIM} \quad \text{(A28)}$$

$$SS_{Error(Bet)} = SS_{Bet} - SS_T - SS_I - SS_E - SS_{TI} - SS_{TE} - SS_{EI} - SS_{TEI} \quad \text{(A29)}$$

$$SS_{Error(WS)} = SS_{WS} - SS_M - SS_{TM} - SS_{IM} - SS_{EM} - SS_{TEM} - SS_{TIM} - SS_{EIM} - SS_{TIME} \quad \text{(A30)}$$

Degrees of Freedom

The following terms are defined for the *df* reported below:

t = total number of conditions for θ,

e = total number of conditions for θ estimation,

i= total number of conditions for item selection,

m = total number of conditions for item misfit,

U = independent number of simulees in the model (60,000),

N = total number of observations in the model (300,000).

$$df_T = (t-1) \quad \text{(A31)}$$

$$df_E = (e-1) \quad \text{(A32)}$$

$$df_I = (i-1) \quad \text{(A33)}$$

$$df_M = (m-1) \quad \text{(A34)}$$

$$df_{TE} = (t-1)(e-1) \quad \text{(A35)}$$

$$df_{TI} = (t-1)(i-1) \quad \text{(A36)}$$

$$df_{TM} = (t-1)(m-1) \quad \text{(A37)}$$

$$df_{EI} = (e-1)(i-1) \quad \text{(A38)}$$

$$df_{EM} = (e-1)(m-1) \quad \text{(A39)}$$

$$df_{IM} = (i-1)(m-1) \quad \text{(A40)}$$

$$df_{TEI} = (t-1)(e-1)(i-1) \quad \text{(A41)}$$

$$df_{TEM} = (t-1)(e-1)(m-1) \quad \text{(A42)}$$

$$df_{TIM} = (t-1)(i-1)(m-1) \quad \text{(A43)}$$

$$df_{EIM} = (e-1)(i-1)(m-1) \quad \text{(A44)}$$

$$df_{TIME} = (t-1)(i-1)(m-1)(e-1) \quad \text{(A45)}$$

$$df_{Error(Bet)} = U - (t \times i \times e) \quad \text{(A46)}$$

$$df_{Error(WS)} = N - U - [t \times i \times e \times (m-1)] \quad \text{(A47)}$$