Differential Item Functioning in Computerized Adaptive Testing: Can CAT Self-Adjust Enough?

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Chayut Piromsombat

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ernest C. Davenport, Jr., Adviser

April, 2014

Acknowledgements

Working on this dissertation was one of the most difficult yet happiest times in my life. It made me realize why most dissertations have at least one page devoted to thanking people. I could not complete my dissertation without help and support from many people around me. First, I would like to express my deepest gratitude to my advisor, Professor Ernest Davenport, with my whole heart. He guided me with his dedication, patience, and intellectual knowledge throughout my doctoral study. His continuous support and encouragement were the key factors that enabled me to complete my degree.

My appreciation extends to the examining committee members, Professor David Weiss, Professor Mark Davison, and Professor Robert delMas. Their comments and suggestions have significantly improved the quality of this dissertation. Their genuine willingness to join the committee in summer, with short notice, will be forever imprinted on my heart. A special thank you goes to Professor Weiss who suggested that I apply for a grant from the Applied Psychological Measurement Inc. to partially support my study.

Having to finish my preliminary and final oral examinations within a restricted time period would be impossible without help from several staff members at the University of Minnesota. A special thank you goes out to all the helpful staff, especially Kathleen Walter, Stacia Madsen, Peggy Ferdinand, and Sharon Sawyer. Finally, I would like to thank the Institute for the Promotion of Teaching Science and Technology (IPST) of Thailand for their financial support throughout my study, and the Office of Educational Affairs Royal Thai Embassy in Washington DC for taking such good care of me during my long journey in the States.

Dedication

To my parents, Suchet and Lamai Piromsombat, for their endless love and support.
To all Thai taxpayers for the scholarship throughout my study.

Abstract

Two issues related to differential item functioning (DIF) in the context of computerized adaptive testing (CAT) were addressed in this study: 1) the effect of DIF in operational items on the accuracy of the ability estimate ($\hat{\theta}_{CAT}$) and 2) the accuracy of detecting DIF in pretest items when DIF occurred in operational items and examinees were matched on the number-correct score (NCS), the ability estimate obtained from nonadaptive computer-based testing ($\hat{\theta}_{CBT}$), and $\hat{\theta}_{CAT}$. To investigate the first issue, a series of simulations were conducted by varying the level of DIF magnitude (0, .4, 1, and 1.6); DIF type (uniform and nonuniform); DIF contamination or the number of DIF items (6, 15, and 24 items out of the 30-item test); and DIF occurrence (first, middle, last, and across stages of CAT). For the latter issue, test impact ($\mu_R - \mu_F$ = 0 and 1) and sample size ratio (N_R:N_F = 1:1 and 9:1) were also added to the simulation.

It was found in the first simulation that CAT could adjust for the effect of DIF in operational items if DIF occurred in the early stages of CAT, with some restrictions though. Specifically, CAT successfully adjusted for the effect of DIF at the earlier stages if the number of DIF items and the magnitude of DIF were moderate. In other situations, CAT seemed to reduce the effect of DIF as seen in the trend of SEs which increased when DIF items were delivered and decreased after CAT administered a new DIF-free item. However, the self-adjustment of CAT was not enough to recover $\hat{\theta}_{CAT}$ from DIF effects.

The results from another simulation suggested that matching examinees on $\hat{\theta}_{CAT}$ did not provide impressive advantages over the NCS and $\hat{\theta}_{CBT}$ in most of the simulation conditions. Overall, when operational items were contaminated with moderate DIF magnitude, the three matching variables yielded comparable results of DIF detection in pretest items. However, when the level of DIF contamination in operational items increased, matching examinees on $\hat{\theta}_{CAT}$ led to the worst situation of detecting DIF in pretest items, especially when large-uniform DIF items were used in the operational test. It was also evident that DIF in operational items, especially CAT items, led to false identification of DIF type. Specifically, pretest items exhibiting uniform DIF were mistakenly identified as nonuniform DIF if the matching variable was obtained from nonuniform-DIF operational items.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Differential item functioning (DIF) occurs when examinees of equal ability, but belonging to different groups, have unequal probabilities to succeed on the same item. In practice, DIF conventionally refers to the difference of item parameters between groups of examinees (Holland & Thayer, 1988; Raju, 1990; Swaminathan & Rogers, 1990; Embretson & Reise, 2000; Herrera & Gómez, 2008). If an item exhibits DIF due to factors irrelevant to the target ability, the item is said to be biased. In other words, DIF is a statistical signal of potential item bias. Hence, it is recommended to screen biased items by first detecting DIF in pretest items, and then conducting a sensitivity review of the pretest items flagged as DIF. After the review, only high quality pretest items should be formed into an operational test and delivered to examinees.

Recently, the choice of test delivery methods has shifted from paper-and-pencil testing (PPT) to computer-based testing (CBT) and computerized adaptive testing (CAT). These delivery methods are different in several ways, for instance, the test mode (paper versus computer), the underlying test theory (classical test theory versus item response theory), and the strategy used to assign items to examinees. Because of these differences, several researchers have addressed concerns about detecting DIF in items delivered by different methods (Zwick, Thayer, & Wingersky, 1994a; Zwick, 2010; Nandakumar & Roussos, 2001).

Based on the item administration, a test can be adaptive or nonadaptive. Typically, PPT and CBT are nonadaptive because they deliver the same set of operational items (fixed-length and fixed-order) to all examinees. In contrast, CAT gives an examinee items that best match the examinee's ability estimate obtained from previous items. Thus, different examinees may receive different sets of items, yielding different total scores across examinees. Hence, matching examinees on total score as done in traditional DIF detection (e.g. the Mantel-Haenszel statistic, SIBTEST, and logistic regression) is meaningless in the context of CAT (see a more comprehensive discussion on this issue in Zwick, 2002). As a result, detecting DIF in CAT is more challenging than in PPT and CBT where examinees can be matched directly on their total scores or

ability estimates because all examinees receive the same test items.

There are several studies on DIF detection in the context of CAT (e.g., Zwick, Thayer, & Wingersky, 1994a; Nandakumar & Roussos, 2001; and Lei, Chen, & Yu, 2006). However, previous studies primarily focused on the application of traditional DIF detection methods used in PPT to detecting DIF in pretest items for CAT. Typically, pretest items are delivered to examinees as a nonadaptive test for item bank development purposes. In the literature, there is a lack of studies that specifically investigate DIF in operational items during the CAT process. Therefore, the primary purpose of the present study is to expand the understanding of DIF effects in such a context.

1.1 Problem Statement

Detecting DIF items in CAT is more challenging than that in PPT and CBT for several reasons (Lei, Chen, & Yu, 2006; Zwick, 2010). First, there tends to be fewer items administered in CAT than in PPT and CBT. As a result, each item may have more impact on the estimation of an examinee's ability. Second, examinees usually receive different sets of operational items in CAT, yielding different possible total scores across examinees. Even if two examinees have the same total score (i.e., they receive the same number of items from CAT), the total score for each examinee has a different meaning because of the different set of items underlying the score. Hence, number-correct scores are inappropriate to serve as the matching variable. Finally, there are some potential sources of DIF that might particularly occur in CAT, including examinees' familiarity with computer usage for testing and anxiety due to computerized testing (Bringsjord, 2001; Wang, Jiao, Young, Brooks, & Olson, 2007, 2008). Unlike gender and ethnicity, these potential sources are rarely considered in DIF studies.

Several researchers proposed statistical methods for detecting DIF in CAT applications. For example, Zwick and colleagues (Zwick, Thayer, & Wingersky, 1994a; Zwick, Thayer, & Lewis, 1997) proposed Zwick-Thayer-Wingersky (ZTW) and Zwick-Thayer-Lewis (ZTL) methods based on the Mantel-Haenszel statistic (Holland & Thayer, 1988) and the standardization method (Dorans & Kulick, 1986). Nandakumar and Roussos (2001) extended the SIBTEST procedure (Shealy & Stout, 1993) to the context

of CAT, yielding the CATSIB procedure. Finally, Lei, Chen, and Yu (2006) proposed two methods based on logistic regression (Swaminathan & Rogers, 1990) and the Item Response Theory (IRT)-based likelihood ratio test (Thissen, Steinberg, & Wainer, 1988), called CAT-LR and CAT-IRTLR.

Although purportedly developed to detect DIF in CAT, the methods listed above do not detect DIF in operational CAT items or during the CAT process. In fact, they detect DIF in pretest items which are typically administered as a nonadaptive test. Basically, these methods consist of two key steps. First, examinees from the reference and focal groups are matched on the IRT-based ability estimate obtained from CAT ($\hat{\theta}_{CAT}$). This means that detecting DIF starts only after CAT is terminated, not during the CAT process. Next, the pretest items are examined by applying traditional detection methods to the responses of matched examinees.

Some questions arise from the fact that the currently available methods for detecting DIF in CAT detect DIF in nonadaptive pretest items by matching examinees on $\hat{\theta}_{CAT}$. For instance, is it worth using $\hat{\theta}_{CAT}$ as the matching variable for detecting DIF in nonadaptive pretest items? Compared to the number-correct score (NCS) and the IRT-based ability estimate ($\hat{\theta}$) obtained from conventional or nonadaptive testing (either PPT or CBT), does $\hat{\theta}_{CAT}$ provide more accurate results of DIF detection? These questions have not been answered in previous studies.

To date, statistical methods for detecting DIF in operational items during the CAT process are not yet available. Two theoretical issues probably make detecting DIF in such cases either unnecessary or complicated. First, the item selection algorithm in CAT may alleviate the effect of DIF in CAT items. Hypothetically, the effect of DIF items selected by CAT in proceeding stages may be reduced by DIF-free items administered in subsequent stages. If that is the case, detecting DIF in operational items during the CAT process may be unnecessary.

Another issue is that different examinees tend to receive different items at the same stage of CAT, depending on their current ability estimates. For example, in a particular stage of CAT, an item with difficulty level of 1.5 is more likely to be delivered to examinees whose ability estimates at the current stage of CAT are about 1.5. It is rare to

find such examinees coming from both reference and focal groups with a sufficient sample size for each group, even in large-scale CAT administrations (Zwick, 2010), because examinees are unlikely to have the same item at the same stage. As a result, responses for detecting DIF in an operational CAT item in each stage are usually insufficient. This limitation makes DIF detection in CAT, if needed, even more complex because most available detection methods require a sufficient number of responses from each group to detect DIF in an item.

The issues of DIF effects and detection methods in the context of CAT are both important as discussed above. However, understanding the nature of DIF effects in CAT and whether CAT can adjust for the effect of DIF is presumably a good start for developing DIF detection methods for CAT. Hence, the primary focus of this study was to investigate the effect of DIF in operational CAT items on the ability estimate and how CAT can adjust for the effect of DIF.

In typical CAT, the ability estimate of an examinee is updated after the examinee answers each item. CAT uses this updated estimate to select the next item for the examinee. Although many item selection algorithms are available, they basically select the item that maximizes information at the examinee's ability estimate. Typically, an item with a high discrimination index and a difficulty level around the ability estimate is chosen as the subsequent item.

If CAT selects a DIF item for a focal group examinee, the ability estimate for this examinee tends to be decreased because the DIF item usually appears more difficult to the examinee. That is, the examinee is likely to answer the item incorrectly in this stage. As a result, CAT will select an easier item as the next item. If the subsequent item does not exhibit DIF, the examinee will have a higher chance to answer the new item correctly. The reason is that CAT selects the new item based on the biased ability estimate which is typically lower than the unbiased ability estimate. In other words, the item difficulty of the new item is less than the examinee's unbiased ability estimate. Figure 1 illustrates the just described scenario.

In sum, if the subsequent item is DIF-free, it may adjust for bias in the ability estimate obtained from the prior DIF item. That is, the adaptive nature and the item selection algorithm in CAT can, in theory, reduce the effect of DIF in operational items

on examinees' ability estimates. If the mechanism of CAT can effectively reduce DIF effects, the development of new statistical methods specifically for detecting DIF in operational CAT items may be unnecessary. However, there are also some cases that the effect of DIF may remain.

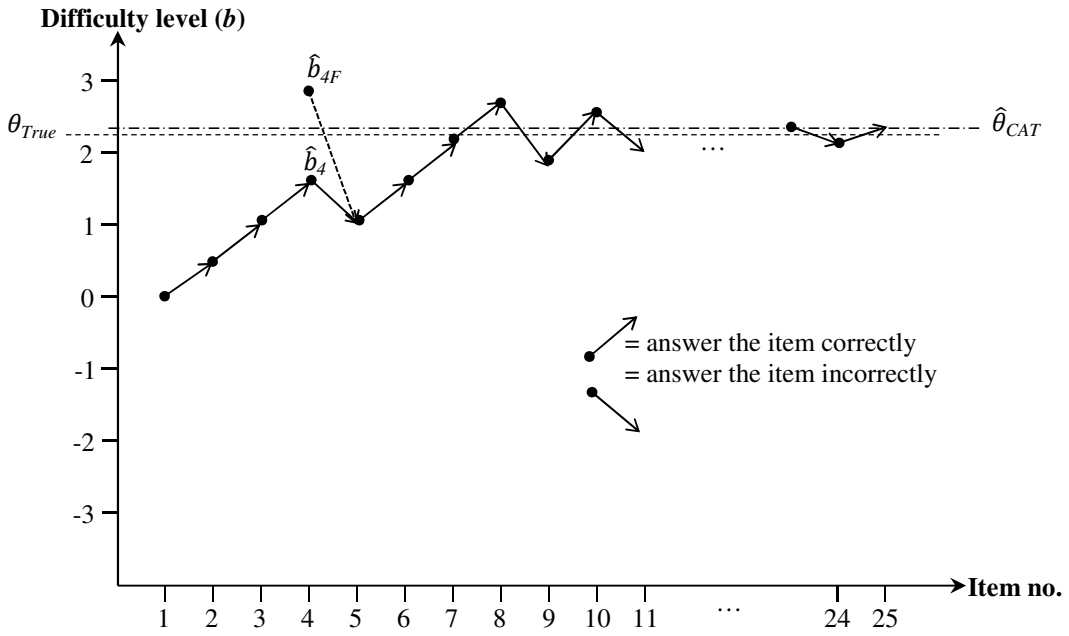


Figure 1. A hypothetical illustration of how CAT can reduce the effect of a DIF item (Item no. 4). At the end of CAT, the IRT-based ability estimate ($\hat{\theta}_{CAT}$) can converge to the true ability level (θ_{True}).

The possible cases where CAT may fail to adjust for the effect of DIF include: (1) all items in the item bank are DIF items, (2) DIF items are selected in the first stages and the remaining operational items are not enough to correct the biased estimate (e.g., a short fixed-length CAT), or (3) DIF items are selected in the last stages. If CAT cannot reduce DIF effects in such cases, statistical methods for detecting DIF in operational CAT items are indeed needed. Previous studies have not investigated the effect of DIF on $\hat{\theta}_{CAT}$ and if the item selection algorithm can alleviate the effect. Therefore, the present study aimed to address this issue to get a better understanding of DIF in CAT.

1.2 Review of Related Literature

1.2.1 Components of CAT and Their Implementation

The popularity of CAT has increased in recent years due not only to the advancement in computer technology, but also the advantages of CAT itself. One of the advantages is that at each stage of CAT, an item is selected based on an examinee's performance obtained from previous stages. Specifically, by matching between the item's difficulty and the examinee's current ability estimate, CAT can achieve an acceptably accurate ability estimate with fewer items and less testing time (Rudner, 1998). This feature of CAT makes it more efficient and precise than its counterparts.

However, no advantage can be obtained unless five basic components of CAT are carefully designed. These components are the item bank, starting point, item selection algorithm, ability estimation, and termination rule (Thompson & Weiss, 2011). After assuring that CAT is needed and resources for CAT development are available, test developers in a testing program need to thoroughly consider each CAT component before an operational CAT can be launched.

Item bank. Every CAT needs at least one item bank consisting of a sufficient number of precalibrated items. In practice, an item bank is usually built based on IRT, rather than CTT, because item and examinee parameters can be matched on the same scale. Generally, after items are collected (newly written or obtained from existing tests), item parameters are estimated for the pilot sample using the selected IRT model. In this step, it is also recommended to examine model-data fit and dimensionality of the items. Next, only the items that satisfy some prespecified psychometric criteria will be added into the bank. For example, if the item bank is developed for classification purposes, items that maximize test information around the cutscore will be retained. On the other hand, if the item bank is built for general assessment purposes, high discriminating items with a wide range of difficulty levels are needed for the bank to obtain high test information across ability levels (Thompson & Weiss, 2011).

Starting point ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{0}}$). At the beginning of CAT, a starting ability estimate for each examinee is required in order to select the first item and continue the CAT process. There are several options for $\hat{\theta}_0$ (Guyer, 2008; Thompson & Weiss, 2011) such as (1) a

fixed value at the population mean, (2) random values drawn from a small range of ability levels, (3) examinees' ability estimate from previous tests, and (4) predicted values based on other information related to the target ability. Each option has its advantages and disadvantages. For example, assigning a fixed value to all examinees may lead to the situation that CAT selects the same first item for every examinee. Unless the item bank has various items that match with the fixed value, this option will overexpose some items. On the other hand, using external information to assign $\hat{\theta}_0$ does not guarantee the accuracy of $\hat{\theta}_0$ because the information may be biased. In sum, the choice of $\hat{\theta}_0$ depends on the testing situation (e.g., size of item bank, nature of the test) and whether related information is available to reasonably assign a specific $\hat{\theta}_0$ to the examinee.

Item selection algorithm. Several algorithms have been developed for selecting the "best" subsequent item for each examinee. Fundamentally, the algorithms first evaluate some type of "information" obtained from each item if the item is administered. Next, the item that provides the maximum information at the examinee's ability estimate is selected and delivered to the examinee in the next stage. There are various types of information proposed for item selection purposes, including Fisher information, Kullback-Leibler information, Fisher information interval, and likelihood weighted information (see details in Guyer, 2008; van der Linden & Pashley, 2010). Although some algorithms provide advantages over others, previous studies (Chang & Ying, 1996; Chen, Ankenmann, & Cheng, 2000; Chen & Ankenmann, 2004; Guyer, 2008) consistently reported that there appeared to be no precision advantage for any of the algorithms after 10-15 items are administered in CAT.

Ability estimation. For IRT-based CAT, several ability estimation methods are available, including maximum likelihood estimation (MLE), weighted maximum likelihood (WLE), Bayesian modal a posteriori (MAP), and Bayesian expected a posteriori estimation (EAP). van der Linden and Pashley (2010) reviewed previous studies on the precision of these estimation methods. The conclusion is very similar to the comparison of item selection algorithms; that is, the difference in precision of different ability estimation methods is only severe for short tests (10 items). For longer

tests (20 items or more), the ability estimation methods are apparently able to recover from a biased start and provide similar ability estimates at the end of CAT.

In practice, MLE seems to be the most popular choice because it is relatively less biased and widely available in general IRT software such as BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996, as cited in du Toit, 2003); Xcalibre 4.1 (Guyer & Thompson, 2011); the R packages *ltm* (Rizopoulos, 2006), irtoys (Partchev, 2011), *MiscPsycho* (Doran, 2010); ScoreAll 4 (Guyer, 2010); and some CAT-specific software such as FastTEST Pro 2 (Weiss, 2008), CATSim (Weiss & Guyer, 2010), and SimulCAT (Han, 2011). However, MLE cannot provide a solution if responses have a non-mixed pattern. This response pattern occurs when only the first item is administered or the examinee is scored 0 or 1 for all administered items. The test developer has three options to solve this problem: (1) switch to one of the Bayesian estimation methods, (2) assign an arbitrary ability estimate until the response pattern is non-mixed, or (3) force CAT to select an item with extreme difficulty level for the next stage, i.e., a very difficult item for the all-1 response pattern or a very easy item for the all-0 pattern (Guyer, 2008). In practice, the last solution seems to be favorable because it is less computationally intensive and most likely to provide a mixed response pattern.

Termination rule. After the initial stages of CAT, one or more termination rules should be evaluated to check if enough information is obtained from administered items and to assure that no items are unnecessarily administered. Typically, CAT can stop for one or a combination of the following conditions: (1) maximum test length is reached, (2) the precision of the ability estimate is satisfied (i.e., the error of the ability estimate is acceptably minimized), (3) the accuracy of examinee classification is satisfied, or (4) the desired test information is obtained (Linacre, 2000; Thompson & Weiss, 2011). Similar to the other components, a choice of termination rule also depends on the testing situation. For example, if the primary goal of the testing program is to classify examinees based on their performance on CAT items, the accuracy of classification results obtained from CAT is important and it should be included in the final termination rule.

In addition to the five basic components above, there are other issues that the CAT developer needs to consider, including content balancing, item exposure, and enemy

items (Weiss & Guyer, 2010). Briefly, the idea of content balancing is to deliver items assessing different content domains to each examinee in the desired proportions based on the test specification of content domains. For item exposure, the idea is to assure that CAT will not select some specific items (e.g., an item with a high discrimination which usually provides the maximum item information) too often. Finally, enemy items refer to items that provide examinees clues for answering other items. These items should not be delivered to the same examinee. In practice, CAT can take these issues into account by adding additional constraints in the item selection algorithm.

1.2.2 Previous Studies on DIF in the Context of CAT

As discussed earlier, several researchers proposed statistical methods for detecting DIF in the context of CAT. Among them, ZTW and ZTL were the first attempts (Lei, Chen, & Yu, 2006; Zwick, 2010). The ZTW and ZTL methods were basically the Mantel-Haenszel statistic based on the IRT-based ability estimate obtained from CAT ($\hat{\theta}_{CAT}$) rather than the number-correct score (NCS). Specifically, both ZTW and ZTL divide the continuum of $\hat{\theta}_{CAT}$ into multiple intervals. Next, examinees from the reference and focal groups are matched if their $\hat{\theta}_{CAT}$ fall into the same interval. Finally, the conventional Mantel-Haenszel statistic can be applied to evaluate DIF for each item. The key difference between ZTW and ZTL is that the ZTL method uses an empirical Bayes estimation to provide more stable results for small-sample cases. Nandakumar and Roussos (2001) proposed a modified version of the SIBTEST, called CATSIB, for detecting DIF in CAT items. In CATSIB, examinees are matched on multiple intervals of the regression-corrected version of $\hat{\theta}_{CAT}$. After matching examinees on these intervals, a typical SIBTEST procedure is applied to assess DIF.

An important limitation of ZTW, ZTL, and CATSIB is that these methods were developed solely for detecting uniform DIF. To overcome this limitation, Lei, Chen, and Yu (2006) proposed CAT-LR and CAT-IRTLR (modified versions of logistic regression and IRT-based likelihood-ratio test, respectively) for detecting nonuniform DIF in CAT contexts. The CAT-LR procedure replaces the total score with $\hat{\theta}_{CAT}$ in the logistic regression equation. For CAT-IRTLR, the application is slightly complicated due to the

fact that the original version of the likelihood ratio test requires the full data matrix in order to use joint maximum likelihood estimation, but CAT usually yields incomplete or sparse data. Therefore, CAT-IRTLR needs to impute responses of unadministered items before applying the traditional likelihood ratio test to detect DIF items in CAT. As originally proposed, CAT-IRTLR implemented the missing response imputation based on the probabilities of getting the items correct, computed from the three-parameter logistic model using the known item parameters and CAT ability estimates.

It should be emphasized again that the ZTW, ZTL, CATSIB, CAT-LR, and CAT-IRTLR procedures only detect DIF in pretest items by matching examinees on the ability estimate obtained from CAT. These methods were not developed for detecting DIF in operational items during the CAT process. Nevertheless, several studies revealed that DIF detection methods listed above were effective in detecting DIF and useful for CAT development.

Statistical properties of DIF detection methods in the context of CAT were investigated in several studies. For instance, Zwick, Thayer, and Wingersky (1994a) evaluated statistical properties of the ZTW method by simulating responses to three 75-item banks (i.e., Bank 1 had no DIF items, Bank 2 had DIF items that were correlated to item difficulty, and Bank 3 had DIF items that were uncorrelated with item difficulty). Using a CAT system based on the item information, each examinee was assigned 25 items from one of the three banks. Examinees from the reference and focal groups were matched on the expected true score on the entire 75-item bank, which was estimated from the expected score on the 25 administered items and the estimated item parameters. The performance of ZTW was compared with that of the traditional Mantel-Haenszel statistic which matched examinees on the summed score from non-adaptive testing administration. The results showed that DIF estimates obtained from the ZTW method were highly correlated to the true DIF magnitudes. In addition, Type I error rates of the ZTW method were acceptable in most of the simulation conditions.

Using similar simulation factors, Zwick and colleagues (1994b) conducted a simulation study to investigate the performance of the ZTW method in detecting DIF in 15 pretest CAT items. The pretest items were all generated to have uniform DIF, and were assigned to all simulated examinees. The researchers compared the results of the

ZTW method when using different matching criteria: ability estimates obtained from existing items in the bank (pre-calibrated CAT items), and ability estimates obtained from the pretest items (non-calibrated CAT items). The results showed that matching examinees on the pretest items led to inflated Type I error rates. In addition, the same researchers (1995) investigated the effect of model-data misfit on the performance of ZTW. In the study, item parameters were generated using a three-parameter logistic model, but estimated with a one-parameter logistic model. As expected, the accuracy of the ZTW method was significantly reduced by the model-data misfit.

For the ZTL method, Zwick, Thayer, and Lewis (1997) conducted a simulation to investigate the validity of the ZTL method by manipulating sample sizes, ability distributions, and underlying IRT models. It should be noted that this simulation did not manipulate any CAT-related factors. Zwick and Thayer (2002) later investigated the performance of the ZTL method in the context of CAT by simulating data using various CAT item banks, sample sizes, test lengths, magnitudes of uniform DIF, and magnitudes of test impact (i.e., group mean difference on the target/true ability). The results showed that the ZTL method was effective in detecting uniform DIF in pretest items, especially in small-sample cases.

There were some questionable findings from the simulations discussed above. For example, results from the simulations by Zwick, Thayer, and Wingersky (1994a, 1994b) revealed that matching examinees on the total score worked as well as the new matching procedure (i.e., matching on $\hat{\theta}_{CAT}$ intervals). Moreover, the standard errors obtained for the ZTW method were much larger than those obtained for the traditional Mantel-Haenszel statistic (Zwick, Thayer, & Wingersky, 1994a, 1994b; Zwick, 1997), even when the IRT model used for the ZTW method fit the data. These findings were suspicious because, in theory, $\hat{\theta}_{CAT}$ should be a more accurate estimate of an examinee's ability than the total score when the model fits the data well. Therefore, the ZTW method should work better and yield smaller standard errors than the standard Mantel-Haenszel statistic in the context of CAT.

Regarding the performance of CATSIB, Nandakumar and Roussos (2001; 2004) conducted simulation studies imitating a pretest scenario of CAT settings. In their

simulations, 25 items from a bank of 1,000 well-calibrated items were adaptively delivered to each examinee, and 16 pretest items (i.e., items with unknown parameters) were conventionally delivered to all examinees. The pretest items were assessed for DIF using the CATSIB procedure. The results showed that CATSIB with the regression correction yielded satisfactory Type I error rates; however, its power ranged widely from .17 to 1. In addition, CATSIB was less sensitive when (1) the magnitude of test impact was large; (2) the sample size of each group was small or about 250 examinees; and (3) the group sample sizes were unequal. In 2006, Roussos, Nandakumar, and Banks proposed a formula for adjusting the bias in CATSIB due to discretization of the ability scale. These researchers also proposed a Kernel-smoothed version of CATSIB for detecting DIF in small-sample CAT cases (Nandakumar, Banks, & Roussos, 2006).

The performance of CATSIB was also compared to that of SIBTEST in the simulation study by Walker, Beretvas, and Ackerman (2001). The researchers concluded that CATSIB and SIBTEST were comparable in terms of their power, but Type I error rates of CATSIB tended to be higher than those of SIBTEST. Although these researchers intended to compare the performance of CATSIB with that of SIBTEST, their findings might not be generalized as such. One of the limitations in their study was that the data were generated for a fixed-length nonadaptive test instead of an adaptive test. Matching examinees on the ability estimate obtained from such a test was indeed to match examinees on either NCS or $\hat{\theta}$, not $\hat{\theta}_{CAT}$. Consequently, the study actually examined the accuracy of SIBTEST when using NCS and $\hat{\theta}$ obtained from a nonadaptive test as the matching variable.

There were also other limitations of CATSIB simulations discussed above. For example, in the simulation by Nandakumar and Roussos (2001), the true item parameters used in data generation were also used in all computations of the CATSIB procedures. That is, examinees' $\hat{\theta}_{CAT}$ and the regression correction were computed using the true item parameters, not the estimated item parameters. This computation approach seemed to be unrealistic because item parameters are usually unknown in practice. Another limitation was found in the study of Walker and colleagues (2001). Although the authors claimed to investigate the performance of CATSIB in CAT, they did not simulate data in

accordance with the key components of adaptive testing. That is, factors related to the CAT environment (e.g., size of item bank, item selection algorithm, and termination rule) were not manipulated. The simulation results may not be widely generalizable.

In the literature of DIF detection in the context of CAT, only the simulation by Lei, Chen, and Yu (2006) investigated the performance of CATSIB, CAT-LR, and CAT-IRTLR procedures simultaneously. This simulation was conducted to mimic a CAT environment in which pretest items were assumed to be seeded in the bank but were not used to estimate $\hat{\theta}_{CAT}$. Data were simulated under the conditions of sample size ratios and magnitudes of test impact. According to this simulation, CATSIB, CAT-LR, and CAT-IRTLR were comparable in terms of their power in detecting uniform DIF. However, CAT-LR and CAT-IRTLR were more powerful than SIBTEST in detecting nonuniform DIF. In addition, CAT-IRTLR provided Type I error rates about .05 under several conditions (e.g., unequal sample sizes and large test impact). Although the simulation results suggested that the CAT-IRTLR procedure was favorable in detecting DIF pretest items, the generalizability of this simulation may be limited because the imputation procedure used in this study was conducted using true item parameters (a similar limitation found in the CATSIB simulations discussed above).

1.3 Research Purposes, Questions, and Hypotheses

Based on the literature review above, there are two major limitations of previous studies on DIF detection in the context of CAT. First, the focus of these studies was on how to detect DIF in pretest items during the development of the CAT item bank. None of them focused on the effect of DIF in operational items that can occur during the actual CAT administration on $\hat{\theta}_{CAT}$. The first purpose of the present study was therefore to address this issue by investigating the effect of DIF in operational items on the accuracy of $\hat{\theta}_{CAT}$ under various scenarios. Four factors were expected to affect the recovery of $\hat{\theta}_{CAT}$: (1) DIF contamination or number of DIF items in the operational test; (2) DIF occurrence or stages of CAT that DIF occur, i.e., whether DIF occurs early, middle, late, or randomly throughout the test; (3) type of DIF; and (4) magnitude of DIF.

As stated by van der Linden and Pashley (2010) and illustrated in Figure 1 above,

CAT may effectively recover from biased estimates obtained in the early stages of CAT. Hence, this study expected that CAT could adjust for the effect of DIF in operational items if DIF occurred in the first stages of CAT. However, CAT may not always alleviate DIF effects. Specifically, based on the results from studies on the effect of DIF in the context of PPT (e.g., Roznowski & Reith, 1999; Li, 2009; Li & Zumbo, 2009), it has been shown that bias in the ability estimate tended to increase when (1) DIF contamination increased, (2) uniform and nonuniform DIF occurred simultaneously, and (3) magnitude of DIF increased. In such cases, CAT was not expected to effectively adjust for the effect of DIF even if DIF occurred at early stages of CAT. Therefore, it was also hypothesized that the four factors listed above interactively affected the recovery of the ability estimate from DIF items in CAT. This research question and its expected results were examined in Study 1.

Another limitation found in the previous studies is that most studies proposed an extension of traditional detection methods to detect DIF in pretest items by matching examinees on $\hat{\theta}_{CAT}$. The pretest items, however, are usually administered as a nonadaptive test in the precalibration process of CAT. Hence, the CAT-based extensions proposed in previous studies still detect DIF in nonadaptive items like the original versions do, but use $\hat{\theta}_{CAT}$ as the matching variable rather than NCS as traditionally used in the original versions. The question is: to detect DIF in nonadaptive items, does matching examinees on $\hat{\theta}_{CAT}$ provide more accurate results than matching examinees on NCS and $\hat{\theta}$ obtained from a conventional or nonadaptive test (either PPT or CBT)?

As reported in Zwick, Thayer, and Wingersky (1994a, 1994b) and Walker, Beretvas, and Ackerman (2001), matching examinees on NCS, $\hat{\theta}$, and $\hat{\theta}_{CAT}$ provided similar DIF detection results. However, it should be noted that these studies assumed that all ability estimates (NCS, $\hat{\theta}$, and $\hat{\theta}_{CAT}$) were not affected by DIF in operational items (i.e., no DIF was manipulated in operational items). Assuming such scenarios, it is not surprising that the three types of matching yielded similar results of DIF detection. In real CAT settings, however, operational items may be contaminated by DIF, resulting in a biased ability estimate which, in turn, decreases the accuracy of DIF detection in pretest items.

Hence, Study 2 was conducted to expand the results of the previous studies to a more realistic CAT scenario, i.e., DIF in operational items may occur and affect DIF detection in pretest items. The second purpose of this research, then, was to investigate the accuracy of DIF detection in pretest items when operational CAT items exhibited DIF. Specifically, Study 2 compared the results of DIF detection in pretest items when NCS, $\hat{\theta}$, and $\hat{\theta}_{CAT}$ serve as the matching variable. In this substudy, the factors used in Study 1 and two additional factors (test impact and sample size ratio) were manipulated to generate testing scenarios when DIF occurred in operational items of CAT.

For Study 2, it was expected that when CAT can effectively adjust for the effect of DIF in operational items (as hypothesized in Study 1), $\hat{\theta}_{CAT}$ should be a more accurate estimate of examinees' ability than NCS and $\hat{\theta}$. Consequently, detecting DIF in pretest items using $\hat{\theta}_{CAT}$ as the matching variable was expected to yield more accurate results (i.e., provide higher detection power and lower Type I error rates) than NCS and $\hat{\theta}$. However, this expected result may not be true if DIF contamination increased, magnitude of DIF increased, and/or both types of DIF occurred at the same stage of CAT. As discussed above, such conditions can increase bias in the ability estimate obtained from both adaptive and nonadaptive tests. Therefore, the performance of DIF detection regardless of matching variable type was expected to be less accurate in such conditions. Regarding the effect of test impact and sample size ratio, it was hypothesized that larger test impact and unbalanced sample size would increase Type I error rates and decrease power of DIF detection as shown in previous studies in the context of PPT (e.g., Kennedy, 1994; Fidalgo, Mellenbergh, and Muñiz, 2000; Herrera & Gómez, 2008).

Chapter 2: Research Methods

This dissertation consists of two substudies corresponding to the two research purposes addressed in the previous section. Specifically, Study 1 was designed to examine the effect of DIF in operational items on the IRT-based ability estimate obtained from CAT ($\hat{\theta}_{CAT}$). Study 2 was designed to compare the accuracy of DIF detection in pretest items using three types of matching variable: the number-correct score (NCS) and the IRT-based ability estimate ($\hat{\theta}$) obtained from nonadaptive testing (either PPT or CBT) and $\hat{\theta}_{CAT}$ when operational items also exhibited DIF. Both substudies implemented a simulation approach; therefore, the main concern was to design the simulation to realistically represent various CAT scenarios so that the results can be generalized to typical CAT in practice. Keeping this in mind, the present study carefully designed CAT and manipulated simulation factors as described in the following sections.

2.1 Study 1: The effect of DIF in operational items on $\hat{\theta}_{CAT}$

CAT administration. In this simulation study, CAT administered 30 operational items to each examinee. As shown in previous studies, ability estimates obtained from different item selection algorithms and ability estimation methods tended to converge for a test of 10 operational items or more (Chang & Ying, 1996; Chen, Ankermann, & Cheng, 2000; Chen & Ankermann, 2004; Guyer, 2008; van der Linden & Pashley, 2010). Hence, a test of 30 operational items was chosen to control for bias in the ability estimate due to the CAT administration. In addition, this choice of test length was similar to that used in previous studies which ranged from 15 to 30 items (e.g., Zwick, Thayer, & Wingersky, 1994; Feng, 2003; Nandakumar & Roussos, 2004; Lei, Chen, & Yu, 2006).

Regarding the ability estimation, CAT updated the examinee's ability estimate ($\hat{\theta}_{CAT}$) based on the current response pattern. Specifically, if the pattern was mixed correct and incorrect responses, $\hat{\theta}_{CAT}$ was updated using maximum likelihood estimation. For the non-mixed response pattern and the mixed pattern with multiple maxima of likelihood (i.e., MLE could not converge to a unique ability estimate), the new $\hat{\theta}_{CAT}$ was obtained by adjusting the current estimate with an appropriate constant, i.e., –.5 for an

incorrect response and +.5 for a correct response of the current item. The final $\hat{\theta}_{CAT}$ for each examinee after responding to 30 items was limited in the range of –3.5 to 3.5.

To begin the CAT process, the starting point ($\hat{\theta}_0$) for each examinee was randomly drawn from U(–.5, .5). CAT then selected the first item that maximized the Fisher information function at $\hat{\theta}_0$. Next, the examinee's $\hat{\theta}_{CAT}$ was re-estimated using the just described estimation rule. The updated $\hat{\theta}_{CAT}$ was then used to select the next item such that the Fisher information function was maximized at this new estimate. These steps of selecting item and updating $\hat{\theta}_{CAT}$ were repeated until CAT administered 30 items to each examinee.

As discussed in the literature review, the combination of CAT components described above, given the length of CAT, not only provided accurate ability estimates, but also represented CAT systems found in real testing programs and previous simulation studies (Chang & Ying, 1996; Chen, Ankenmann, & Cheng, 2000; Chen & Ankenmann, 2004; van der Linden & Pashley, 2010). Therefore, the effect of CAT on the ability estimate in this study should be minimized to allow the effect of DIF in operational items to be clearly observed.

Item bank. The simulated item bank consisted of 500 dichotomous items assuming a three-parameter logistic model. This bank size was a compromise among the bank sizes used in previous simulation studies on DIF detection in CAT, for example, 75 items in Zwick, Thayer, and Wingersky (1994a; 1994b); 360 items in Lei, Chen, and Yu (2006); 700 items in Nandakumar, Banks, and Roussos (2006); and 1,000 items in Nandakumar and Roussos (2001).

The mathematical equation of a three-parameter logistic model is defined by (Embretson & Reise, 2000):

$$\mathrm{P}\left(X_{ij}=1\middle|\theta_j,a_i,b_i,c_i\right)=c_i+(1-c_i)\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]} \quad (1)$$

where X_{ij} = 1 denotes a correct response made by the j^{th} examinee on the i^{th} item; θ_j denotes the latent ability of the examinee; a_i, b_i, and c_i represent the item discrimination, item difficulty, and lower asymptote/pseudo-guessing parameters of the item; and D is a scaling constant (1.7 in this study) which makes the logistic form equivalent to

the normal ogive form of the model.

For the 500 items in the bank of this study, their item discrimination (a), item difficulty (b), and pseudo-guessing (c) were generated from the following distributions: $a \sim \ln N(-.1, .1)$; $b \sim U(-3.6, 3.6)$; and $c \sim U(0, .3)$. In the logistic metric, the mean and standard deviation of a were .909 and .091, respectively. The uniform distribution from –3.6 to 3.6 was chosen for b to cover a wide range of true ability levels, resulting in a comparable test information function across ability levels. These distributions also represented the wide ranges of item parameters found in practice and in previous CAT simulations (Chang & Ying, 1996; Wang & Vispoel, 1998; Chen, Ankenmann, & Chang, 2000; Feng, 2003; Guyer, 2008). Moreover, the distributions yielded an item bank with high information functions across the ability continuum (Figure 2), which are typically used for general testing purposes (Thompson & Weiss, 2011). A complete list of the generated item parameters for the 500-item bank is provided in Appendix A.

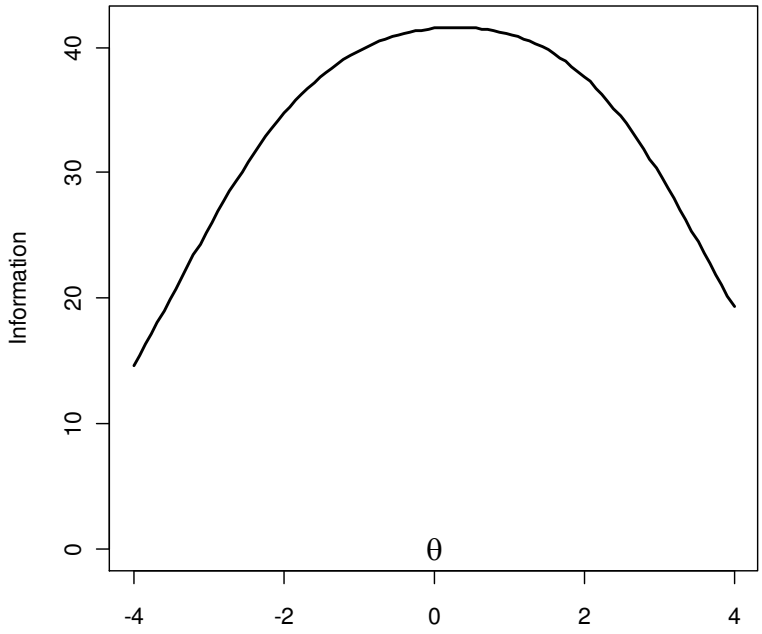


Figure 2. The test information function of 500 items in the simulated bank.

Manipulated factors. Four DIF-related factors including DIF contamination, DIF occurrence, type of DIF, and magnitude of DIF were manipulated in this study. The details of each factor are described as follows.

DIF contamination. As observed by Budgell, Raju, and Quartetti (1995), the

percentage of DIF items in real tests ranged from approximately 1% to 70%. In this study, the percentage of DIF items observed in real tests was divided using three cutoff levels: 20%, 50%, and 80% to represent low, moderate, and high DIF contamination in a test. That is, for each simulated 30-item CAT, there were 6, 15, or 24 DIF items which favored the reference group.

DIF occurrence or *DIF location.* For each operational test, DIF items were delivered at first stages, middle stages, last stages, or randomly distributed across stages of CAT. The general steps to administer DIF items during the process of CAT were as follows: (1) selected an item that maximized the Fisher information function at the current $\hat{\theta}_{CAT}$; (2) embedded DIF into the item by adjusting the item parameters in accordance with the condition of DIF type and magnitude; (3) generated 0/1 response of the item for the examinee (Note: the details of DIF manipulation and data generation are described in later sections); and (4) updated $\hat{\theta}_{CAT}$ and used it to select the next item. To administer six DIF items at the first stages of CAT, for example, steps 1-4 above were repeatedly applied to Items 1-6, while, only steps 1, 3, and 4 were repeatedly applied to Items 7-30.

Type and Magnitude of DIF. In this study, the magnitude of DIF was simply defined as the difference in item parameters between the reference and focal groups. As observed in previous studies that examined DIF in nonadaptive items, various ranges of DIF magnitude were reported, for example, .03 to .80 in Raju (1990); .10 to .76 in Walstad and Robson (1997); .06 to .83 in Gratia (1997); and .01 to .77 in Bao, Dayton, and Hendrickson (2009). In addition to these ranges, DIF magnitude could sometimes be very large or greater than 1. For instance, Raju (1988) reexamined DIF items reported in the study by Linn, Levine, Hastings, and Wardrop (1981), and found that some items exhibited very large DIF. The item discrimination of one item, for instance, was 1.8 for the reference group and .5 for the focal group, so the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = |1.8 - .5| = 1.3$. In addition, the item difficulty of the same item was 3.5 for the reference group and 5.0 for the focal group, so the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = |3.5 - 5| = 1.5$. As Raju reported, the signed and unsigned area indices for this item were 1.2 and 1.4, respectively. Based on either simple difference in

item parameters across groups or Raju's area indices, this item exhibited both nonuniform and uniform DIF with very large magnitudes. Such extreme magnitudes of nonuniform and uniform DIF might be uncommon; however, they really occurred in practice (see other examples in Raju, 1990; and Hidalgo & López-Pina, 2004).

Unlike DIF in nonadaptive items, neither type nor magnitude of DIF in operational CAT items has been reported due to the lack of statistical methods for specifically detecting DIF in such contexts. In order to investigate the effect of DIF in operational CAT items on $\hat{\theta}_{CAT}$, this study manipulated two types of DIF (nonuniform and uniform) and three levels of DIF magnitude (.4, 1, and 1.6). Such DIF magnitudes were chosen to represent medium, large, and very large DIF. Instead of using a small DIF magnitude such as .01 or .2, the three magnitudes were used to maximize the effect of DIF and provide a better observation of the effect.

DIF manipulation. Previous simulations on DIF in nonadaptive items (e.g., Swaminathan & Rogers, 1990; Zwick, Thayer, & Wingersky, 1994a; Lei, Chen, & Yu, 2006) usually manipulated DIF type and magnitude by first generating the item parameters for the reference group, and then adjusting these parameters with the desired magnitude and type of DIF to obtain the item parameters for the focal group. These generated item parameters were then used to compute the probability of a correct answer and generate 0/1 responses for the corresponding groups. Unfortunately, these steps cannot be applied to manipulating DIF in operational CAT items.

In CAT, item parameters are used to (1) select the subsequent item and (2) compute the probability of a correct answer, and then generate the responses. As discussed earlier, DIF in operational CAT items occurs if CAT selects the subsequent item using the item parameters obtained from the precalibration based on data from both reference and focal groups; but the item actually exhibits DIF (i.e., its parameters are different for each group of examinees). Typically, DIF items favor the reference group. If the subsequent item shows uniform DIF, this item will appear to be easier for the reference group but more difficult for the focal group. If the subsequent item shows nonuniform DIF, on the other hand, its discriminating power in the focal group will be smaller than in the reference group.

As a result, this study embedded DIF in operational CAT items by generating three

sets of item parameters. First, the initial item parameters were generated as they were obtained from the precalibration process. These initial item parameters, saved in the item bank, were used in the item selection algorithm. Second, the item parameters of the same item but adjusted for the desired type and magnitude of DIF in favor of the reference group were generated and used to compute the probability of a correct answer and generate the 0/1 response for the reference examinees. Finally, the item parameters of the same item but adjusted for the desired type and magnitude of DIF against the focal group were generated and used to compute the probability of a correct answer and generate the 0/1 response for the focal examinees.

For example, to manipulate uniform DIF with a magnitude of .4, the R code (Appendix E) will subtract .2 from the initial item difficulty generated for the item bank (b_{bank}) to obtain the item difficulty for the reference group ($b_{reference} = b_{bank} - .2$) and add .2 to b_{bank} to obtain the item difficulty for the focal group ($b_{focal} = b_{bank} + .2$). Consequently, the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = |(b_{bank} - .2) - (b_{bank} + .2)| = .4$. Similarly, to manipulate nonuniform DIF with a magnitude of 1.6, the program will add .8 to the initial item discrimination (a_{bank}) to obtain the item discrimination for the reference group ($a_{reference} = a_{bank} + .8$) and subtract .8 from a_{bank} to obtain the item discrimination for the focal group ($a_{focal} = a_{bank} - .8$). Consequently, the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = |(a_{bank} + .8) - (a_{bank} - .8)| = 1.6$.

Using the manipulation described above could lead to negative item discrimination for the focal group in some cases. For example, an item with the discrimination parameter of .708 was manipulated to exhibit very large nonuniform DIF ($|a_{reference} - a_{focal}| = 1.6$), the adjusted item discrimination would be .708 + .8 = 1.508 for the reference group and .708 – .8 = –.092 for the focal group. In practice, items with negative discrimination are hardly seen because such items are classified as low quality items and discarded during the test development step. Nevertheless, items with negative discrimination can still be found in the literature (e.g., Carter & Wilkinson, 1984; Kraja et al., 2007).

In the context of CAT, low quality items are screened during the precalibration process. Thus, items in the item bank are typically assumed to be high quality items. However, items examined in the precalibration process (or pretest items) are usually

administered as a nonadaptive test and sometimes as a paper-and-pencil test. As reported by Goldberg and Pedulla (2002) and Gu, Drake, and Wolfe (2006), item parameters of the same item can be changed due to the test mode. Thus, a good item defined by the precalibration process can probably perform worse when the item is administered in a live CAT. Therefore, this study retained all items that might lead to the case of negative discrimination when very large nonuniform DIF occurred in order to represent as many possible CAT settings as possible.

Data generation. To examine the effect of DIF on $\hat{\theta}_{CAT}$, 15 values between –3.5 to 3.5 in steps of 0.5 (i.e., ±3.5, ±3, ±2.5, ±2, ±1.5, ±1, ±0.5, and 0) were used as the true or generating ability levels (θ) of examinees in the reference and focal groups across all simulation conditions. As mentioned earlier, this range of true ability level matched the range of item difficulty in the item bank. Under each simulation condition, the probability of a correct response of an examinee with θ was computed for both groups using the three-parameter logistic model. This probability was then compared with a random number drawn from the standard uniform distribution, i.e., $U(0, 1)$. If the computed probability was less than the random number, a score of 0 was assigned to the examinee. Otherwise, a score of 1 was assigned (Harwell et al., 1996).

Based on the pilot simulations, the results obtained from 100 replications and beyond were stable. Appendix B presents the sample plots of average signed bias obtained from some extreme conditions, using data from 100-500 replications (the number on each line represents the number of replications in a hundred unit). As seen in the plots, the patterns of bias for both reference and focal groups were stable across true ability levels after 100 replications. Thus, the results discussed in the remaining chapters were based on the data obtained from 500 replications in each simulation condition. That is, a total of 15,000 examinees (15 true/generating ability levels × 2 groups of examinee × 500 replications) were generated for each cell. It should be noted that data for the baseline condition (no DIF factors were manipulated) were also replicated 500 times for each group of examinees.

Data Analysis. In this study, 180 simulation conditions were created (3 DIF contamination levels × 4 stages of DIF occurrence × 15 combinations of uniform and

nonuniform DIF magnitudes. For each condition, the average signed bias (BIAS), root mean squared error (RMSE), and empirical standard error (SE) of the estimate were computed across 500 replications using the following formulas (Chen, Ankenmann, & Chang, 2000):

$$\text{BIAS}(i,\theta) = \frac{1}{500}\sum_{n=1}^{500}\left(\hat{\theta}_{CAT(i,n)} - \theta\right) \tag{2}$$

$$\text{RMSE}(i,\theta) = \sqrt{\frac{1}{500}\sum_{n=1}^{500}\left(\hat{\theta}_{CAT(i,n)} - \theta\right)^2} \tag{3}$$

and

$$\text{SE}(i,\theta) = \sqrt{\frac{1}{500}\sum_{n=1}^{500}\left(\hat{\theta}_{CAT(i,n)} - \frac{1}{500}\sum_{n=1}^{500}\hat{\theta}_{CAT(i,n)}\right)^2} \tag{4}$$

where i denotes the i^{th} simulation condition and θ is the true ability level.

Generally, these indices provide descriptive information about the recovery of θ across all simulation conditions. As described earlier, each component of CAT was designed to have no effect on $\hat{\theta}_{CAT}$. Therefore, poor recovery of θ or bias in $\hat{\theta}_{CAT}$ should reflect the effect of DIF in operational items as simulated in each condition. In addition to the three indices, ANOVA models were applied to examine the effect of DIF contamination, DIF occurrence, type of DIF, and magnitude of DIF on the accuracy of $\hat{\theta}_{CAT}$.

2.2 Study 2: The accuracy of DIF detection in pretest items using NCS, $\widehat{\theta}_{CBT}$, and $\widehat{\theta}_{CAT}$ as the matching variable.

In this study, three types of scores on an operational test for each simulated examinee were: the number-correct score (NCS) and the IRT-based ability estimate ($\hat{\theta}$) from nonadaptive testing (either PPT or CBT); and the IRT-based ability estimate from computerized adaptive testing ($\hat{\theta}_{CAT}$). To obtain NCS and $\hat{\theta}$, 30 operational items were randomly drawn from the item bank generated in Study 1. These 30 items were formed as a fixed-order test and delivered to all examinees. The 0/1 responses of this test were generated using a three-parameter logistic model and the data generation procedure as described in Study 1. For each examinee, NCS was simply the sum of correct or 1 responses, while $\hat{\theta}$ was estimated using maximum likelihood estimation for a three-parameter logistic model. To obtain $\hat{\theta}_{CAT}$, additional 30 operational items from the same bank were adaptively delivered to the examinee using the same CAT administration as implemented in Study 1.

In addition to the 30 operational items, a fixed test of 16 pretest items was delivered to each examinee. The responses of the pretest items were generated using the item parameters provided by Lei, Chen, and Yu (2006) given in Appendix C. These item parameters resulted in both uniform and nonuniform DIF with small to large magnitude (.3 to 1.06). In order to compare the results with previous studies on DIF detection in pretest items (e.g. Zwick, Thayer, & Wingersky, 1994a, 1994b; Nandakumar & Roussos, 2001, 2004; Lei, Chen, & Yu, 2006), the responses on pretest items were not used in the ability estimation.

This study not only manipulated the same factors used in Study 1, but also included two additional factors: test impact and ratio of group sample sizes. In this study, test impact was defined as the group difference on the mean of true/generating ability levels. When no test impact existed, both reference and focal group were sampled from a population with a standard normal distribution, $N(0, 1)$. When test impact was present the population mean of the focal group was assumed to be one standard deviation below the population mean of the reference group, $N(-1 , 1)$. For the ratio of group sample sizes, two combinations of reference and focal group sample sizes (N_R:N_F = 1000:1000

and 1800:200) were used. These factors were similar to those manipulated in previous simulations on DIF detection in CAT (e.g., Zwick, Thayer, & Wingersky, 1994a; Feng, 2003; Nandakumar & Roussos, 2004; Lei, Chen, & Yu, 2006).

Data for each simulation condition in Study 2 were redrawn from the data generated for the similar condition in Study 1. To do so, each of 15,000 examinees in each cell of Study 1 was first assigned a probabilistic weight using a probability density function of a normal distribution. The mean of such a distribution for the reference group was controlled by the condition of test impact (0 for no test impact and 1 if test impact existed). Then, examinees for the corresponding condition in Study 2 were simulated by resampling the weighted examinees, depending on the condition of sample size of each group.

For instance, the responses for the condition of "6 uniform DIF items with large magnitude in the first stages of CAT, without test impact and N_R:N_F = 1000:1000" in Study 2 were redrawn from the responses for the condition of "6 uniform DIF items with large magnitude in first stages of CAT" in Study 1. First, each examinee in such a condition in Study 1 was assigned a probabilistic weight obtained from a standard normal distribution (i.e., no test impact) separately for the reference and focal group examinees. Next, based on the assigned weight, 1,000 examinees were randomly drawn from each group and their responses were used as the responses of the examinees on operational items in Study 2. Appendix D provides example distributions of original and resampled true ability levels. Finally, the true ability levels of these examinees were used to generate the responses on the 16 pretest items, using the item parameters described above.

The responses of these resampled examinees were then used to evaluate DIF in the pretest items by means of the Mantel-Haenszel statistic and logistic regression with the NCS, $\hat{\theta}$, and $\hat{\theta}_{CAT}$ as the matching variable. Power and Type I error rates of each detection method were investigated.

The Mantel-Haenszel (MH) statistic. First, examinees from the reference and focal groups were matched on 15 intervals of all possible NCS. Next, a 2×2 table was generated for each score interval and for each of the 16 pretest items as follows:

Group	Score on the i^{th} Item		
	1	0	
Reference	A_{is}	B_{is}	n_{Ris}
Focal	C_{is}	D_{is}	n_{Fis}
	m_{1is}	m_{0is}	n_{is}

where n_{is} is the total number of examinees in the s^{th} score interval for the i^{th} item; A_{is}, B_{is}, C_{is}, and D_{is} denote the number of examinee in their corresponding cells; and n_{1is}, n_{0is}, m_{1is}, and m_{0is} denote the marginal sums. The MH statistic for examining DIF in the i^{th} item was then computed by (Holland & Thayer, 1988):

$$MH_i = \left| \sum_{s=1}^{15} [A_{is} - E(A_{is})] \right|^2 \Big/ \sum_{s=1}^{15} Var(A_{is}) \tag{5}$$

where

$$Var(A_{is}) = \frac{n_{Ris} n_{Fis} m_{1is} m_{0is}}{n_{is}^2 (n_{is} - 1)} \tag{6}$$

and

$$E(A_{is}) = \frac{n_{Ris} m_{1is}}{n_{is}} \tag{7}$$

The MH statistic defined above is distributed approximately as a χ^2 statistic with one degree of freedom. If the MH statistic was statistically significant at the alpha level of .05, the item was identified as uniform DIF. Similar steps were also applied when $\hat{\theta}$ and $\hat{\theta}_{CAT}$ served as the matching variable. In such cases, the ability estimates were grouped into 15 intervals by dividing the range of –3.5 to 3.5 with an increment of .5. Note that the MH statistic with the matching variable $\hat{\theta}_{CAT}$ is in fact the ZTW method proposed by Zwick and colleagues (1994a).

Logistic regression (LR). As proposed by Swaminathan and Rogers (1990), the LR model for detecting DIF in the i^{th} item can be expressed by:

$$\text{logit}(\pi_{ij}) = \beta_{i0} + \beta_{i1} M_j + \beta_{i2} G_j + \beta_{i3} (MG)_j \tag{8}$$

where π_{ij} is the probability of the j^{th} examinee correctly answering the i^{th} item; M_j denotes the score on the matching variable (NCS, $\hat{\theta}$, or $\hat{\theta}_{CAT}$); G_j is the group index (1 for the reference group and 2 for the focal group); and $(MG)_k$ denotes the interaction term. After estimating the parameters $\beta_{i0}, \beta_{i1}, \beta_{i2}$, and β_{i3} in the model for each pretest item, uniform

and nonuniform DIF in the item were examined by testing the statistical significance of β_{i2} and β_{i3} with the Wald test. In later sections, the statistical tests for uniform and nonuniform DIF from the LR model are called LR-UDIF and LR-NDIF respectively. Also note that the LR with $\hat{\theta}_{CAT}$ as the matching variable is actually the CAT-LR procedure proposed by Lei, Chen, and Yu (2006).

2.3 Computer Programs

Currently, there are three computer programs specifically designed for simulation studies on CAT, including CATSim(Weiss & Guyer, 2010), SimulCAT (Han, 2010), and an R package called *catR* (Magis & Raîche, in press). However, these simulation programs do not allow users to manipulate DIF in operational items at different stages of CAT as designed for the present study. Therefore, the author developed computer codes in R (R Development Core Team, 2007) to implement CAT and generate responses for all simulation conditions. The R code developed for this study is provided in Appendix E.

To validate the accuracy of the R code, a pilot CAT administration was conducted using CATSim and the R code. First, CATSim and R independently simulated CAT data for 1,500 examinees on a 30-item test from a 500-item bank. In both programs, the generating ability, true item parameters, and CAT administration were identical to those described in Section 2.1. Next, the ability estimates obtained from CATSim and R were evaluated. The results revealed that the estimates from both programs were highly correlated (r = .987). Bias in the ability estimate was also compared.

As seen in Figure 3 below, the average signed bias from both programs varied in a small range of –.1 to .1. Moreover, the magnitude of biases in the ability estimates (the unsigned bias) from both programs were around .2 for the average ability levels (–3 to 3). The only noticeable difference occurred at the extreme ability levels (±3.5) in which CATSim yielded a larger bias. Such a difference was due to the fact that CATSim and the R code defined a range of valid ability estimates differently, i.e., ±4 for CATSim and ±3.5 for the R code. In other words, each program truncated extreme ability estimates at different points, resulting in a large difference in bias at the extreme ability levels.

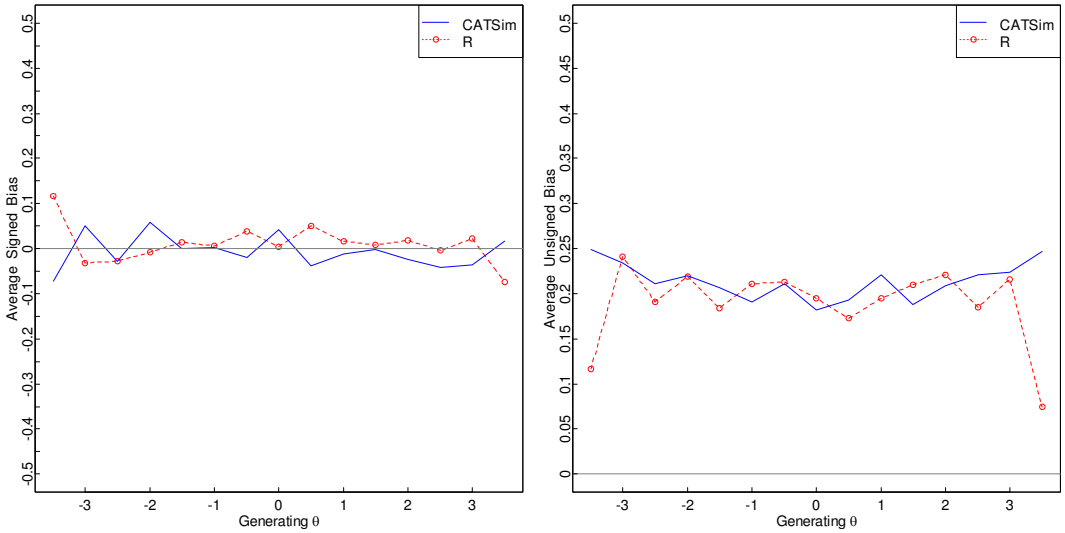


Figure 3. Bias in the ability estimate obtained from CATSim and the R code.

Besides the different range of ability estimates, CATSim and R also used different random number generators. To further evaluate the R code, another comparison was thus conducted by assuming that both programs implemented the same random mechanism. Specifically, using the starting point ($\hat{\theta}_0$) and order of administered items generated by CATSim, the ability estimate and its observed standard error for each examinee at each stage of CAT were recomputed by R. These recomputed estimates were then compared to the original estimates reported by CATSim. Overall, the initial and the recomputed estimates were comparable. Particularly, the average differences between both estimates were about .003 for the ability estimate and –.012 for the standard error of estimate.

Chapter 3: Results of Study 1

Study 1 was designed to examine the effect of DIF in operational CAT items on the ability estimate obtained from CAT ($\hat{\theta}_{CAT}$). In this study, 180 conditions were generated by manipulating DIF contamination (the number of DIF items in a 30-item CAT); DIF occurrence (the stages of CAT that DIF items were administered); DIF type (uniform, nonuniform, or both); and DIF magnitude (the difference in item parameters for the reference and focal groups). The effect of these simulation factors was examined by ANOVA models, using the average signed bias (BIAS) of $\hat{\theta}_{CAT}$ as the dependent variable. The recovery of $\hat{\theta}_{CAT}$ was then examined using the BIAS, root mean squared error (RMSE), and empirical standard error (SE) of $\hat{\theta}_{CAT}$.

3.1 Results from ANOVA models

It should be noted that DIF type and DIF magnitude were not fully crossed with the other simulation factors in this study. As seen in Appendix B, there were some cases when DIF items exhibited only one type of DIF (either uniform or nonuniform). For example, Conditions 2-5 manipulated only uniform DIF. In such conditions, it could be said that the magnitude of nonuniform DIF was 0. Thus, this study technically manipulated four levels of DIF magnitude including 0, .4, 1, and 1.6. When the magnitudes of both types of DIF were 0, there were in fact no DIF items in CAT. Also note that the magnitudes of nonuniform DIF were not nested within uniform DIF, and vice versa. Thus, the overall design of simulations in Study 1 was neither a fully crossed (factorial) nor nested design, but in fact an incomplete design.

However, the magnitudes of each DIF type were fully crossed with DIF occurrence and DIF contamination. Hence, this study applied two separate five-way ANOVA models (i.e., one for each DIF type) to the data to examine the effect of simulation factors on the BIAS of $\hat{\theta}_{CAT}$. Table 1 summarizes the results of a factorial ANOVA for the effect of nonuniform DIF and other simulation factors including DIF contamination, DIF occurrence, examinees' group (reference or focal), and the true ability or generating θ. In addition, the results of a factorial ANOVA for the effect of uniform DIF and those

simulation factors are provided in Table 2.

ANOVA results reveal that all effects were statistically significant at the .001 level, but their contributions to the variation in the BIAS of $\hat{\theta}_{CAT}$ were very small. Specifically, the ANOVA models for nonuniform and uniform DIF could explain only 35% and 45%, respectively, of the total variation in the average signed bias. As seen in Tables 1 and 2, group of examinees (reference versus focal groups) accounted for the most variation in both models (about 20% of the total variation in the average signed bias). In addition, the generating θ, the second-order interactions between DIF contamination and group (CON × Group) and between the magnitude of uniform DIF and group (UDIF × Group), and the third-order interaction between the magnitude of uniform DIF, DIF contamination, and group (UDIF × CON × Group) were the other effects that contributed more than 1% to the total variation of bias in the ability estimate.

Table 1

Summary of ANOVA Results for the Effect of Nonuniform DIF with Other Factors

Effect	*df*	*SS*	*MS*	η^2
Nonuniform DIF magnitude (NMAG)	3	380.406	126.802	.002
DIF Contamination (CON)	2	18.050	9.025	<.001
DIF Occurrence (OCC)	3	261.587	87.196	.002
Group	**1**	**96932.267**	**96932.267**	**.600**
Generating ability (θ)	**14**	**15297.354**	**1092.668**	**.095**
NMAG × CON	6	18.927	3.154	<.001
NMAG × OCC	9	242.755	26.973	.002
NMAG × Group	3	1346.071	448.690	.008
NMAG × θ	**42**	**2276.666**	**54.206**	**.014**
CON × OCC	6	42.674	7.112	<.001
CON × Group	**2**	**25613.088**	**12806.544**	**.159**
CON × θ	**28**	**2107.184**	**75.257**	**.013**
OCC × Group	3	42.401	14.134	<.001
OCC × θ	42	799.427	19.034	.005
Group × θ	**14**	**4919.850**	**351.418**	**.030**
NMAG × CON × OCC	18	35.650	1.981	<.001
NMAG × CON × Group	6	314.521	52.420	.002
NMAG × CON × θ	84	1022.294	12.170	.006
NMAG × OCC × Group	9	106.466	11.830	.001
NMAG × OCC × θ	126	1066.102	8.461	.007
NMAG × Group × θ	42	2634.595	62.728	.016
CON × OCC × Group	6	22.170	3.695	<.001
CON × OCC × θ	84	207.772	2.473	.001
CON × Group × θ	28	1295.938	46.284	.008
OCC × Group × θ	42	1085.674	25.849	.007
NMAG × CON × OCC × Group	18	30.837	1.713	<.001
NMAG × CON × OCC × θ	252	360.215	1.429	.002
NMAG × CON × Group × θ	84	1088.200	12.955	.007
NMAG × OCC × Group × θ	126	1314.371	10.432	.008
CON × OCC × Group × θ	84	243.886	2.903	.002
NMAG × CON × OCC × Group × θ	252	379.032	1.504	.002

The *p* values for all effects are <.001

Table 2

Summary of ANOVA Results for the Effect of Uniform DIF with Other Factors

Effect	*df*	*SS*	*MS*	η^2
Uniform DIF magnitude (UMAG)	3	577.204	192.401	.003
DIF Contamination (CON)	2	17.638	8.819	.000
DIF Occurrence (OCC)	3	271.289	90.430	.001
Group	**1**	**83675.587**	**83675.587**	**.434**
Generating ability (θ)	**14**	**15312.258**	**1093.733**	**.080**
UMAG × CON	6	57.369	9.562	.000
UMAG × OCC	9	90.574	10.064	.000
UMAG × Group	**3**	**44195.090**	**14731.697**	**.229**
UMAG × θ	42	861.856	20.520	.004
CON × OCC	6	44.743	7.457	.000
CON × Group	**2**	**22156.468**	**11078.234**	**.115**
CON × θ	**28**	**2141.666**	**76.488**	**.011**
OCC × Group	3	37.854	12.618	.000
OCC × θ	42	892.031	21.239	.005
Group × θ	**14**	**4885.743**	**348.982**	**.025**
UMAG × CON × OCC	18	13.519	.751	.000
UMAG × CON × Group	**6**	**11800.149**	**1966.692**	**.061**
UMAG × CON × θ	84	425.454	5.065	.002
UMAG × OCC × Group	9	26.553	2.950	.000
UMAG × OCC × θ	126	42.170	.335	.000
UMAG × Group × θ	42	1403.685	33.421	.007
CON × OCC × Group	6	22.127	3.688	.000
CON × OCC × θ	84	238.395	2.838	.001
CON × Group × θ	28	1340.964	47.892	.007
OCC × Group × θ	42	1209.295	28.793	.006
UMAG × CON × OCC × Group	18	9.547	.530	.000
UMAG × CON × OCC × θ	252	31.845	.126	.000
UMAG × CON × Group × θ	84	368.081	4.382	.002
UMAG × OCC × Group × θ	126	111.574	.886	.001
CON × OCC × Group × θ	84	275.732	3.283	.001
UMAG × CON × OCC × Group × θ	252	43.730	.174	.000

Note: The *p* values for all effects are <.001

3.2 BIAS of $\hat{\theta}_{CAT}$

Figures 4-18 show patterns of the average signed bias (BIAS) of $\hat{\theta}_{CAT}$ obtained after 30 operational items were administered. Each figure presents the results from each combination of nonuniform and uniform DIF magnitudes across other simulation conditions including the generating θ (–3.5 to 3.5 with an increment of .5), DIF contamination (the number of DIF items in the operational CAT), and DIF occurrence (stages of CAT that DIF items were administered). Each figure consists of four panels representing each stage of CAT in which DIF items were administered. In addition, within these panels, the bias values when the number of DIF items was varied from 6 to 24 items are plotted against those obtained from the baseline condition (i.e., the simulation condition where all operational CAT items did not exhibit DIF). Also, the bias values for the reference and focal groups are simultaneously presented.

3.2.1 BIAS when DIF items exhibited only uniform DIF

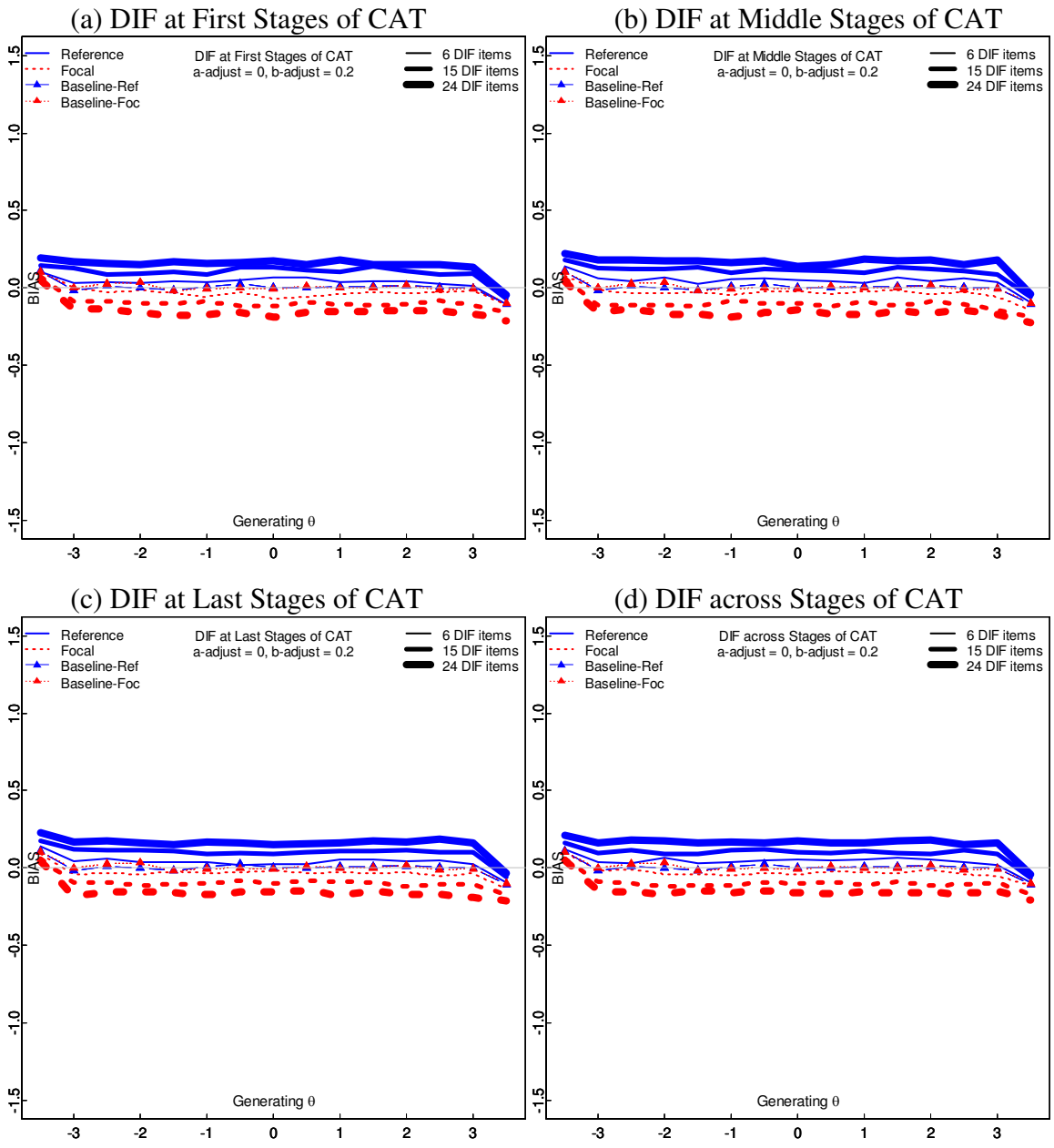
To observe the effect of each simulation condition on the recovery of $\hat{\theta}_{CAT}$, it seems wise to start with the simplest scenario: operational CAT items exhibited only uniform DIF with magnitude of .4 (i.e., $b_{\mathrm{F}} - b_{\mathrm{R}} = .4$). As shown in the four panels of Figure 4, the pattern of average bias are divided into three parts based on the generating ability level and group. Specifically, bias in $\hat{\theta}_{CAT}$ for the reference group with θ = –3.5 to 3 was positive and larger than the bias obtained from the baseline condition. At θ = 3.5, the bias for this group decreased into negative values. In contrast, the focal group examinees with θ = –3 to 3.5 received negatively biased ability estimates. However, the bias for very low-ability examinees in the focal group climbed into positive values.

It should be noticed that the difference in bias of $\hat{\theta}_{CAT}$ between the reference and focal groups for average ability levels was constant, given the number of DIF items and stage of CAT that DIF occurred. For example, when 6 items exhibited uniform DIF with a magnitude of .4 at the first stages of CAT, the bias was about .1 for the reference group and –.1 for the focal group. This means that, at each ability level, the ability estimate for the reference group was larger than that for the focal group around .2 on the ability scale, or mathematically speaking $\hat{\theta}_R - \hat{\theta}_F = (\theta_R + .1) - (\theta_F + (-.1)) = .2$ when $\theta_R = \theta_F = \theta$.

When the number of DIF items increased, the difference also increased. As seen in the figure, with 24 moderate-DIF items, the difference was up to .6. This pattern was consistent across the stages of CAT.

When magnitudes of uniform DIF increased from .4 to 1.6 (Figures 4-6), the patterns of BIAS were similar to the just described patterns. However, the magnitude of bias and the difference in bias between the reference and focal groups were much clearer. That is, the larger the magnitudes of uniform DIF, the larger the bias and differences in bias between groups. As seen in the figures, the plots of BIAS vertically expanded as the magnitude of uniform DIF and numbers of DIF items increased, regardless of θ and DIF occurrence. In other words, the reference group examinees tended to receive larger positively biased estimates, while the focal group examinee with the same ability received lower negatively biased estimates.

Figure 4. BIAS of $\hat{\theta}_{CAT}$ when the magnitude of uniform DIF was .4.



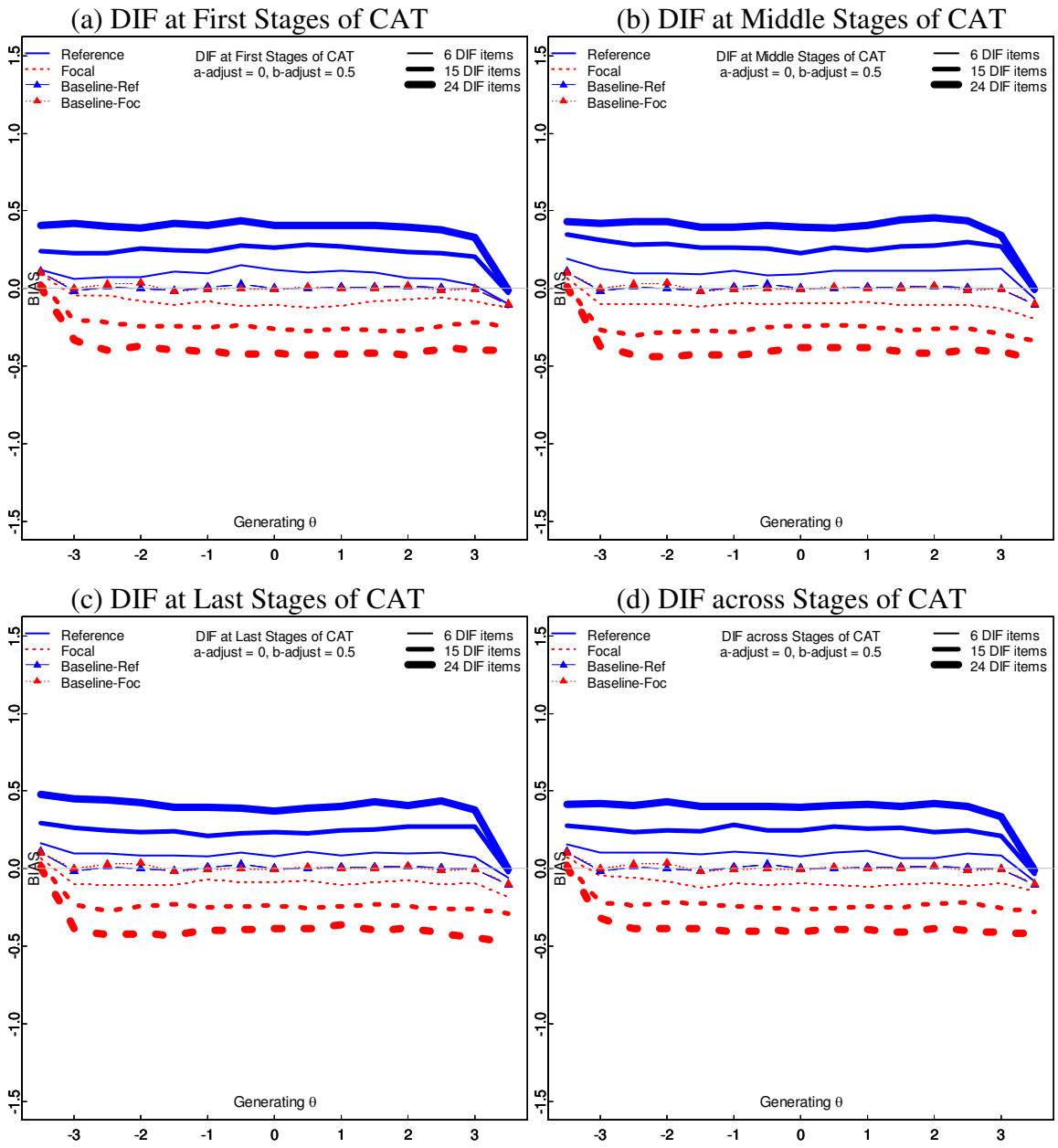
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, a-adjust = 0, meaning that $a_{\text{reference}} = a_{\text{focal}} = a_{\text{bank}}$ or nonuniform DIF did not occur.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .2$ and $b_{\text{focal}} = b_{\text{bank}} + .2$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 5. BIAS of $\hat{\theta}_{CAT}$ when the magnitude of uniform DIF was 1.0.



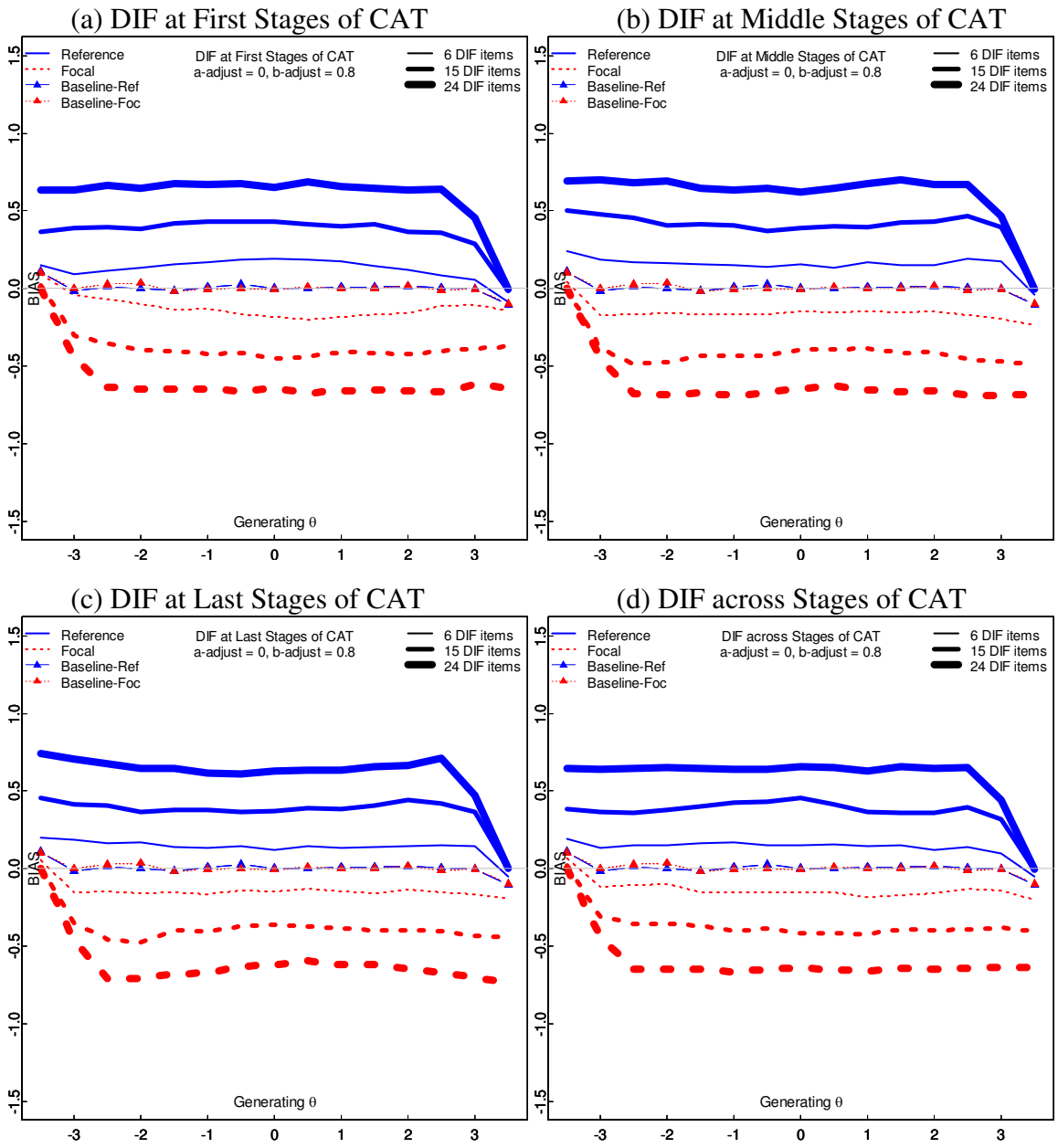
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, a-adjust = 0, meaning that $a_{\text{reference}} = a_{\text{focal}} = a_{\text{bank}}$ or nonuniform DIF did not occur.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .5$ and $b_{\text{focal}} = b_{\text{bank}} + .5$, yielding the magnitude of uniform DIF $= |b_{\text{reference}} - b_{\text{focal}}| = 1.0$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 6. BIAS of $\hat{\theta}_{CAT}$ when the magnitude of uniform DIF was 1.6.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, a-adjust = 0, meaning that $a_{\text{reference}} = a_{\text{focal}} = a_{\text{bank}}$ or nonuniform DIF did not occur.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .8$ and $b_{\text{focal}} = b_{\text{bank}} + .8$, yielding the magnitude of uniform DIF $= |b_{\text{reference}} - b_{\text{focal}}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

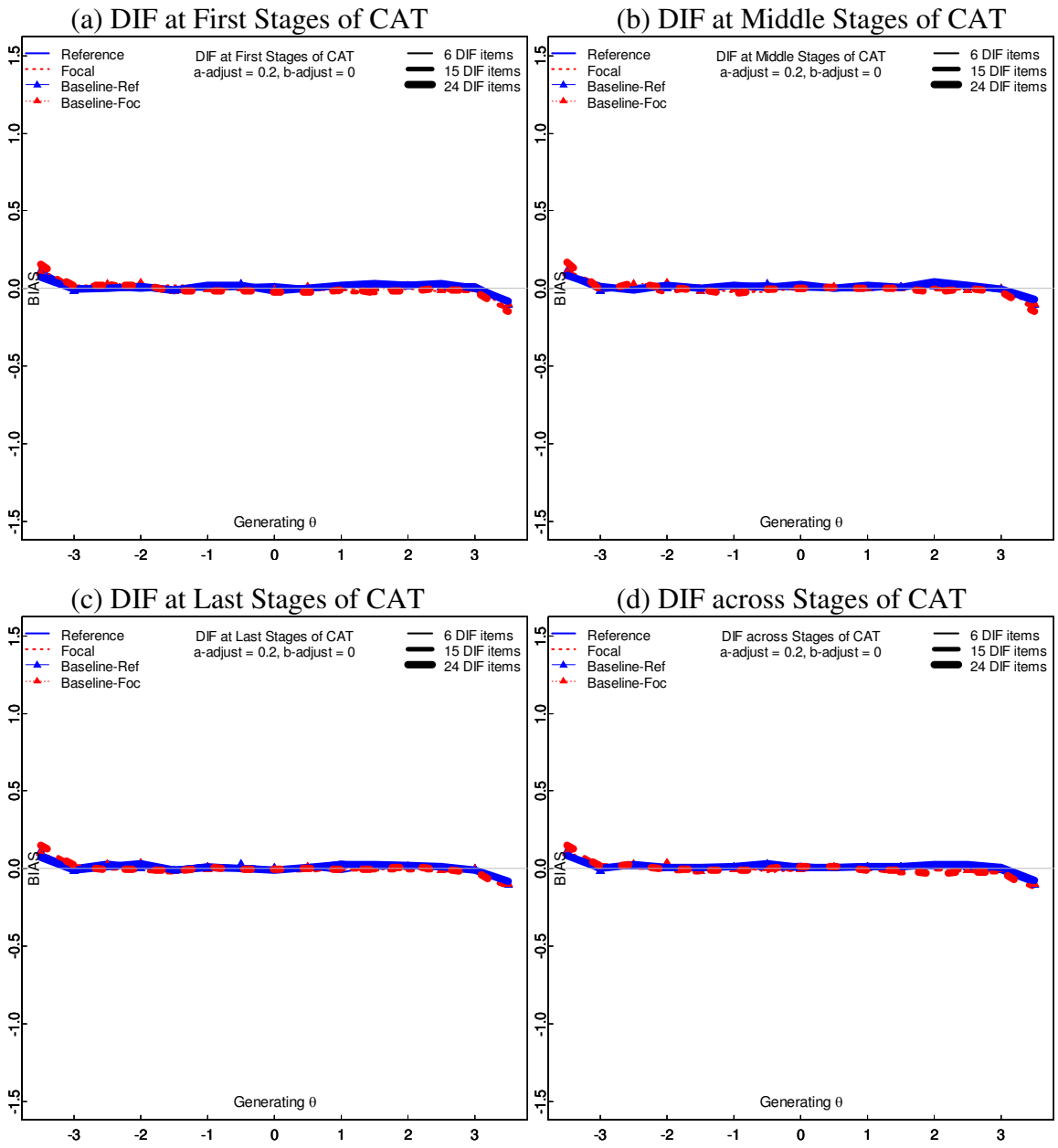
3.2.2 BIAS when DIF items exhibited only nonuniform DIF

In Figure 7, the average bias in ability estimates obtained from CAT with moderate nonuniform DIF items ($a_R - a_F = .4$) are presented. It appeared that both reference and focal group examinees obtained equivalently biased $\hat{\theta}_{CAT}$ across numbers of DIF items, stages of CAT that DIF occurred, and θ levels. Specifically, the bias in $\hat{\theta}_{CAT}$ was approximately .1 for very low-ability examinees, 0 for average examinees, and –.1 for very high-ability examinees. Moreover, the bias values observed in this simulation condition were comparable to those observed in the baseline condition. In other words, CAT with a moderate magnitude of nonuniform DIF seemed to recover well from the effect of DIF, despite the condition of DIF contamination and DIF occurrence.

The patterns of bias interestingly changed when the magnitude of nonuniform DIF increased from .4 to 1.6 (Figures 7-9). It was found that nonuniform DIF had more effect on the bias of ability estimates for the focal group than the reference group. As seen in Figures 8 and 9, the bias values for the reference group were approximately the same with those in DIF-free CAT. On the other hand, for the focal group, examinees with $\theta <$ 0 received higher positive bias, while examinees with $\theta > 0$ received lower negative bias (Figure 9). By this pattern of BIAS, CAT with nonuniform DIF items seemed to favor the focal group examinees with low ability levels more than those with higher ability from the same group.

Regarding the effect of DIF contamination when DIF items only showed nonuniform DIF, the difference in bias of $\hat{\theta}_{CAT}$ between both groups increased when the number of DIF items increased. Precisely, the focal group examinees with $\theta < 0$ tended to receive higher positive bias, while those with $\theta > 0$ received lower negative bias when the number of DIF items increased. In contrast, the bias for the reference group seemed to be stable across the number of DIF items. As for the effect of DIF occurrence, the results surprisingly showed that the bias values obtained when the magnitude of nonuniform DIF was 1.6 and DIF occurred at the first stages of CAT were generally larger than those obtained when the same amount of DIF occurred at the last stages of CAT. In such cases, the magnitude of bias in absolute values for examinees with positive θ was greater than those for examinees with negative θ (Figure 9a vs. 9c).

Figure 7. BIAS of $\hat{\theta}_{CAT}$ when the magnitude of nonuniform DIF was .4.



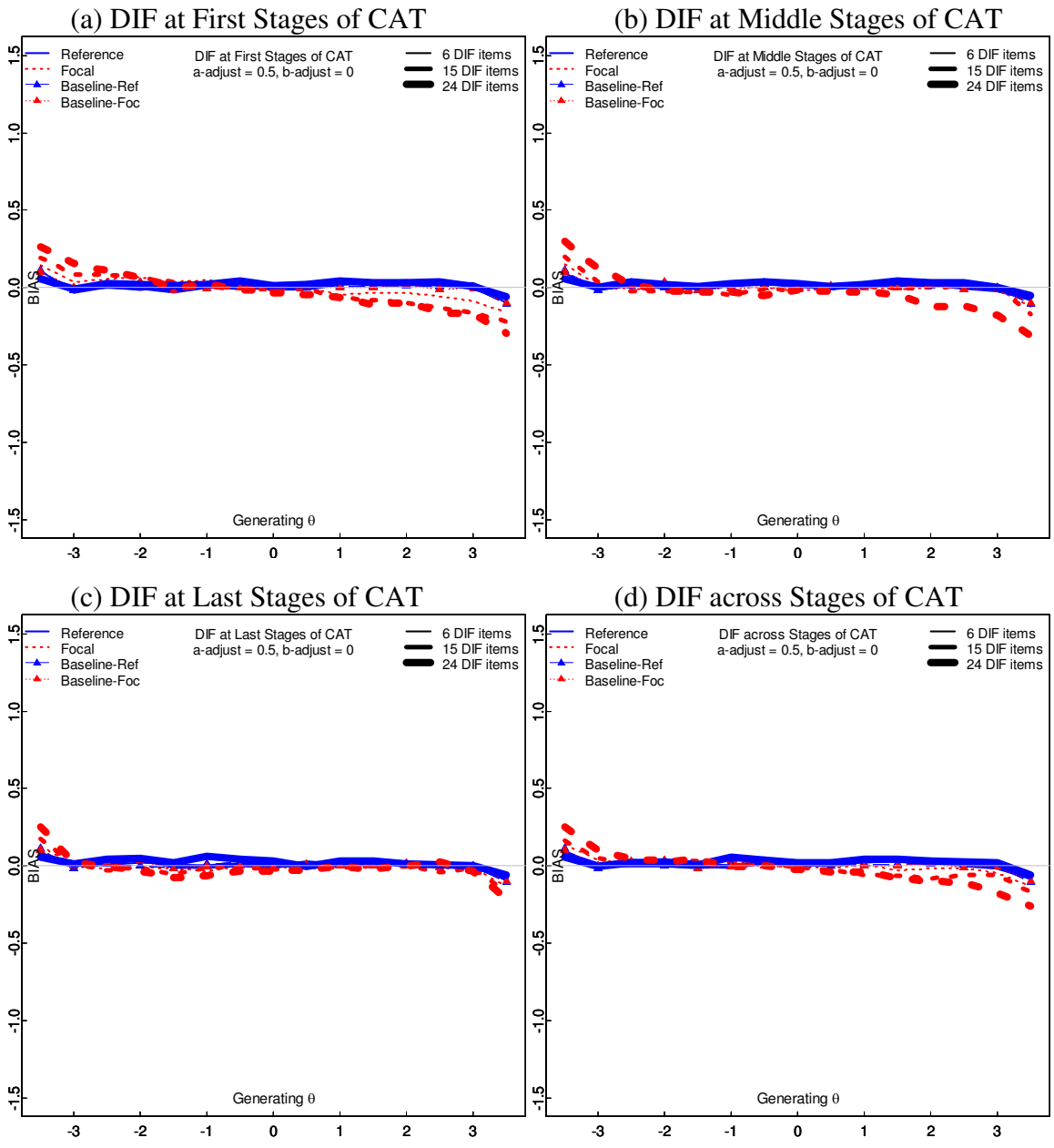
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .2$ and $a_{\text{focal}} = a_{\text{bank}} - .2$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, b-adjust = 0, meaning that $b_{\text{reference}} = b_{\text{focal}} = b_{\text{bank}}$ or uniform DIF did not occur.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 8. BIAS of $\hat{\theta}_{CAT}$ when the magnitude of nonuniform DIF was 1.0.



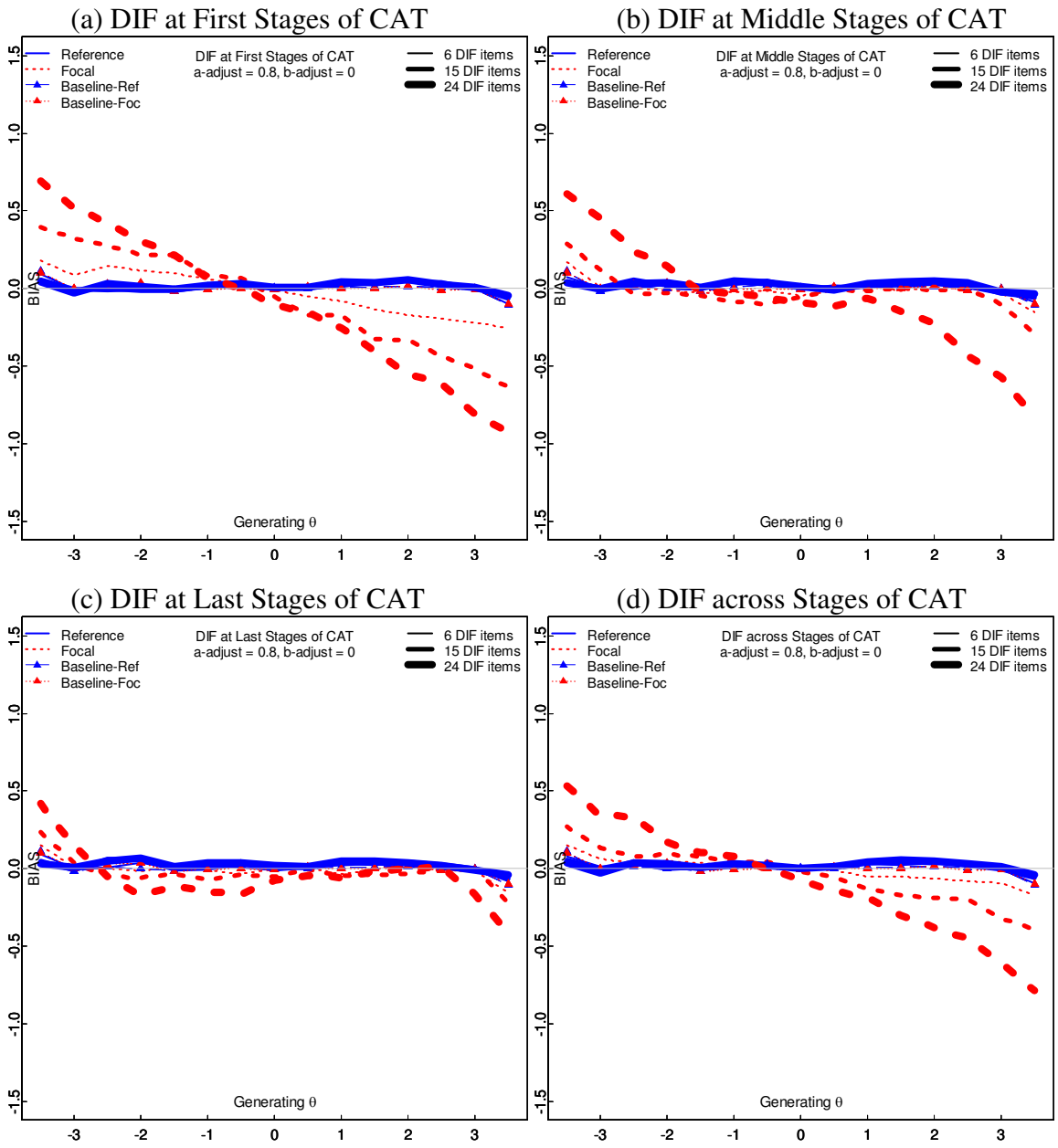
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .5$ and $a_{\text{focal}} = a_{\text{bank}} - .5$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1.0$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, b-adjust = 0, meaning that $b_{\text{reference}} = b_{\text{focal}} = b_{\text{bank}}$ or uniform DIF did not occur.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 9. BIAS of $\hat{\theta}_{CAT}$ when the magnitude of nonuniform DIF was 1.6.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .8$ and $a_{\text{focal}} = a_{\text{bank}} - .8$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, b-adjust = 0, meaning that $b_{\text{reference}} = b_{\text{focal}} = b_{\text{bank}}$ or uniform DIF did not occur.

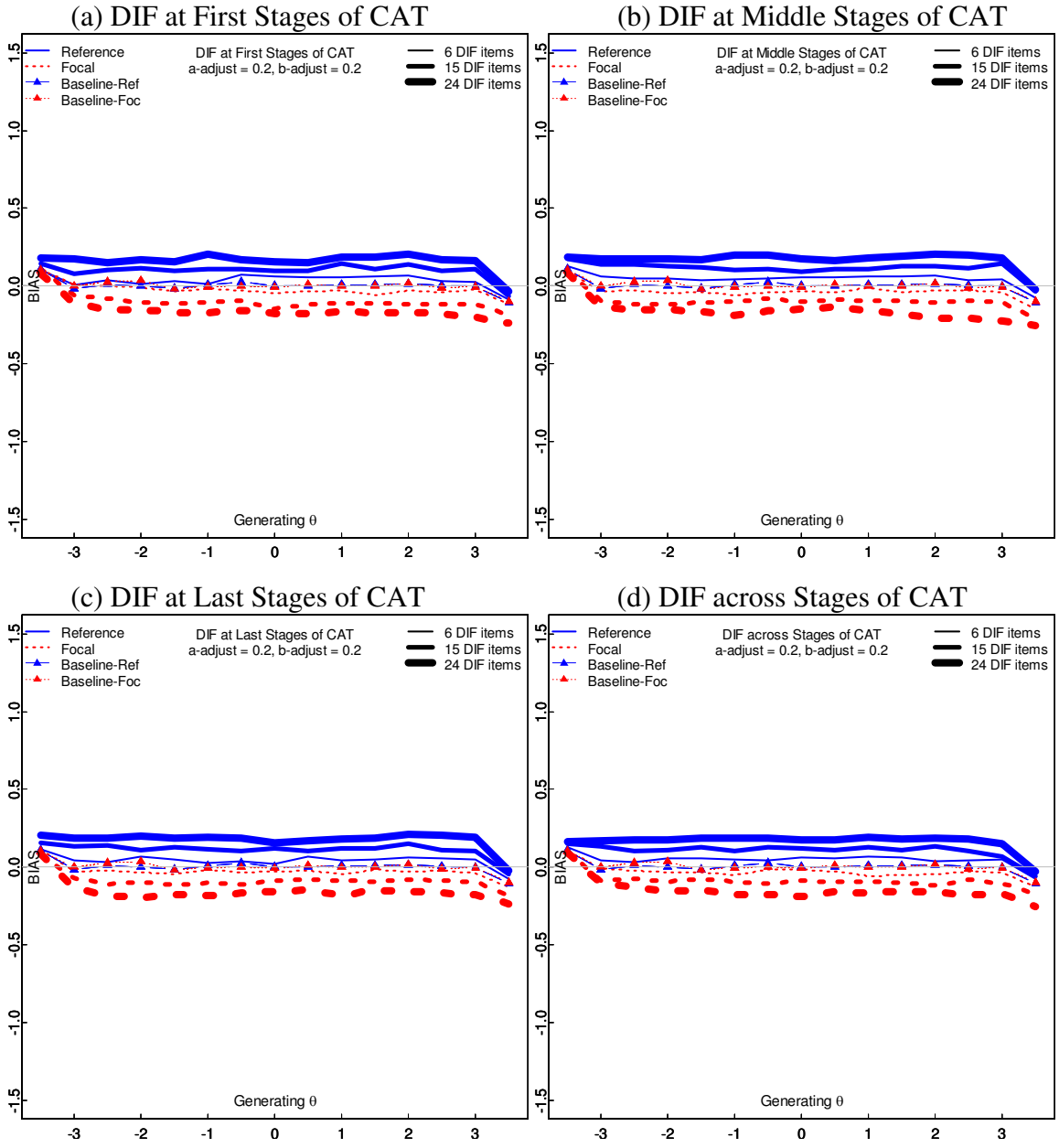
Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

3.2.3 BIAS when DIF items exhibited both types of DIF with the same magnitude

Figures 10–12 illustrate the patterns of BIAS when CAT administered items that exhibit both types of DIF with the same magnitude. Basically, the patterns observed in these conditions were a combination of the patterns described above. First, the bias values increased as the magnitude of DIF and numbers of DIF items increased for average θ levels, regardless of DIF occurrence. Second, the change in magnitudes of uniform DIF consistently affected the bias in $\hat{\theta}_{CAT}$ for both groups. Third, the change in magnitudes of nonuniform DIF only resulted in a rapid change (decrease) in the bias values for the focal group examinees, especially those with extreme θ levels. Finally, for the largest nonuniform DIF cases (Figure 12), the focal group examinees with $\theta > 0$ had lower negative bias when DIF occurred at the first stages of CAT than when the same amount of DIF occurred at the last stages of CAT.

Figure 10. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were .4 and .4, respectively.



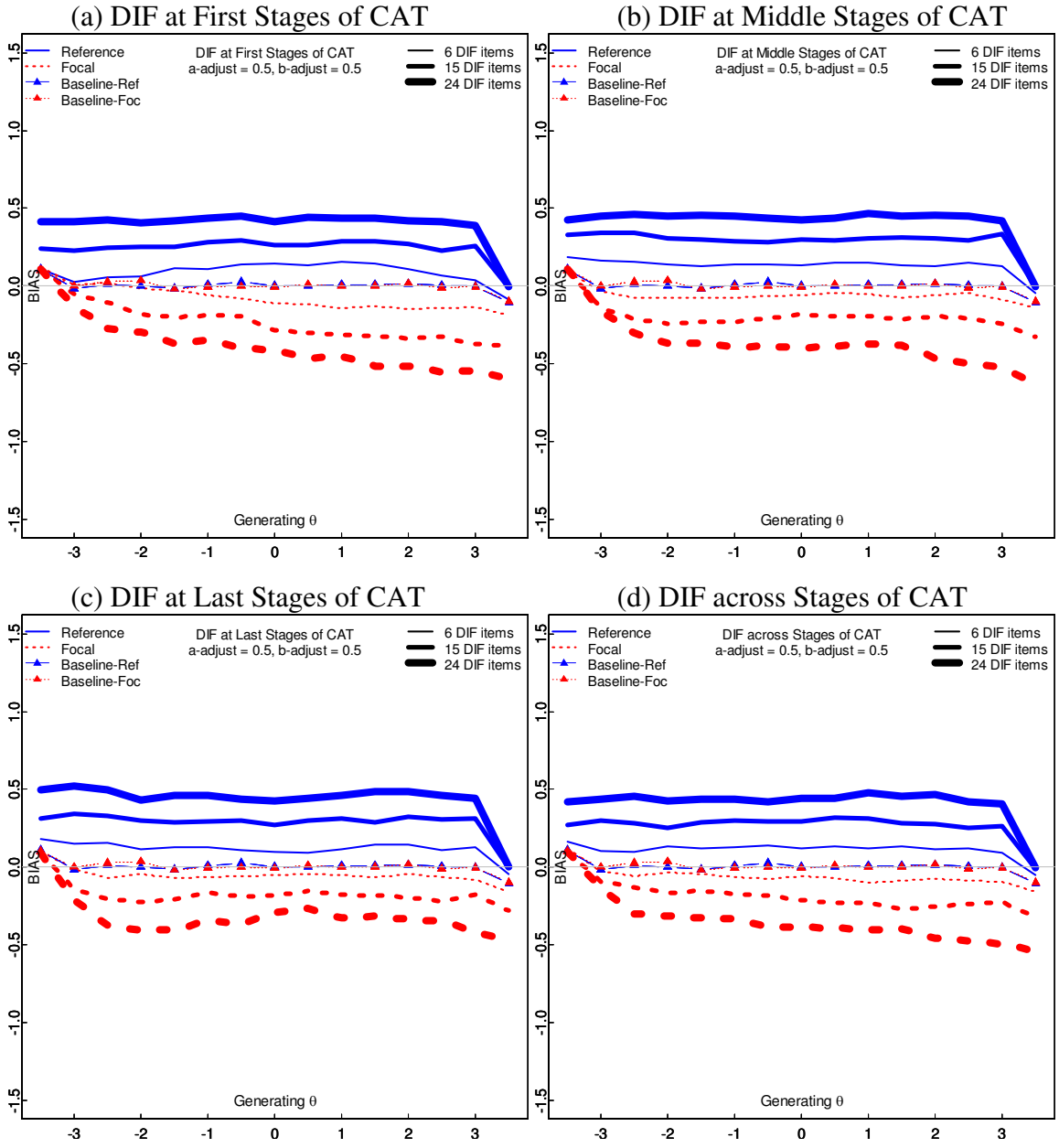
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .2$ and $a_{focal} = a_{bank} - .2$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .2$ and $b_{focal} = b_{bank} + .2$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 11. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.0 and 1.0, respectively.



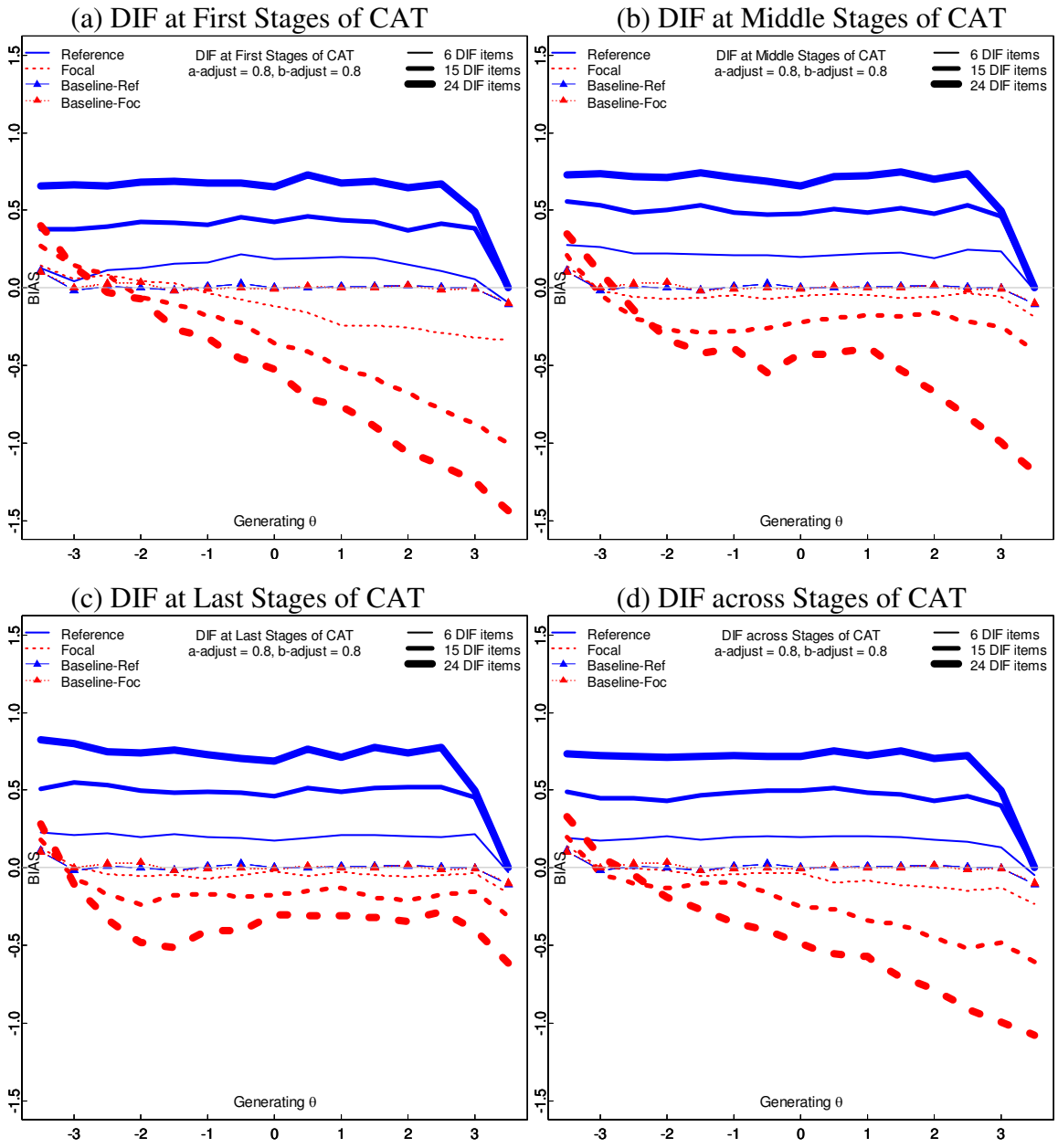
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .5$ and $a_{\text{focal}} = a_{\text{bank}} - .5$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .5$ and $b_{\text{focal}} = b_{\text{bank}} + .5$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 12. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.6 and 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .8$ and $a_{focal} = a_{bank} - .8$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .8$ and $b_{focal} = b_{bank} + .8$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

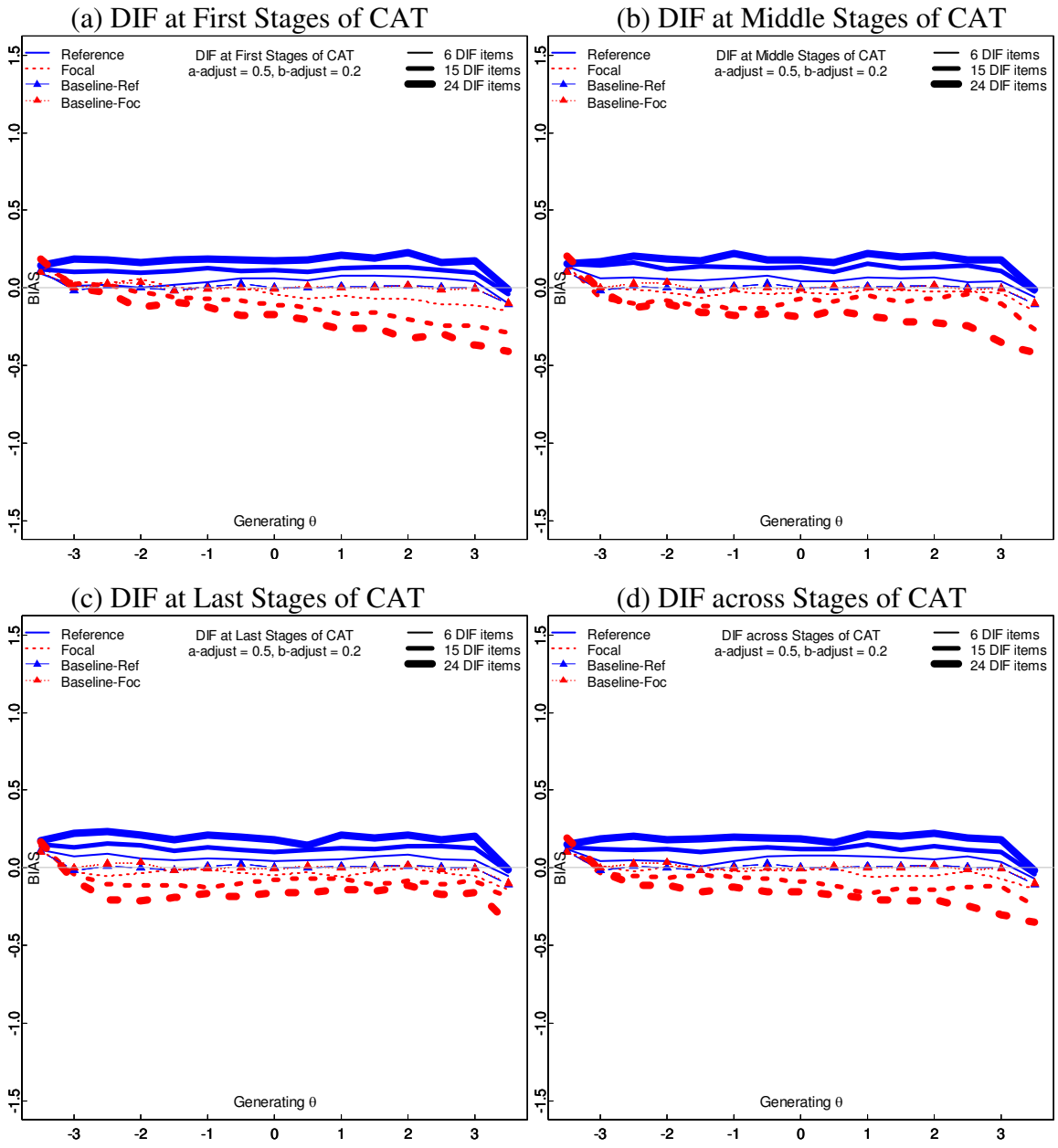
Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

3.2.4 BIAS when DIF items exhibited both type of DIF with different magnitudes

Overall, if the magnitude of nonuniform DIF was greater than the magnitude of uniform DIF, the effect of nonuniform DIF overshadowed the effect of uniform DIF. In Figures 13–15, the magnitudes of nonuniform DIF were higher than uniform DIF. Consequently, the bias in $\hat{\theta}_{CAT}$ for the focal group examinees dramatically changed as the magnitude of nonuniform DIF and the number of DIF items changed. As observed in previous cases of nonuniform DIF, the largest magnitude of nonuniform DIF, regardless of uniform DIF magnitude, yielded the result that CAT with DIF items at the first stages of CAT yielded relatively higher bias of $\hat{\theta}_{CAT}$ for the focal group examinees than CAT with DIF items at other stages (Figures 14a and 15a).

In contrast, the bias in $\hat{\theta}_{CAT}$ shown in Figures 16–18 consistently changed for both groups of examinees. This pattern was apparently due to the fact that the magnitudes of uniform DIF were larger than nonuniform DIF. In addition, as the magnitude of uniform DIF increased, despite the change in nonuniform magnitude, the difference in bias of $\hat{\theta}_{CAT}$ across groups constantly changed. These patterns of bias in $\hat{\theta}_{CAT}$ were observed in all conditions of DIF occurrence. Finally, unlike the case of nonuniform DIF where only the focal group was affected by DIF items, when nonuniform and uniform DIF simultaneously occurred in the same items, both groups were affected by DIF.

Figure 13. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.0 and .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



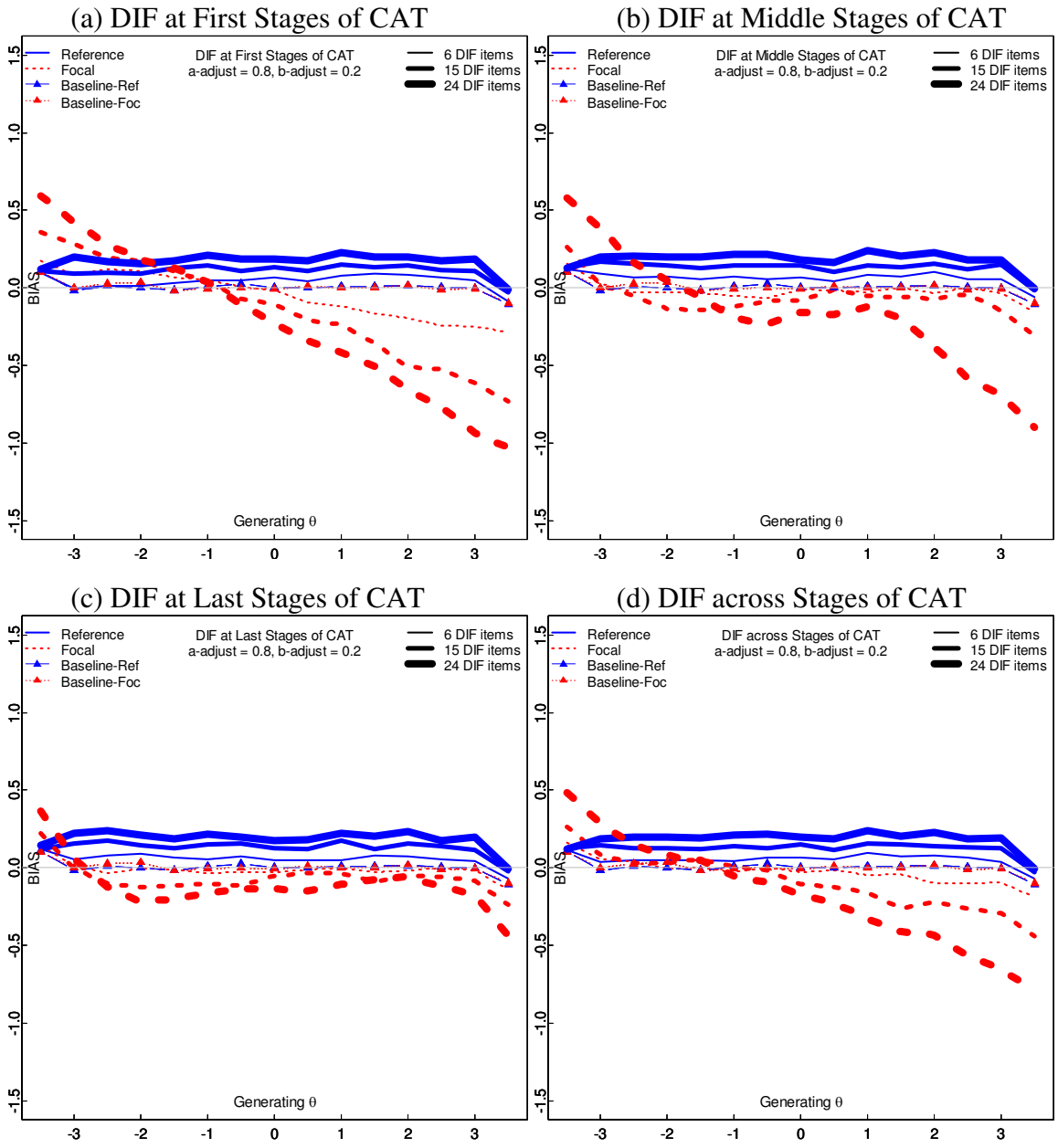
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .5$ and $a_{focal} = a_{bank} - .5$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .2$ and $b_{focal} = b_{bank} + .2$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 14. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.6 and .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



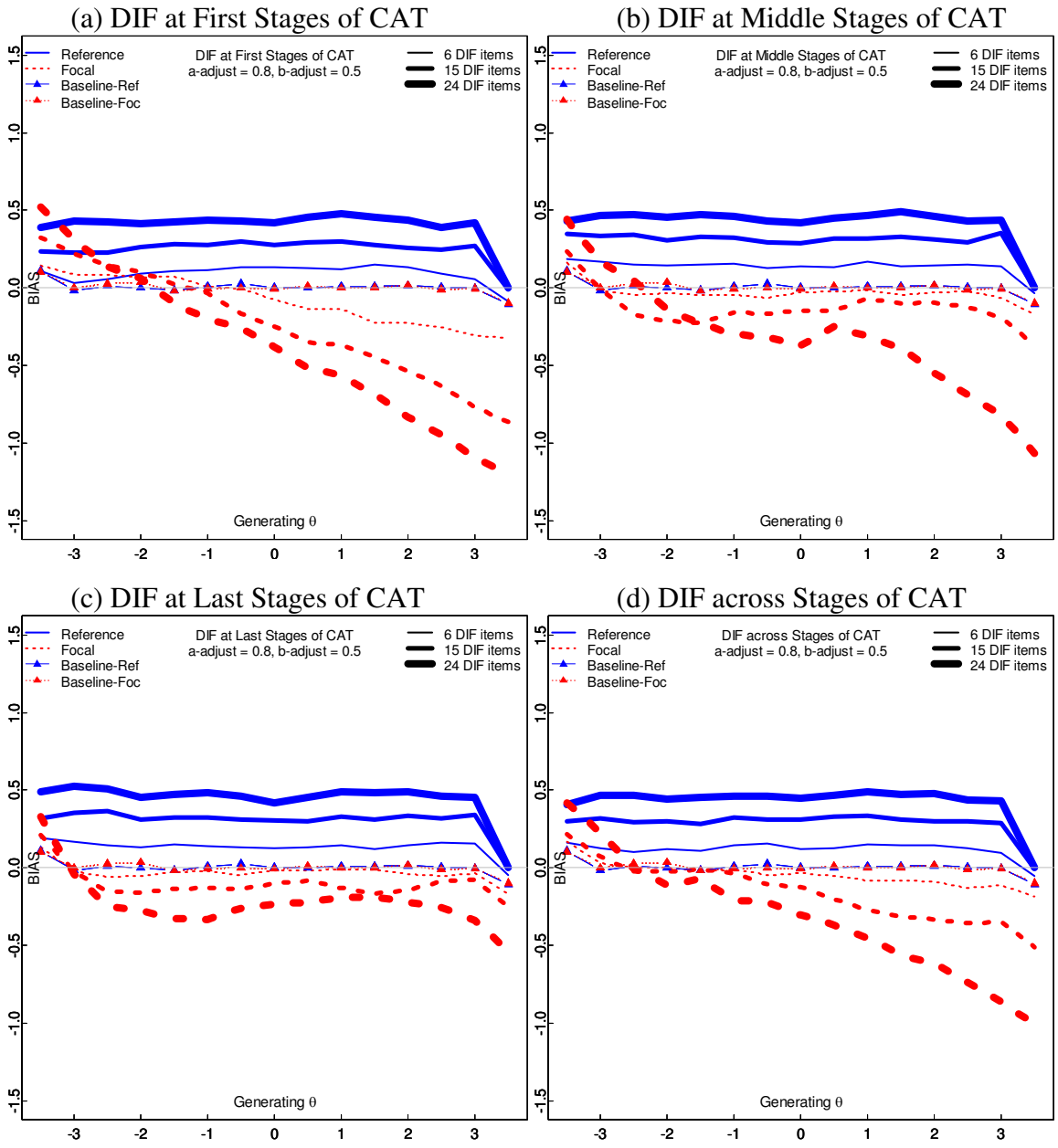
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .8$ and $a_{focal} = a_{bank} - .8$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .2$ and $b_{focal} = b_{bank} + .2$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 15. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.6 and 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



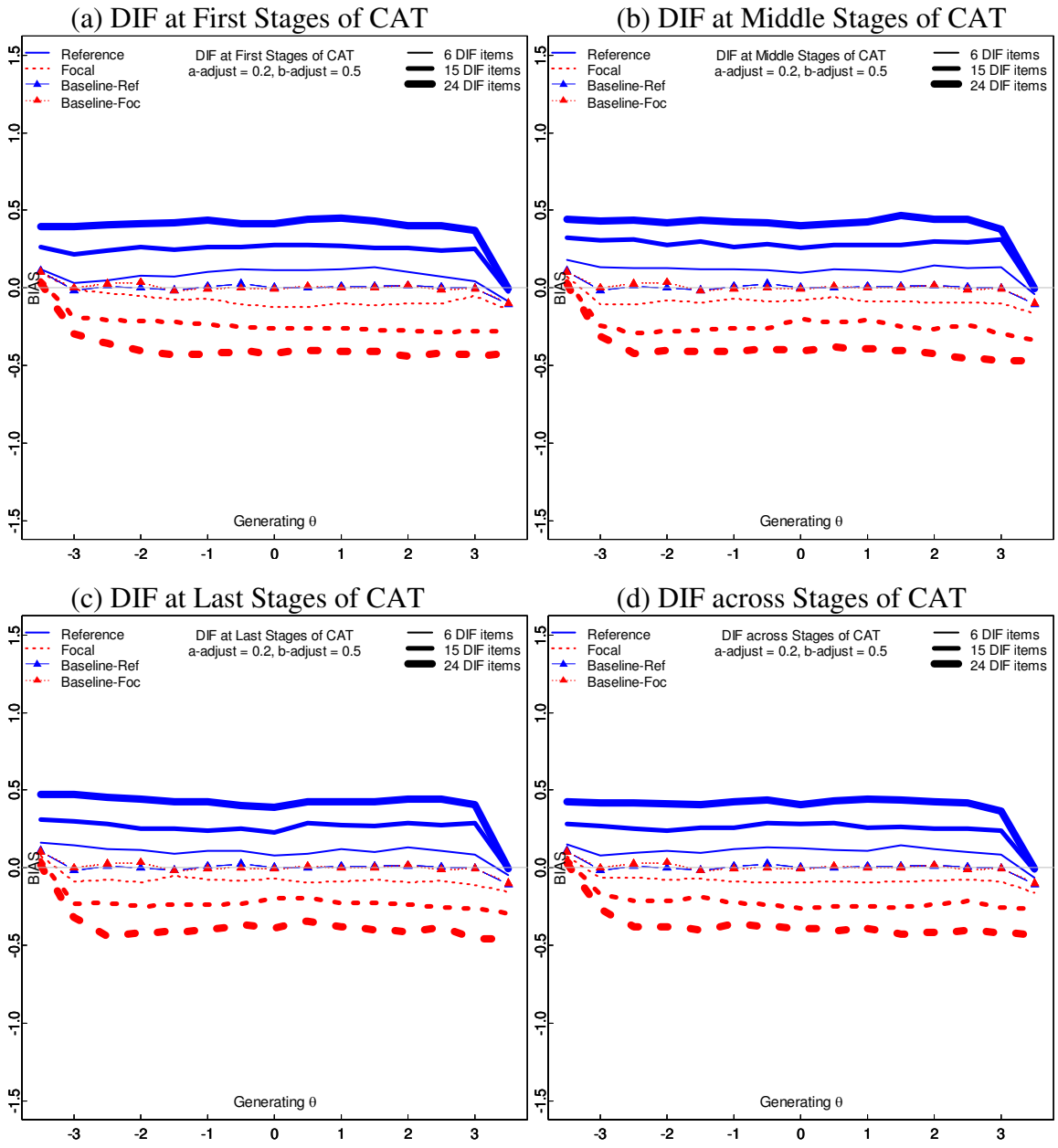
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .8$ and $a_{focal} = a_{bank} - .8$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .5$ and $b_{focal} = b_{bank} + .5$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 16. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were .4 and 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



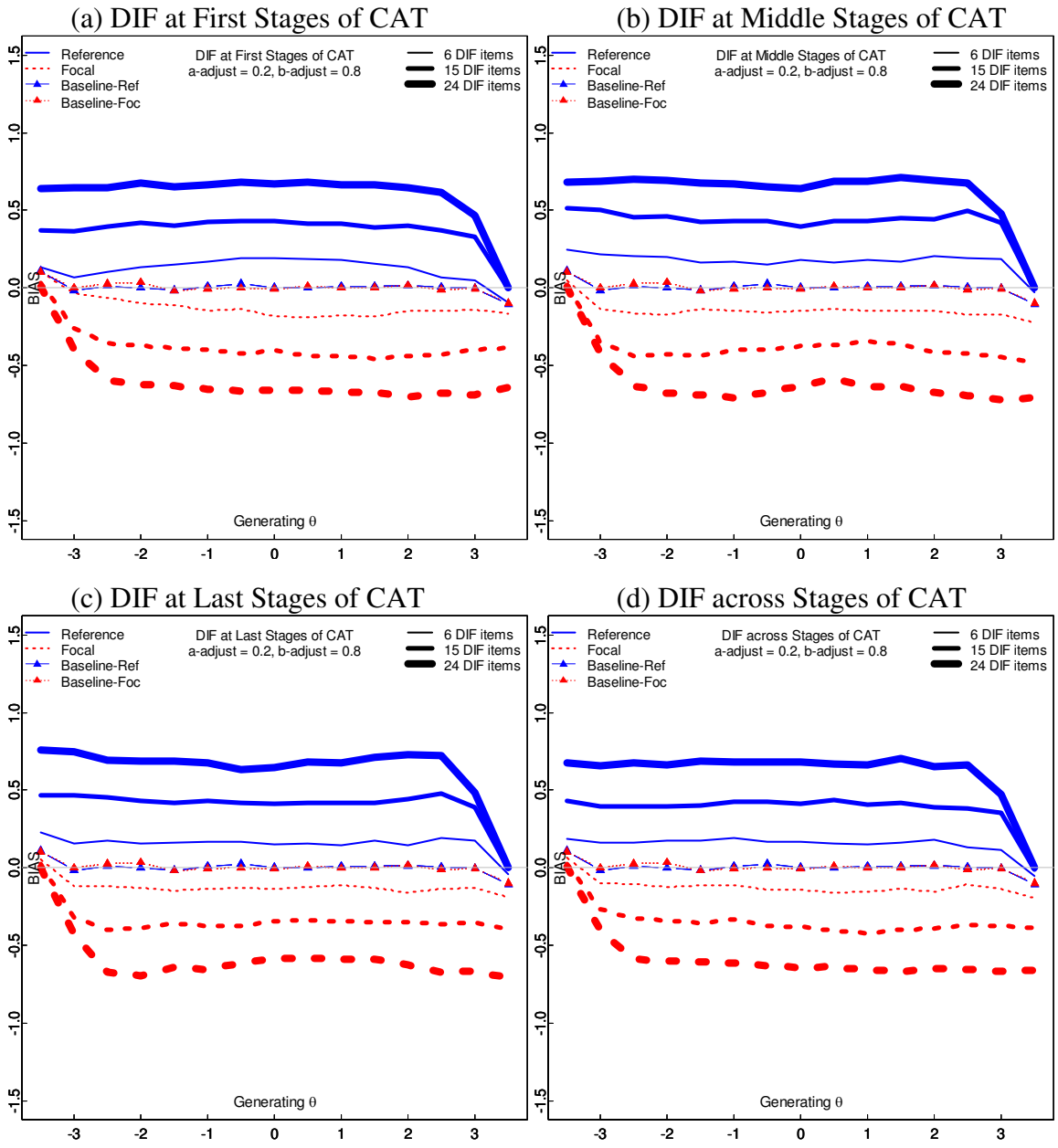
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .2$ and $a_{focal} = a_{bank} - .2$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .5$ and $b_{focal} = b_{bank} + .5$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 17. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were .4 and 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



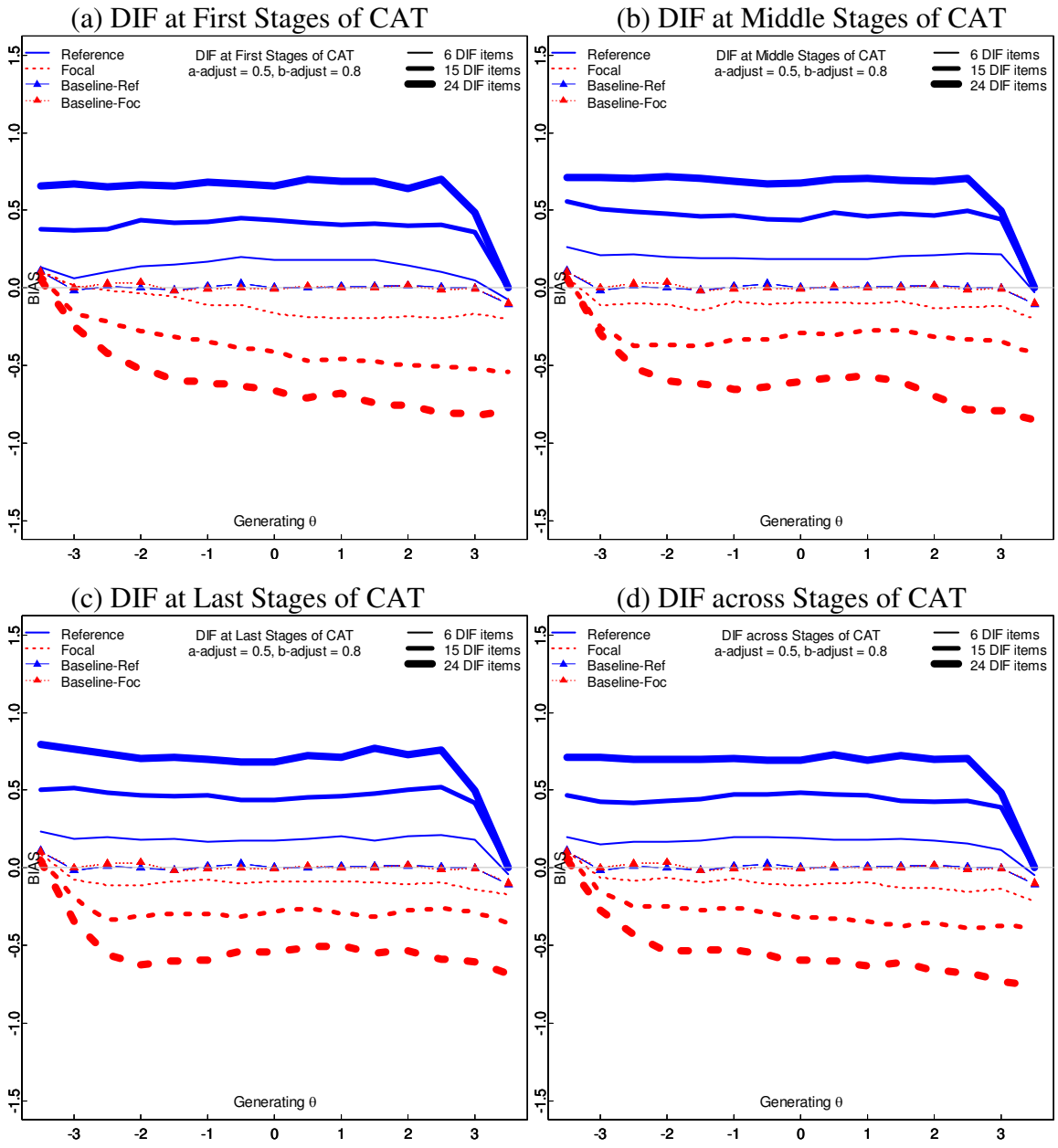
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .2$ and $a_{\text{focal}} = a_{\text{bank}} - .2$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .8$ and $b_{\text{focal}} = b_{\text{bank}} + .8$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 18. BIAS of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.0 and 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .5$ and $a_{focal} = a_{bank} - .5$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .8$ and $b_{focal} = b_{bank} + .8$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

3.3 RMSE of $\widehat{\theta}_{CAT}$

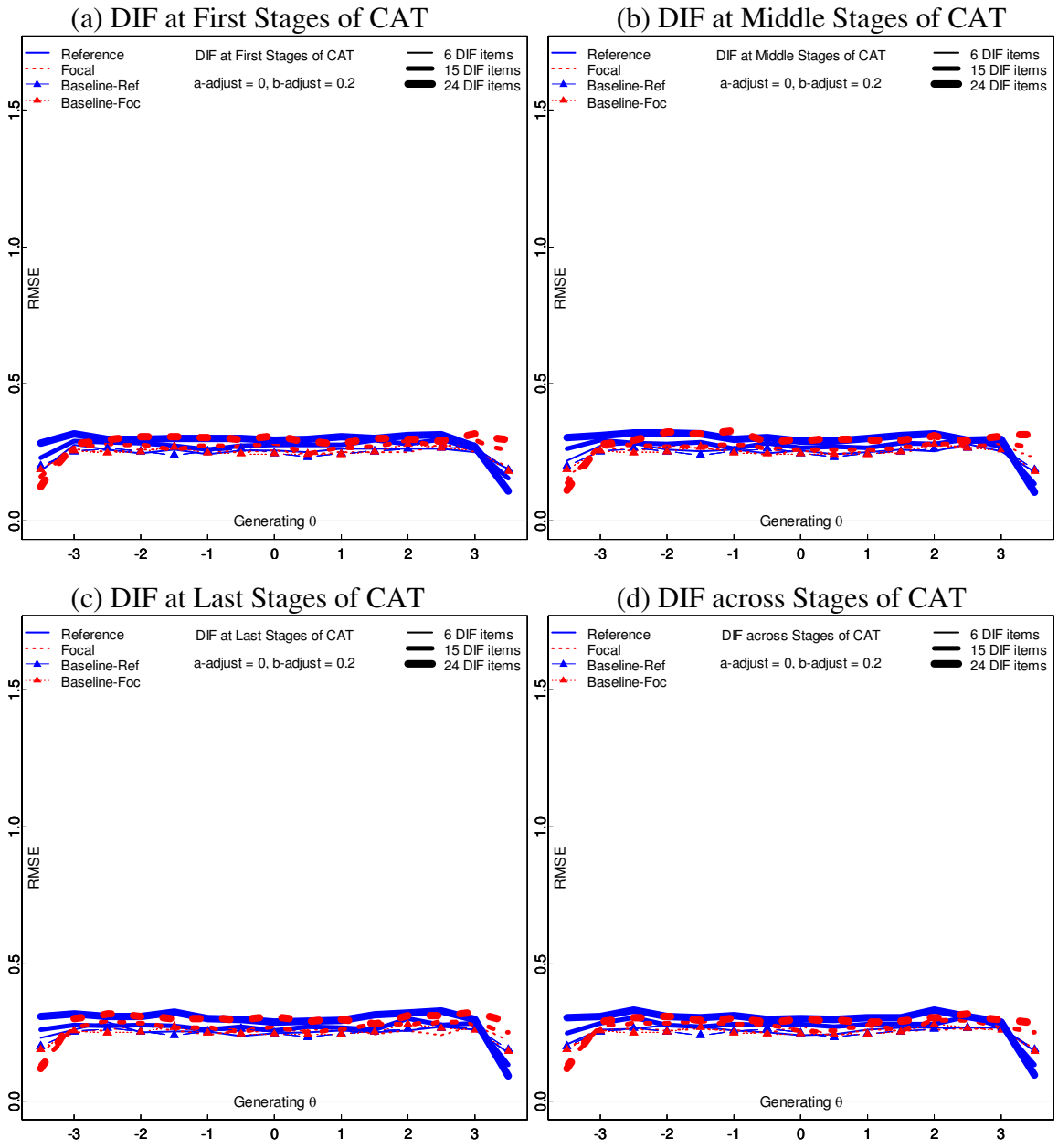
Using a similar organization of the BIAS results above, the observed values of root mean squared error (RMSE) which indicates the deviation of $\widehat{\theta}_{CAT}$ from θ in absolute values are presented in Figures 19-33.

3.3.1 RMSE when DIF Items Exhibited Only Uniform DIF

Overall, when uniform DIF items with magnitude of .4 (Figure 19) were administered, the RMSE values were stable (around .3) for most conditions of the number of DIF items, stages of CAT that DIF occurred, and generating θ. However, even with a moderate magnitude of uniform DIF ($b_F - b_R = .4$), the RMSE for the examinees with $\theta = \pm 3.5$ seemed to be affected. Specifically, the reference group examinees with the very low ability had larger RMSE, while the focal group peer had lower RMSE. In contrast, the reference group examinees at the upper end of ability scale had smaller RMSE than their peers. When the magnitude of uniform DIF increased from .4 to 1.0 and 1.6, the patterns of RMSE (Figures 19-21) changed differently depending on θ and group of examinees. Particularly, the RMSE for the reference group was larger than the focal group when $\theta < -3.5$, but smaller when $\theta = 3.5$. For θ from –3 to 3, examinees from both groups had comparable RMSE.

These patterns of RMSE were also observed across DIF contamination and DIF occurrence, but with clearer trends. That is, the difference in RMSE between the reference and focal groups increased as the number of DIF items increased. As seen in the figures, the RMSE obtained from 24 DIF items essentially departed from the RMSE obtained from DIF-free CAT, specifically around .5 on the ability scale. In addition, when DIF occurred at the first or middle stages of CAT, the difference in RMSE between the very low-ability examinees from the two groups greatly increased.

Figure 19. RMSE of $\hat{\theta}_{CAT}$ when the magnitude of uniform DIF was .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



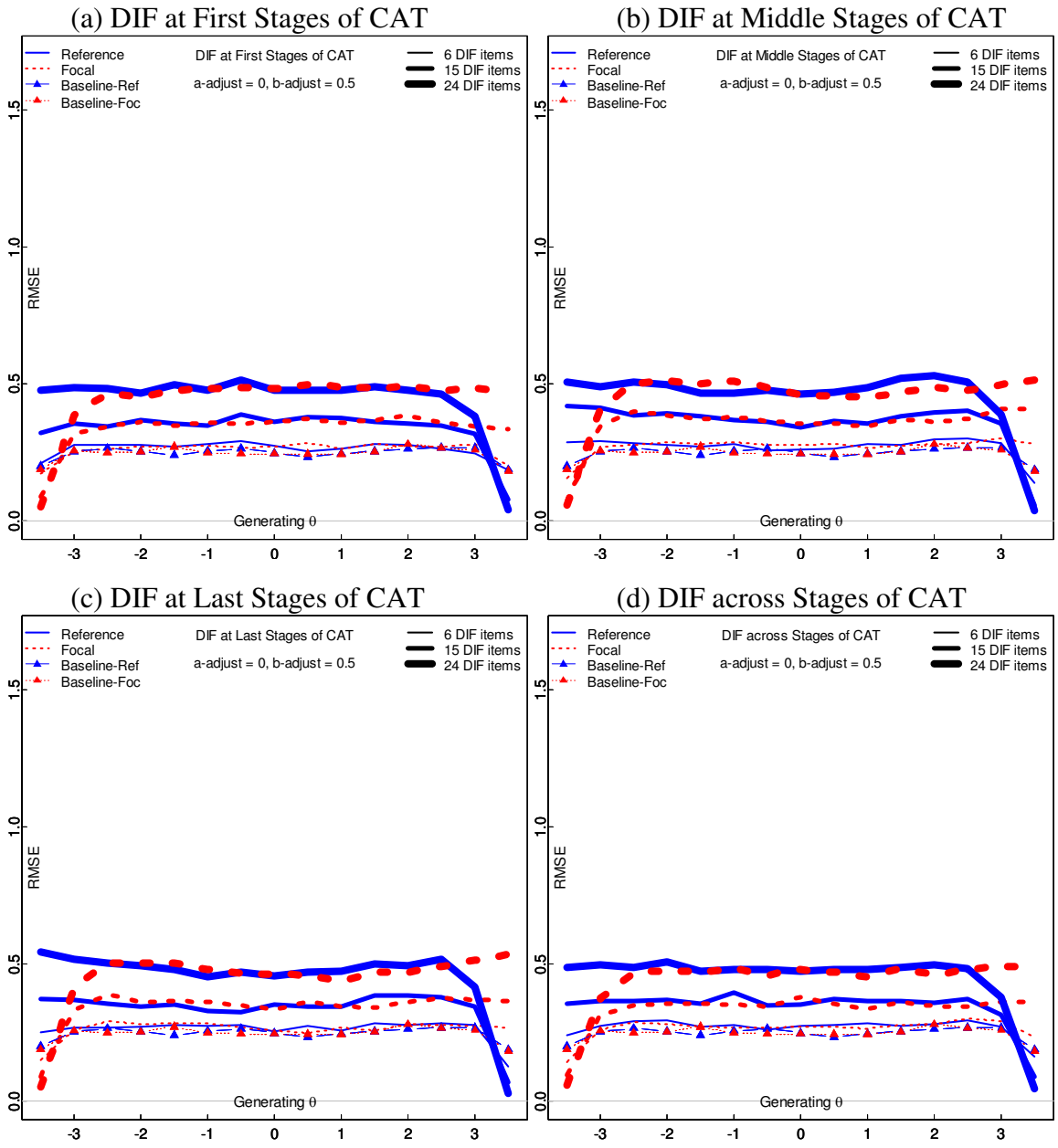
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, a-adjust = 0, meaning that $a_{\text{reference}} = a_{\text{focal}} = a_{\text{bank}}$ or nonuniform DIF did not occur.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .2$ and $b_{\text{focal}} = b_{\text{bank}} + .2$, yielding the magnitude of uniform DIF $= |b_{\text{reference}} - b_{\text{focal}}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 20. RMSE of $\hat{\theta}_{CAT}$ when the magnitude of uniform DIF was 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



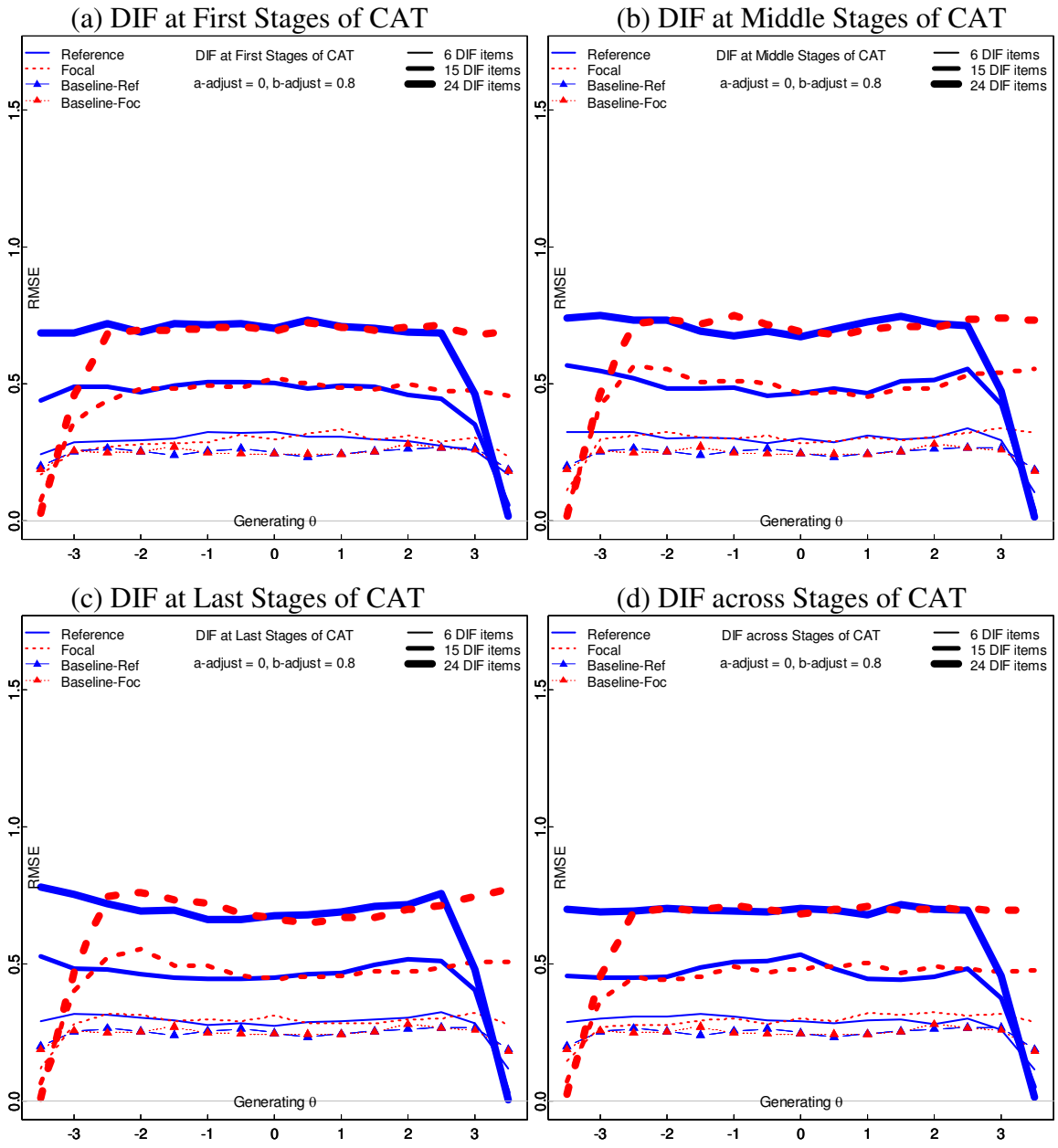
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, a-adjust = 0, meaning that $a_{\text{reference}} = a_{\text{focal}} = a_{\text{bank}}$ or nonuniform DIF did not occur.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .5$ and $b_{\text{focal}} = b_{\text{bank}} + .5$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 21. RMSE of $\hat{\theta}_{CAT}$ when the magnitude of uniform DIF was 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, a-adjust = 0, meaning that $a_{\text{reference}} = a_{\text{focal}} = a_{\text{bank}}$ or nonuniform DIF did not occur.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .8$ and $b_{\text{focal}} = b_{\text{bank}} + .8$, yielding the magnitude of uniform DIF $= |b_{\text{reference}} - b_{\text{focal}}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

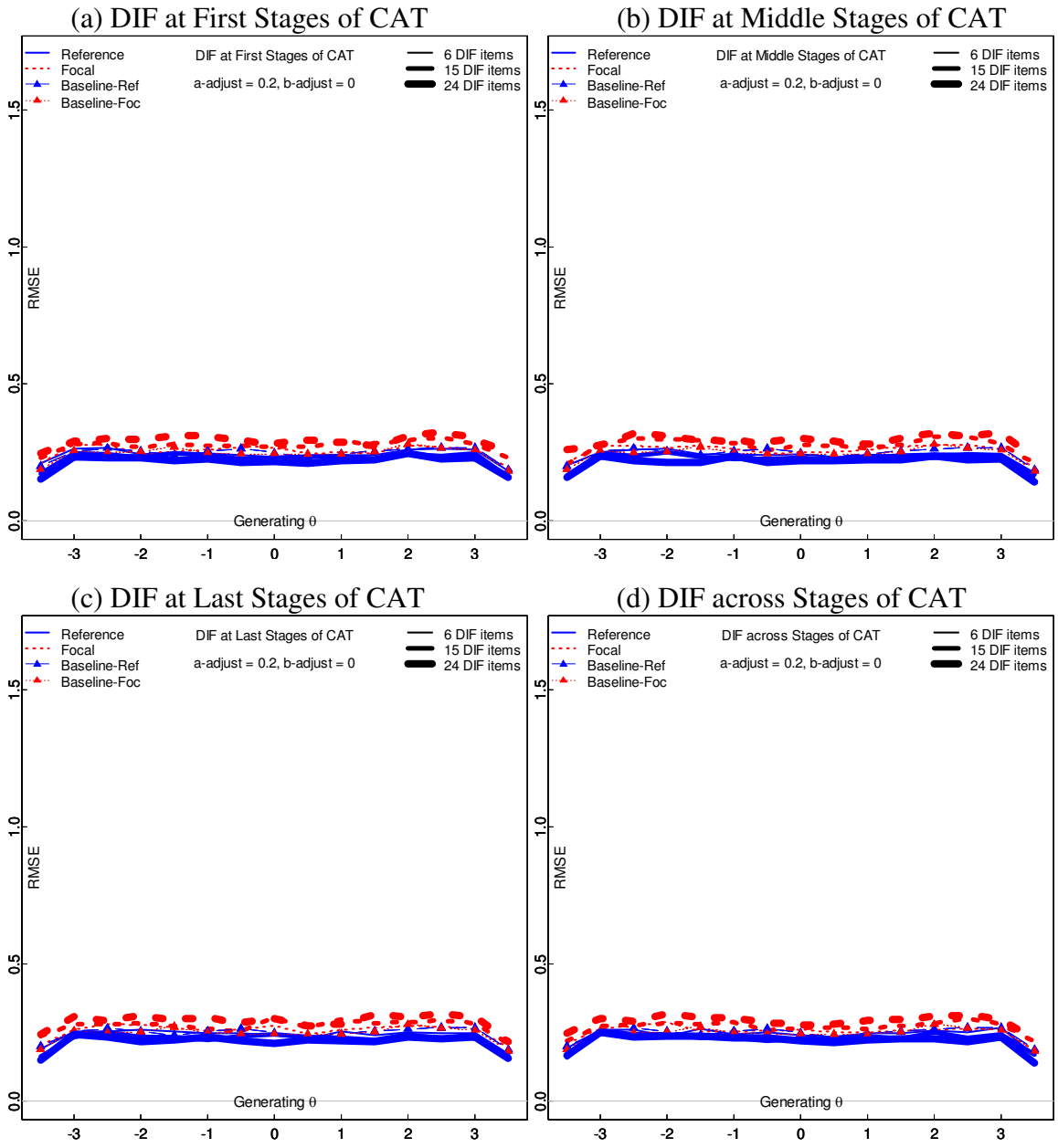
Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

3.3.2 RMSE when DIF items exhibited only nonuniform DIF

Unlike the case of uniform DIF, the values of RMSE for the focal group examinees were apparently larger than those for the reference group examinees for nonuniform DIF. As shown in Figures 22-24, the RMSE for the reference group was lower than for the focal group as the magnitude of nonuniform DIF increased. The RMSE for the focal group, on the other hand, changed dramatically when the magnitude of nonuniform DIF and number of DIF items increased. As seen in Figure 24, the plot of RMSE for the focal group almost formed a U-shape. This means that among examinees from the focal group, those with extreme ability levels had larger RMSE than those in the middle of the ability scale.

Regarding the effect of DIF occurrence, the RMSE obtained when DIF occurred at different stages of CAT tended to be stable when the magnitude was moderate to large in size. However, when the magnitude of uniform DIF was 1.6 (Figure 24), there seemed to be an interaction between the effect of DIF occurrence and generating θ. Specifically, when the largest nonuniform DIF occurred at the first stages of CAT (Figure 24a), the RMSE of the focal group examinees with θ at the lower and upper ends of the ability scale was larger than that of the average-ability examinees. When this magnitude of nonuniform DIF occurred at the last stages of CAT (Figure 24c), the RMSE for focal group examinees was unexpectedly stable across θ levels.

Figure 22. RMSE of $\hat{\theta}_{CAT}$ when the magnitude of nonuniform DIF was .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



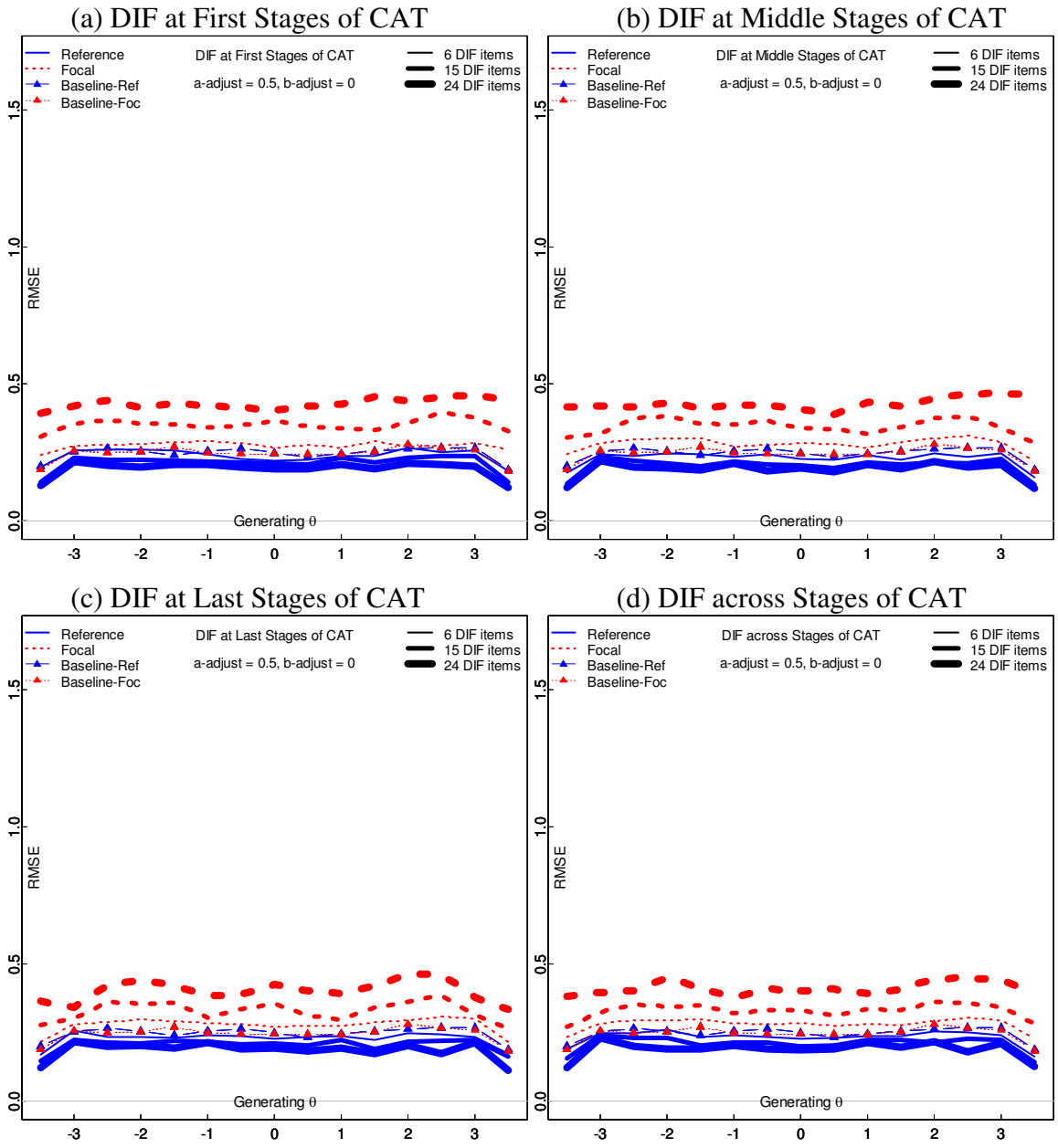
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .2$ and $a_{\text{focal}} = a_{\text{bank}} - .2$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, b-adjust = 0, meaning that $b_{\text{reference}} = b_{\text{focal}} = b_{\text{bank}}$ or uniform DIF did not occur.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 23. RMSE of $\hat{\theta}_{CAT}$ when the magnitude of nonuniform DIF was 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



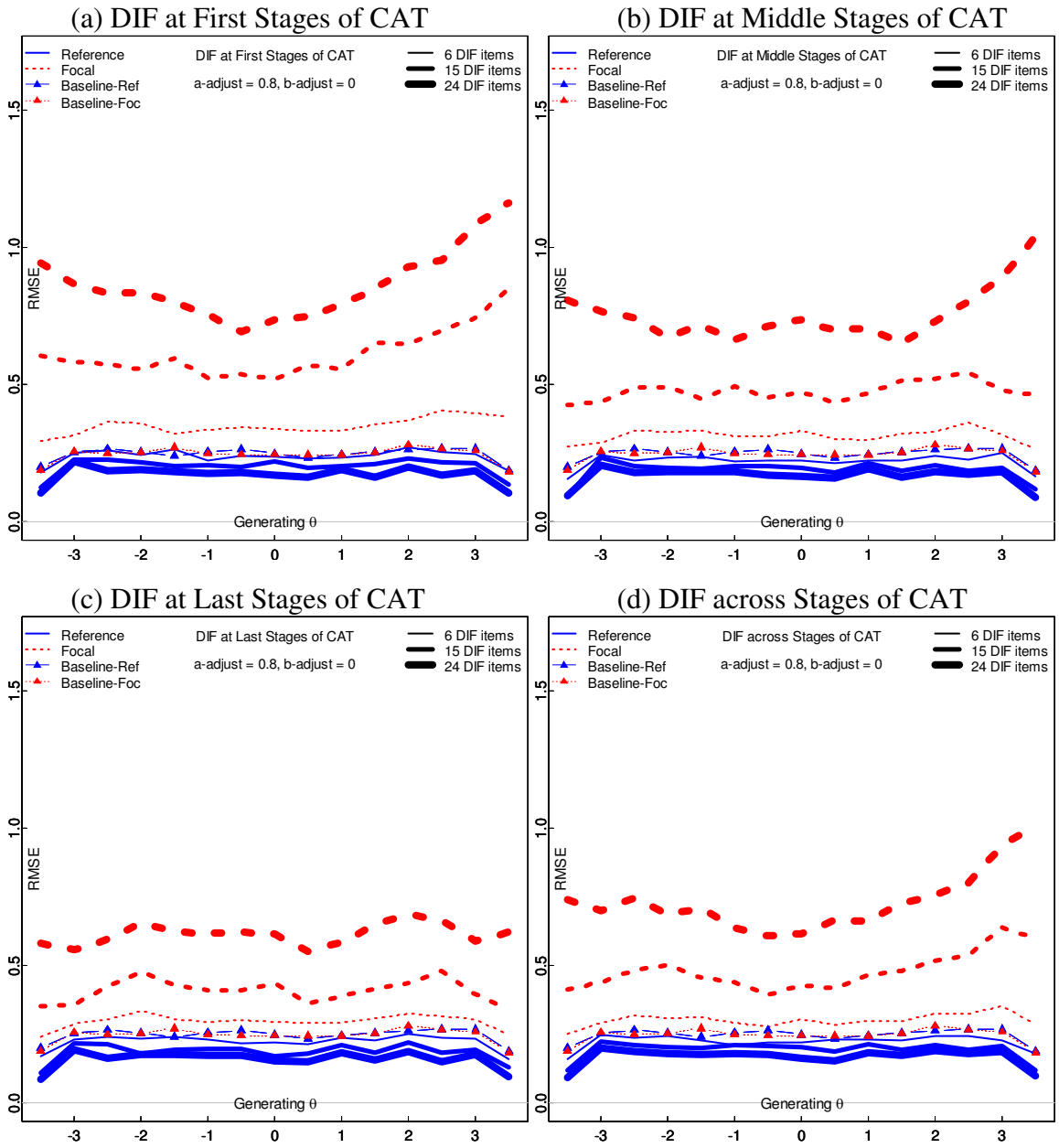
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .5$ and $a_{\text{focal}} = a_{\text{bank}} - .5$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, b-adjust = 0, meaning that $b_{\text{reference}} = b_{\text{focal}} = b_{\text{bank}}$ or uniform DIF did not occur.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 24. RMSE of $\hat{\theta}_{CAT}$ when the magnitude of nonuniform DIF was 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .8$ and $a_{\text{focal}} = a_{\text{bank}} - .8$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, b-adjust = 0, meaning that $b_{\text{reference}} = b_{\text{focal}} = b_{\text{bank}}$ or uniform DIF did not occur.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

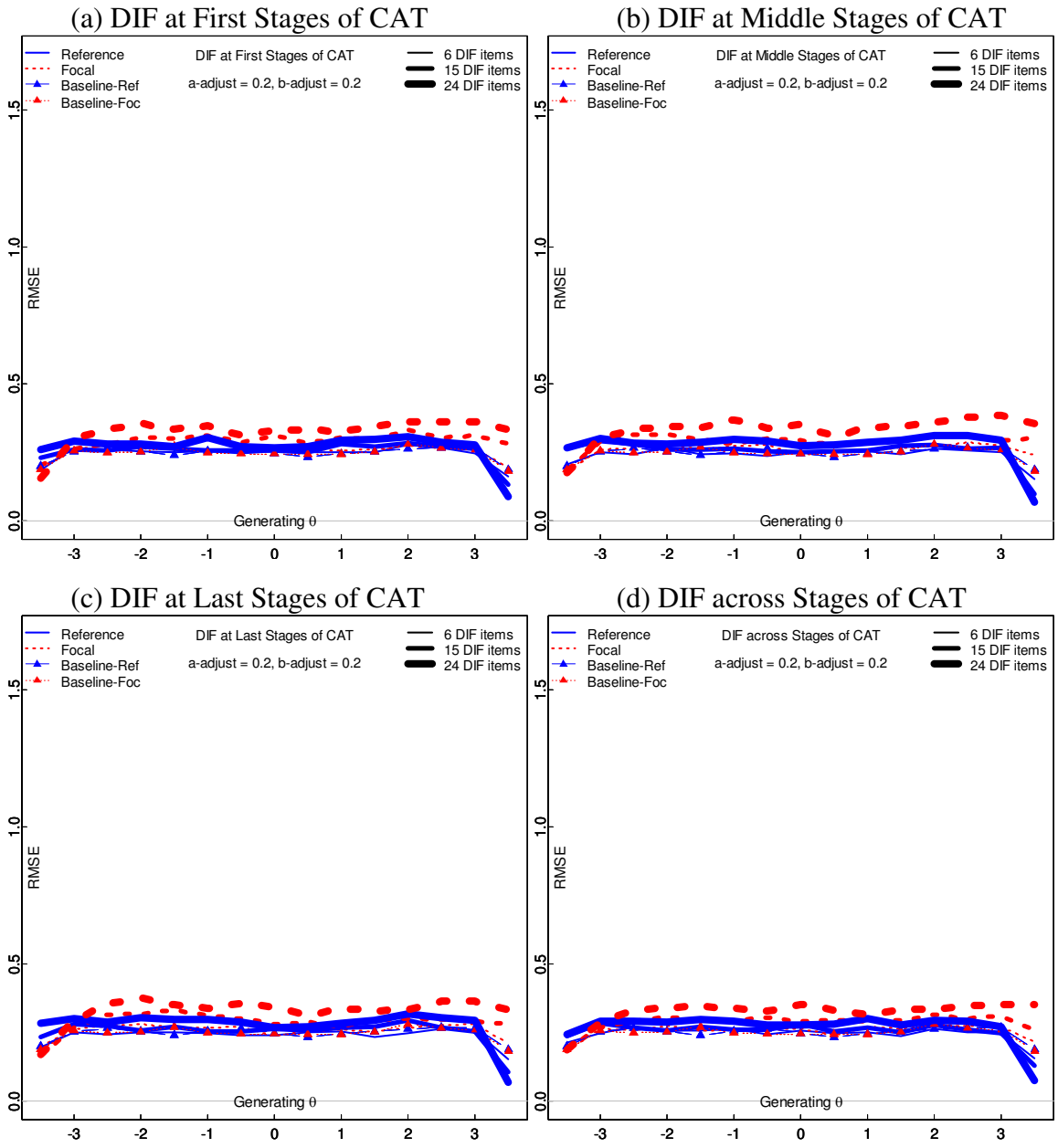
3.3.3 RMSE when DIF items exhibited both types of DIF with the same magnitude

With the smallest magnitude of both types of DIF (Figure 25), the RMSE was stable across numbers of DIF items and stages of CAT that DIF occurred except for the extreme ability levels. In general, examinees from the focal group had higher RMSE for $\theta = -3$ to 3.5. On the other hand, the focal group examinees had lower RMSE at $\theta = -3.5$. Another general observation is that when the number of DIF items increased, the RMSE also increased. However, when the magnitude of DIF changed from .4 to 1.0 and 1.6, the trends of RMSE for each group changed mostly depending on the generating θ.

Particularly when the magnitude of nonuniform and uniform DIF was 1.0 (Figure 26), the focal group examinees with θ between –2 to 2 were likely to have larger RMSE than the reference group with the same θ. In addition, the focal group examinees with very high ability levels tended to have the highest RMSE (.6-.7) compared to other examinees. Finally, the effect of DIF contamination on RMSE was similar as discussed in other simulation conditions. That is, 24 DIF items typically yielded larger RMSE than 6 and 15 DIF items.

Finally, Figure 27 shows the interaction effect among simulation factors. As seen in the figure, the pattern of RMSE varied across the θ level, number of DIF items, stages of CAT that DIF occurred, and group membership. For example, when the magnitudes of nonuniform and uniform DIF were 1.6, the RMSE for the focal group was larger than that for the reference group if 24 DIF items were delivered to extremely high-ability examinees at the first stages of CAT. However, both groups had comparable RMSE when 24 DIF items were administered in the last stages of CAT for examinees with θ between –2 and 2. Lastly, the difference in RMSE between high-ability examinees from both groups was lower when DIF items were administered at the last stages of CAT than other stages.

Figure 25. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were .4 and .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



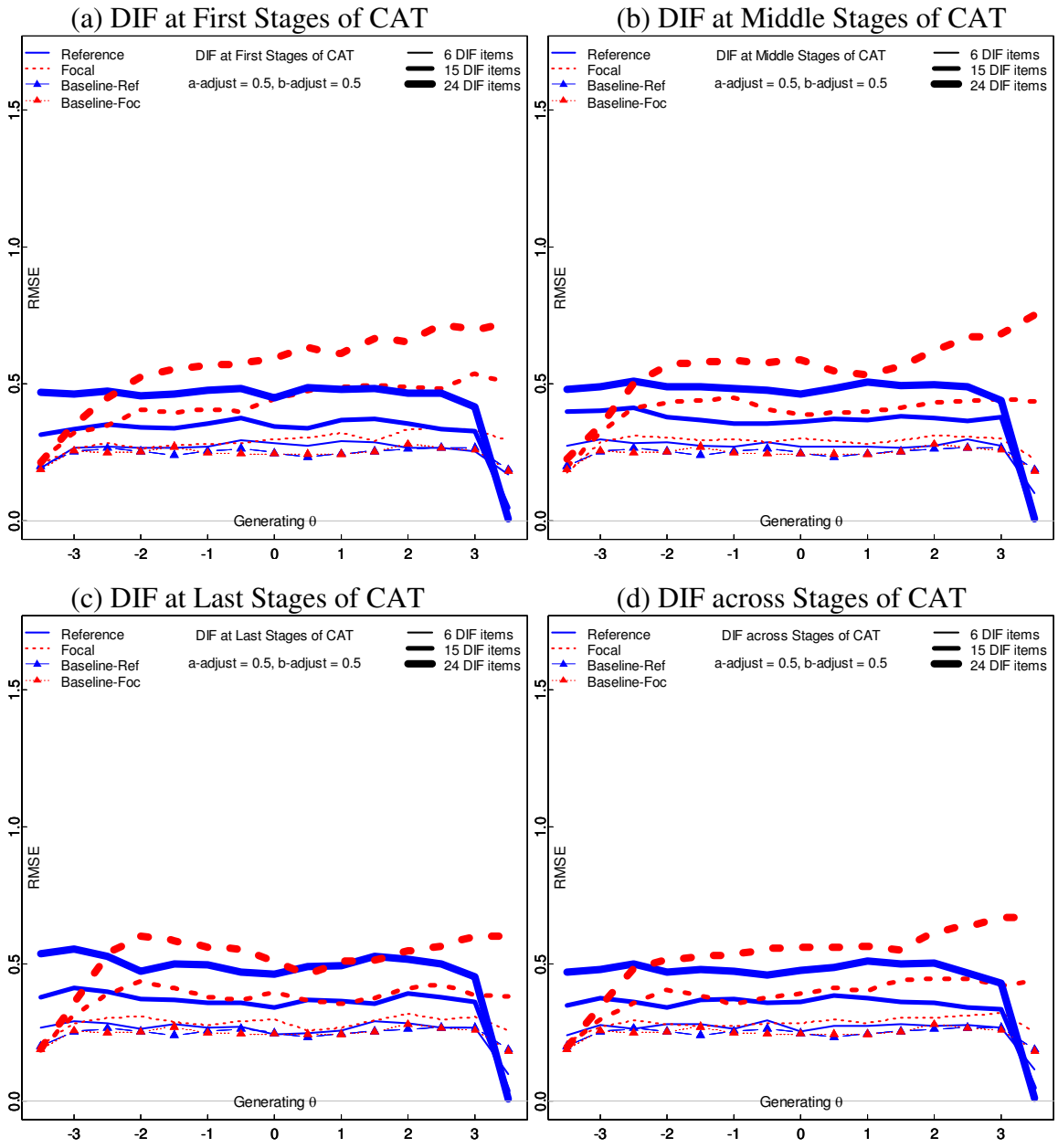
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .2$ and $a_{\text{focal}} = a_{\text{bank}} - .2$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .2$ and $b_{\text{focal}} = b_{\text{bank}} + .2$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 26. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.0 and 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



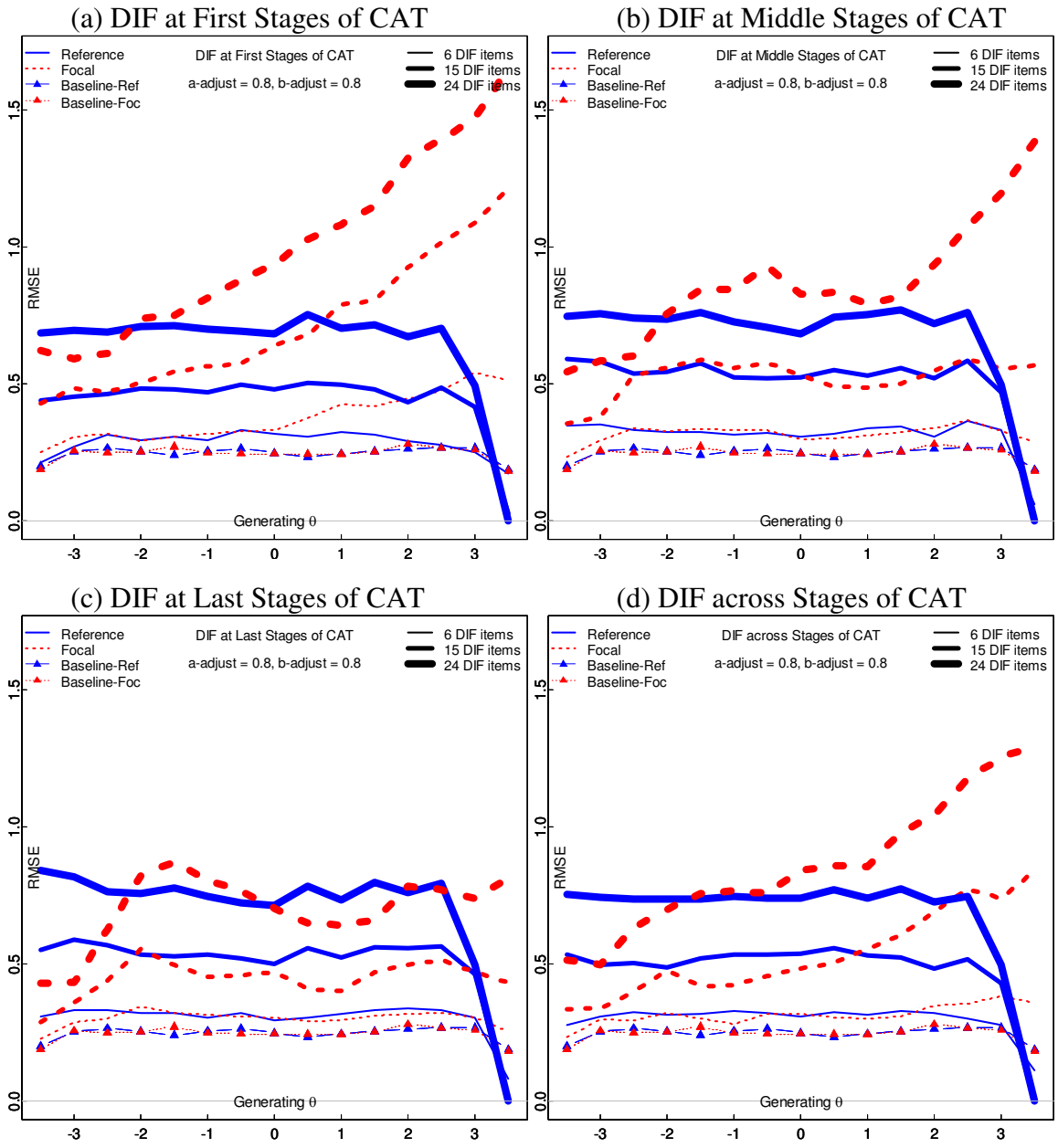
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .5$ and $a_{focal} = a_{bank} - .5$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .5$ and $b_{focal} = b_{bank} + .5$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 27. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.6 and 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .8$ and $a_{\text{focal}} = a_{\text{bank}} - .8$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .8$ and $b_{\text{focal}} = b_{\text{bank}} + .8$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

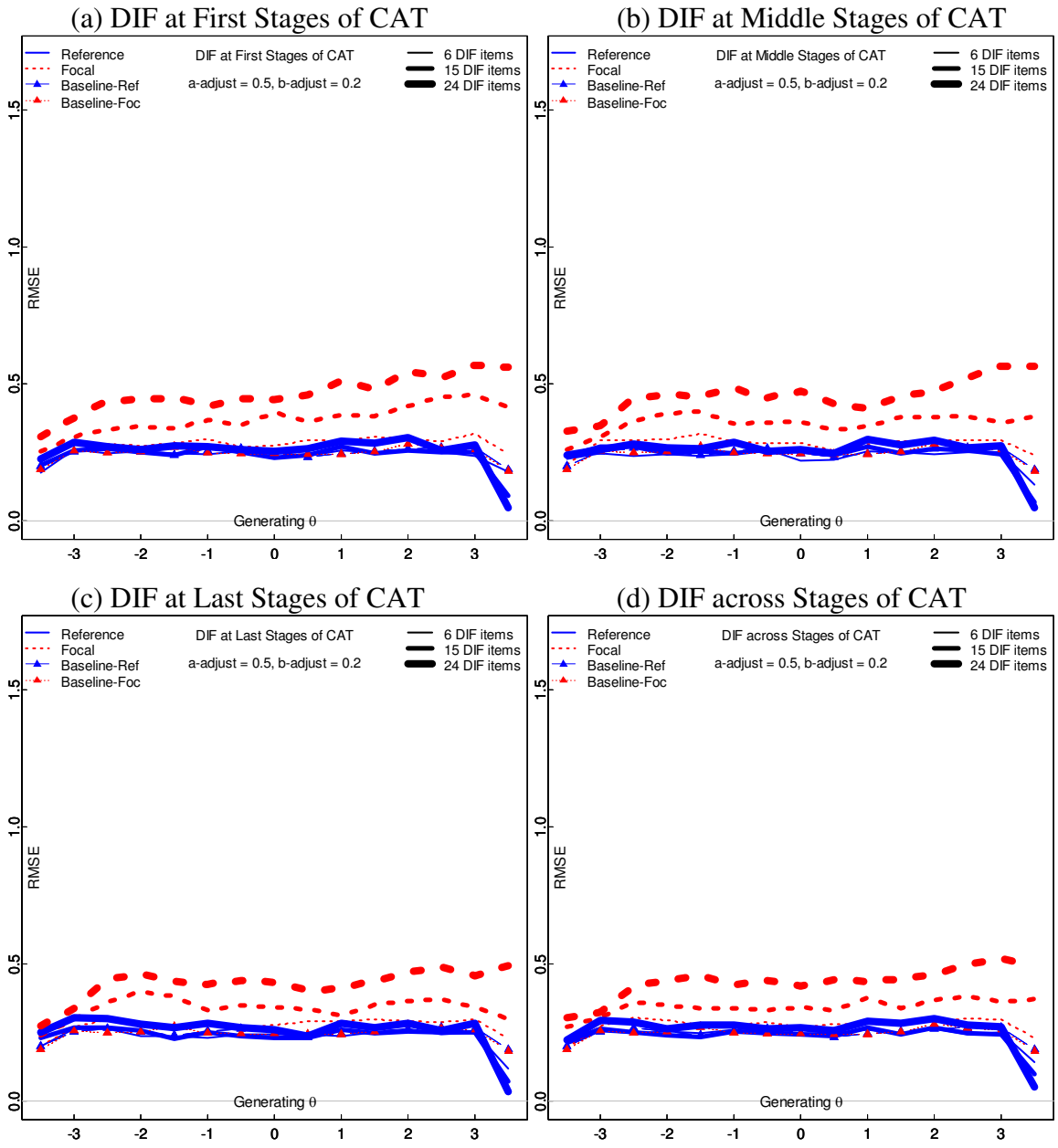
Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

3.3.4 RMSE when DIF items exhibited both types of DIF with different magnitudes

Basically, if the magnitude of nonuniform DIF was greater than the magnitude of uniform DIF, the effect of nonuniform DIF suppressed the effect of uniform DIF. As seen in Figures 28-30, the magnitudes of nonuniform DIF were higher than uniform DIF. The RMSE of $\hat{\theta}_{CAT}$ for the focal group examinees dramatically increased as the magnitude of nonuniform DIF, number of DIF item, and generating θ increased. As observed in previous cases of nonuniform DIF, the largest magnitude of nonuniform DIF, regardless of uniform DIF magnitude, yielded the result that CAT with DIF items at the last stages provided lower RMSE for the focal group examinees than CAT with DIF items at other stages (Figures 29c and 30c).

In contrast, the RMSE of $\hat{\theta}_{CAT}$ shown in Figures 31-33 consistently changed for both groups of examinees with average θ levels because the magnitudes of uniform DIF were larger than nonuniform DIF. In addition, as the magnitude of uniform DIF increased, despite the change in other factors, the difference in RMSE of $\hat{\theta}_{CAT}$ between groups consistently changed across the middle θ levels (–2 to 2). As θ increased and/or the number of DIF items increased, such a difference apparently increased. These patterns of RMSE were observed in all conditions of DIF occurrence.

Figure 28. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.0 and .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



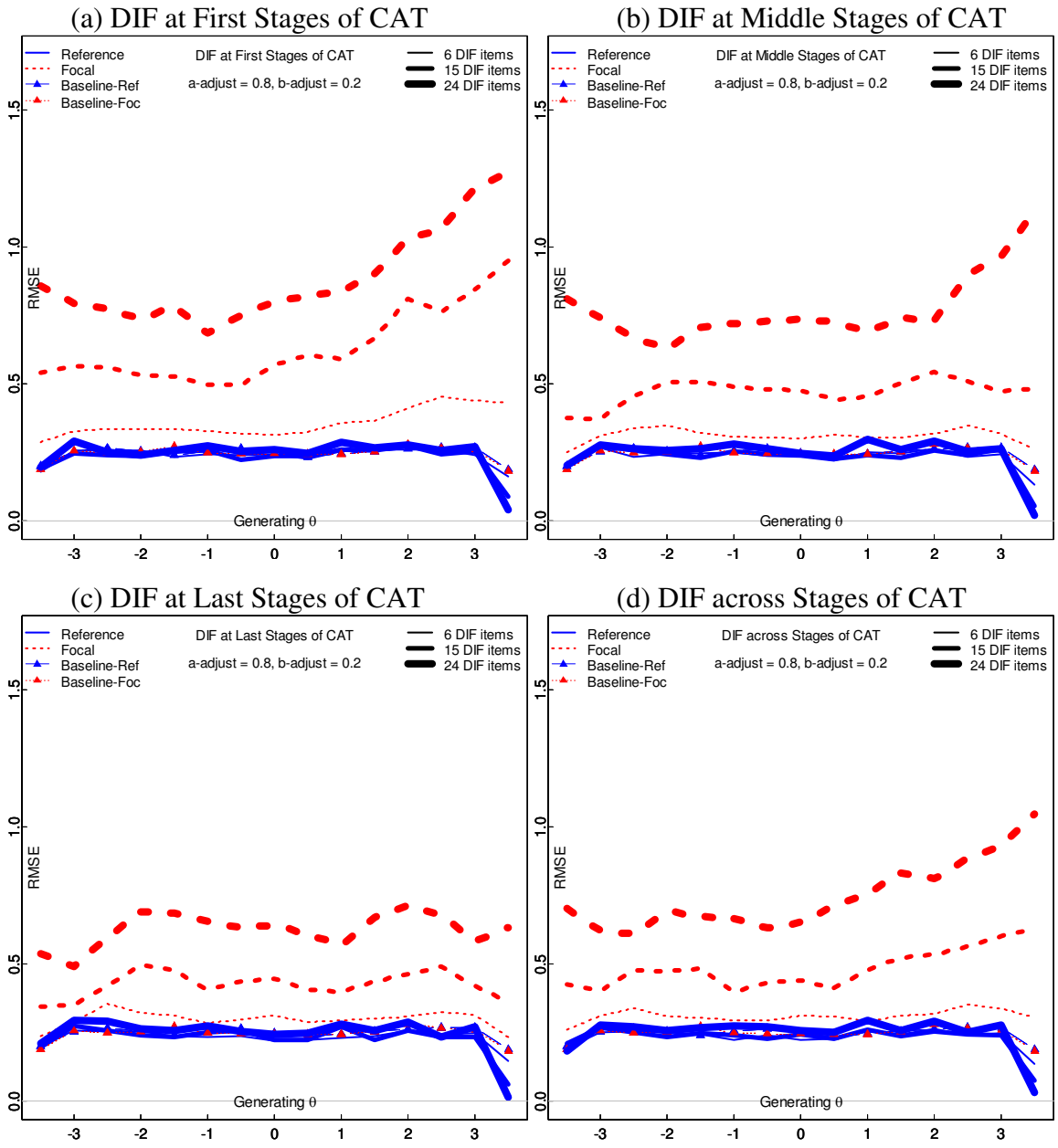
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .5$ and $a_{focal} = a_{bank} - .5$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .2$ and $b_{focal} = b_{bank} + .2$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 29. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.6 and .4 across the conditions of generating θ, DIF contamination, and DIF occurrence.



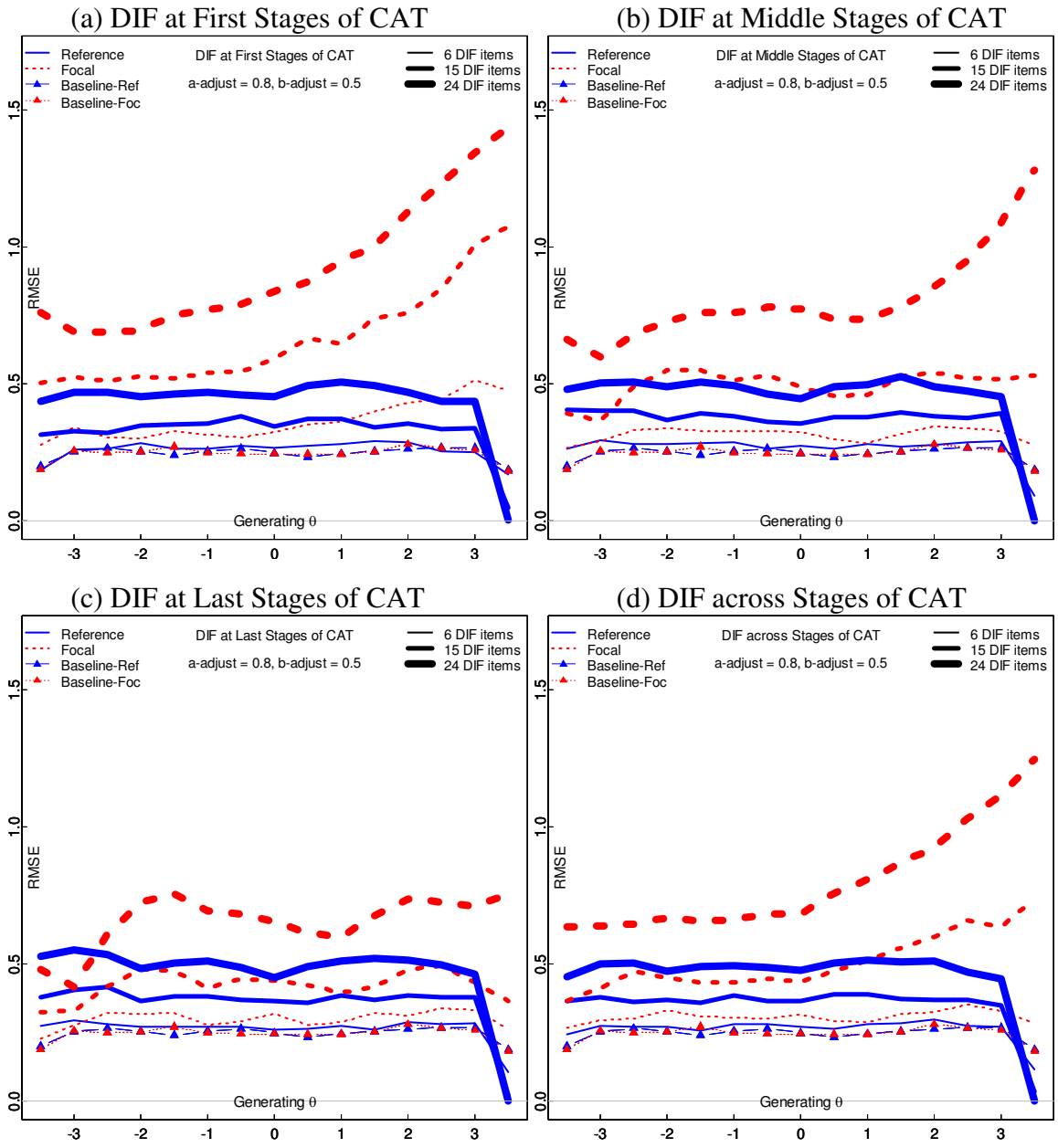
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .8$ and $a_{\text{focal}} = a_{\text{bank}} - .8$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .2$ and $b_{\text{focal}} = b_{\text{bank}} + .2$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = .4$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 30. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.6 and 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



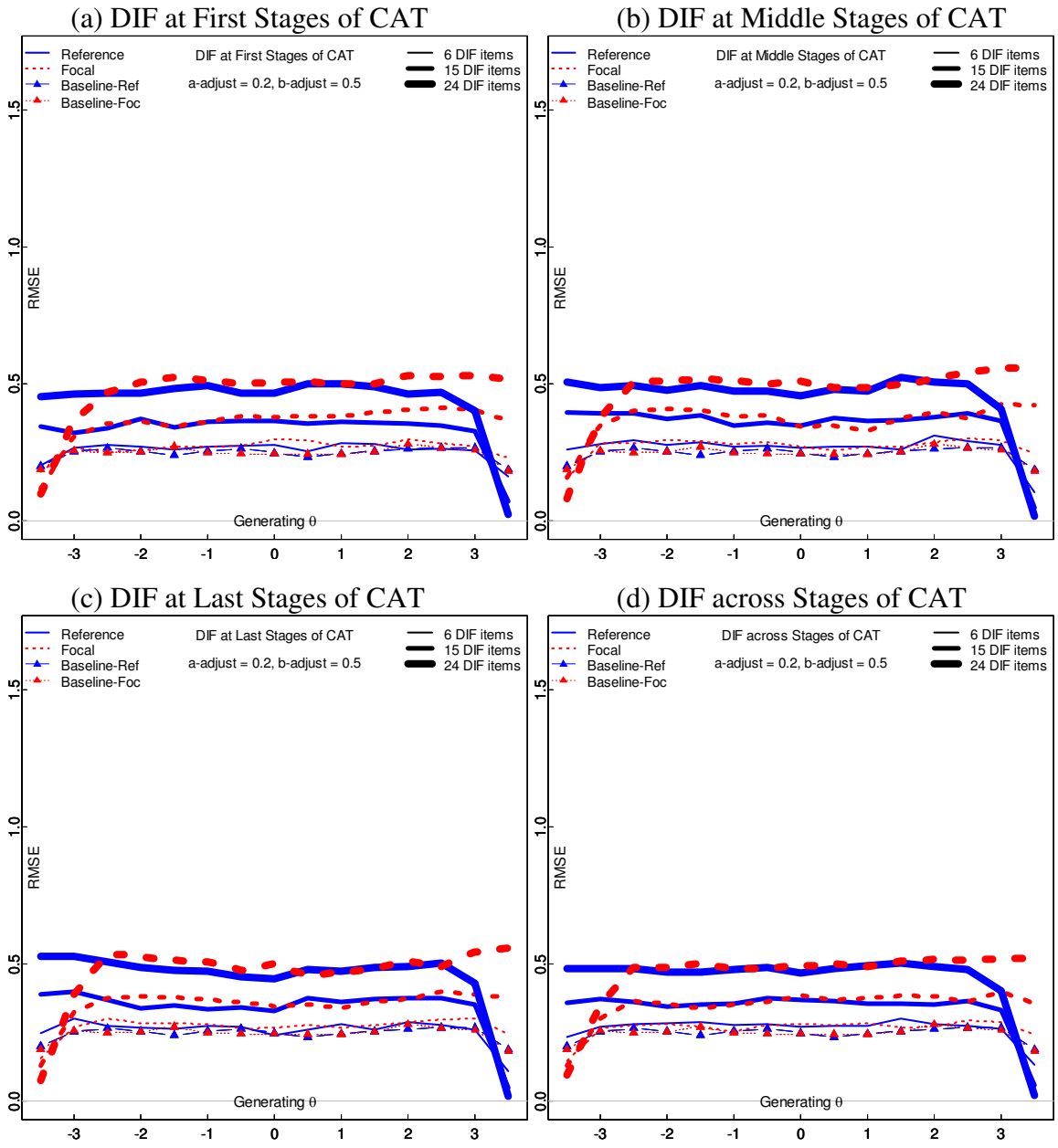
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .8$ and $a_{focal} = a_{bank} - .8$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1.6$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .5$ and $b_{focal} = b_{bank} + .5$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 31. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were .4 and 1.0 across the conditions of generating θ, DIF contamination, and DIF occurrence.



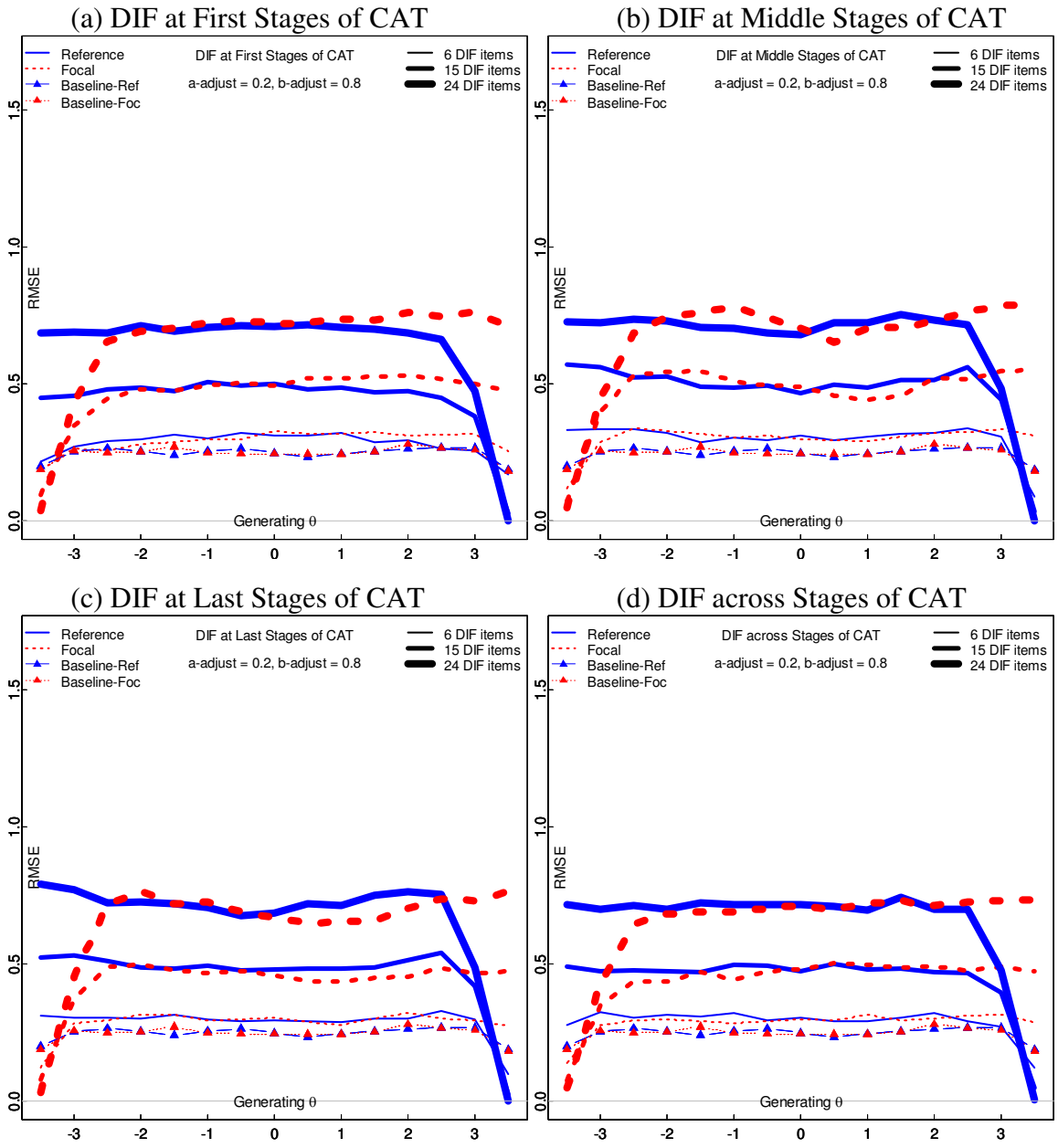
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .2$ and $a_{\text{focal}} = a_{\text{bank}} - .2$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .5$ and $b_{\text{focal}} = b_{\text{bank}} + .5$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = 1$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 32. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were .4 and 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



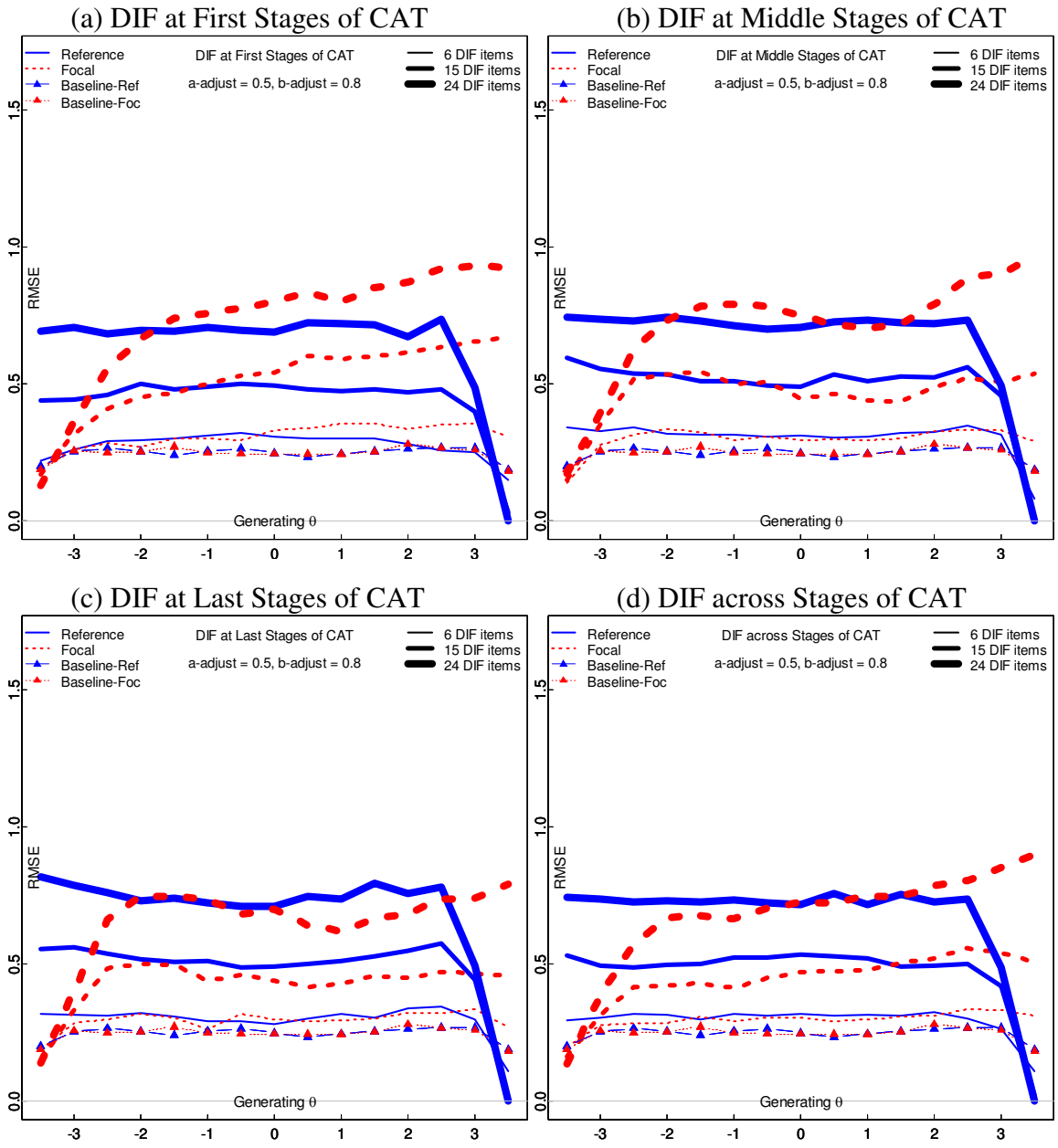
Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{\text{reference}} = a_{\text{bank}} + .2$ and $a_{\text{focal}} = a_{\text{bank}} - .2$, yielding the magnitude of nonuniform DIF = $|a_{\text{reference}} - a_{\text{focal}}| = .4$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{\text{reference}} = b_{\text{bank}} - .8$ and $b_{\text{focal}} = b_{\text{bank}} + .8$, yielding the magnitude of uniform DIF = $|b_{\text{reference}} - b_{\text{focal}}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

Figure 33. RMSE of $\hat{\theta}_{CAT}$ when the magnitudes of nonuniform and uniform DIF were 1.0 and 1.6 across the conditions of generating θ, DIF contamination, and DIF occurrence.



Note: a-adjust = the constant used to adjust the initial discrimination (a_{bank}) to obtain the discrimination of DIF items for the reference and focal groups. Here, $a_{reference} = a_{bank} + .5$ and $a_{focal} = a_{bank} - .5$, yielding the magnitude of nonuniform DIF = $|a_{reference} - a_{focal}| = 1$.

b-adjust = the constant used to adjust the initial difficulty (b_{bank}) to obtain the difficulty of DIF items for the reference and focal groups. Here, $b_{reference} = b_{bank} - .8$ and $b_{focal} = b_{bank} + .8$, yielding the magnitude of uniform DIF = $|b_{reference} - b_{focal}| = 1.6$.

Baseline-Ref = BIAS for the reference group from the baseline or DIF-free CAT condition.

Baseline-Foc = BIAS for the focal group from the baseline or DIF-free CAT condition.

3.4 SE of $\hat{\theta}_{CAT}$

Unlike the BIAS and RMSE that were estimated using the generating θ and $\hat{\theta}_{CAT}$, the empirical SE was solely based on $\hat{\theta}_{CAT}$. Because the generating or true θ is unknown in practice, the empirical SE was the only index that can be computed based solely on $\hat{\theta}_{CAT}$, and expected to signal DIF in operational items in live CAT. Therefore, the SE of $\hat{\theta}_{CAT}$ obtained from each operational item are presented in this section, rather than presenting only the marginal values at the end of CAT as done in the BIAS and RMSE sections. Due to a large number of SE values as a multiplicative factor of the number of operational items, θ points, groups, and the four simulation factors (30×15×2×180), only the results of the simulation conditions for θ at ±2.5 are presented. The results for other θ's also show this similar trend.

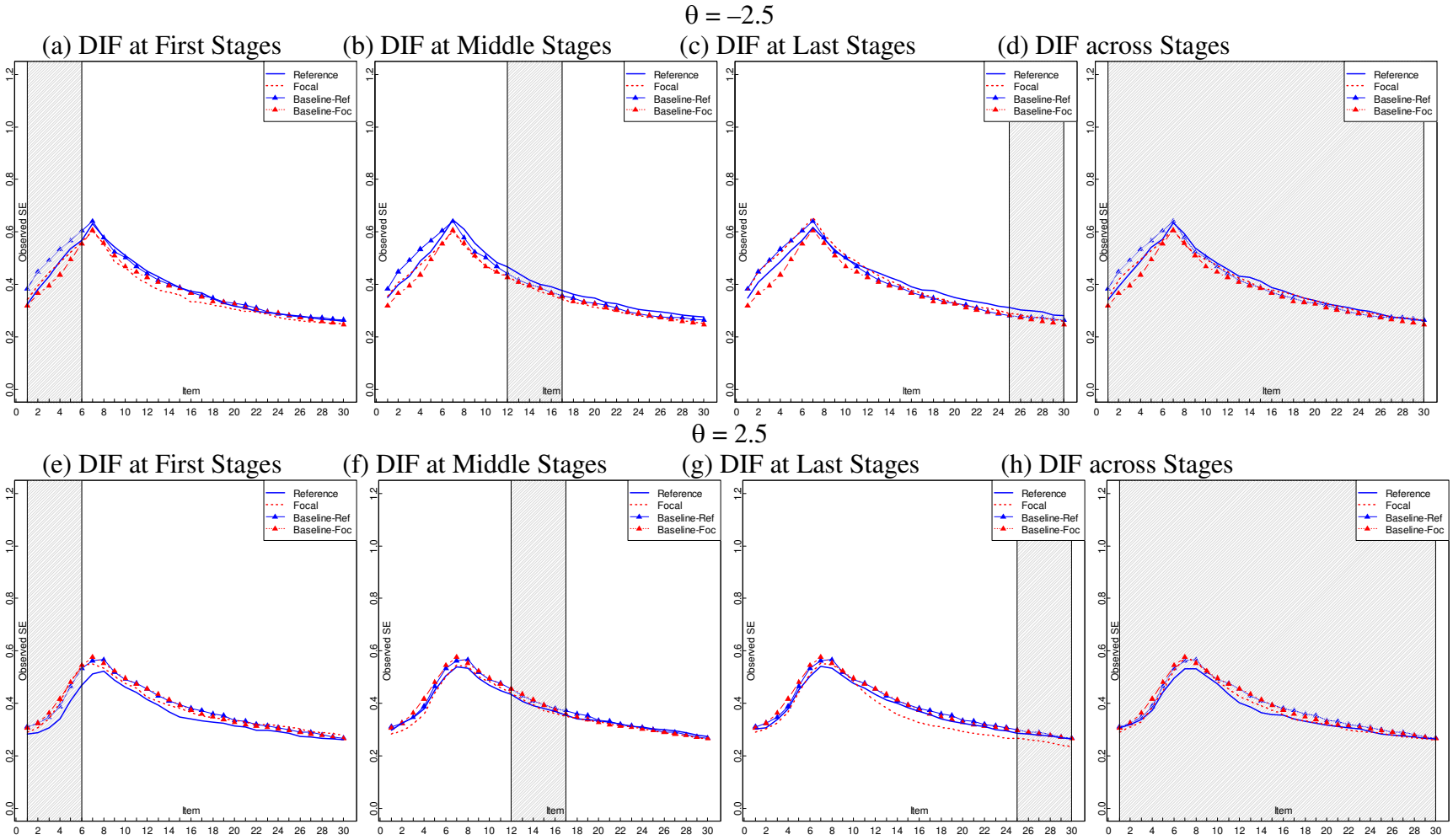
It should be noted that an increase in SE for the first few items was due to the fact that CAT was searching for a better ability estimate from the restricted information obtained from the first few items of CAT. As seen in the baseline condition (i.e., the DIF-free CAT condition), the increase of SE stopped around Item 6 and gradually declined to .2-.3 as the number of administered items increased.

3.4.1 SE when DIF items exhibited only uniform DIF

Figures 34-35 show the observed SE obtained after each operational item was administered for examinees with $\theta = \pm 2.5$ when CAT delivered 6 and 24 uniform DIF items with the magnitude of .4 at different stages. When items exhibited only the uniform DIF with the magnitude of .4, regardless of the number of DIF items and stages of CAT that DIF occurred, the SE of θ at ±2.5 tended to converge to the baseline SE.

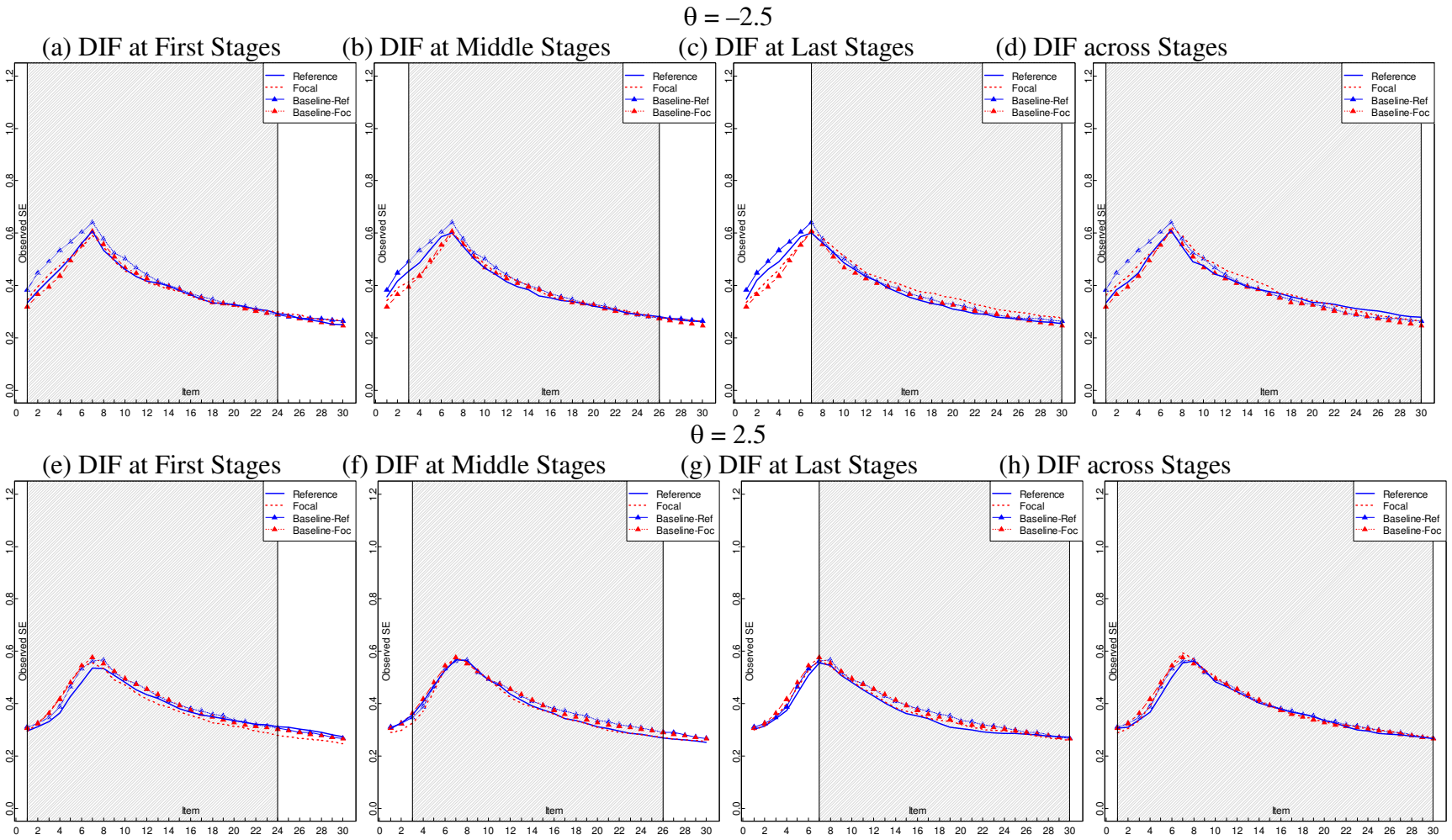
When the magnitude of uniform DIF increased to 1.6, the SE for $\theta = \pm 2.5$ was not greatly affected by DIF effects, even when 24 DIF items were administered. As shown in Figures 36-37, the differences between the SE from CAT with DIF items and the baseline SE were generally small (less than .02), except for the case that 24 uniform DIF items with the magnitude of 1.6 occurred at the beginning or middle stages of CAT (as seen in Figures 37a, 37b, 37e, and 37f).

Figure 34. Observed SE for θ = ±2.5 when 6 items showed uniform DIF with a magnitude of .4.



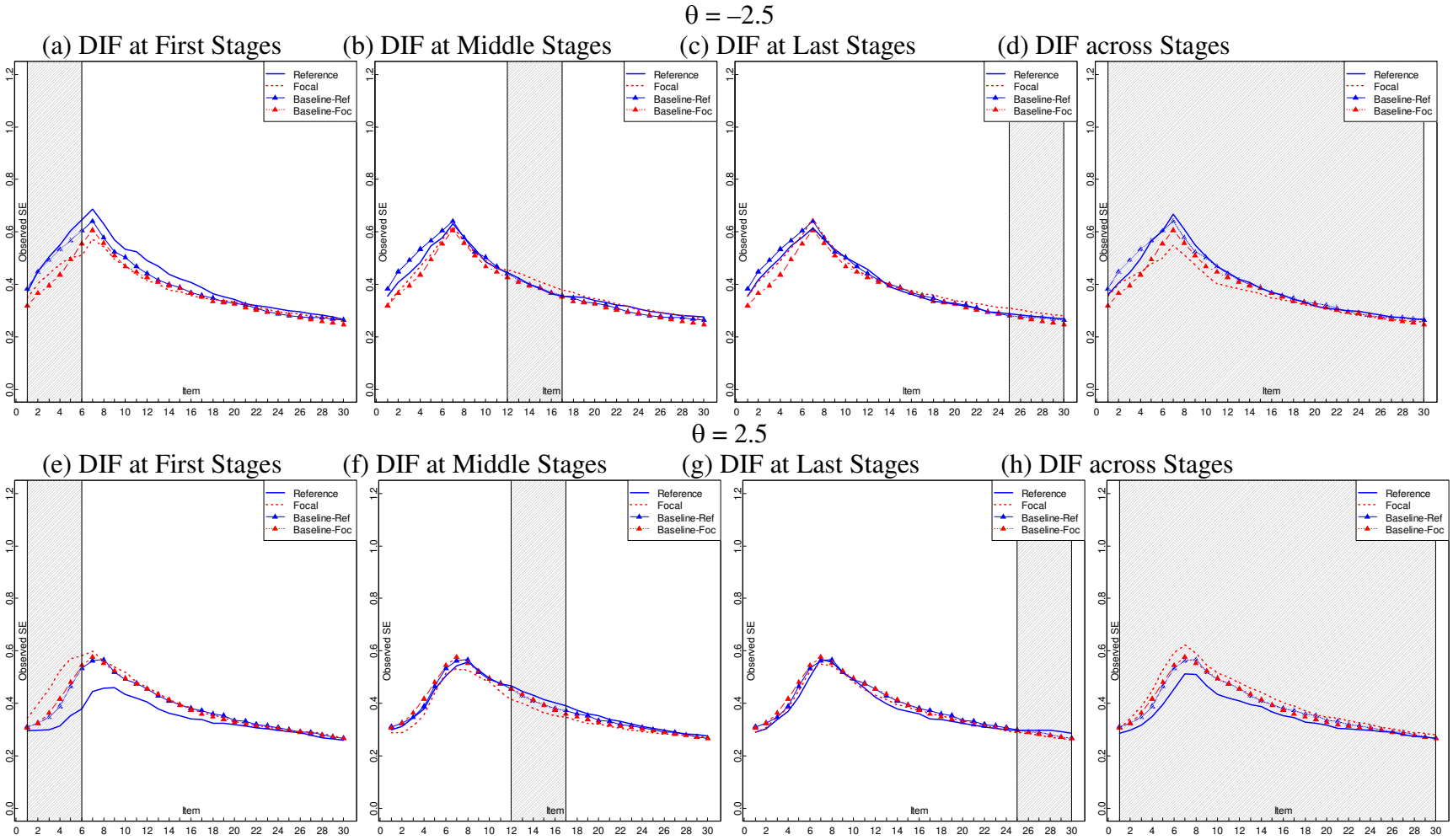
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 35. Observed SE for θ = ±2.5 when 24 items showed uniform DIF with a magnitude of .4.



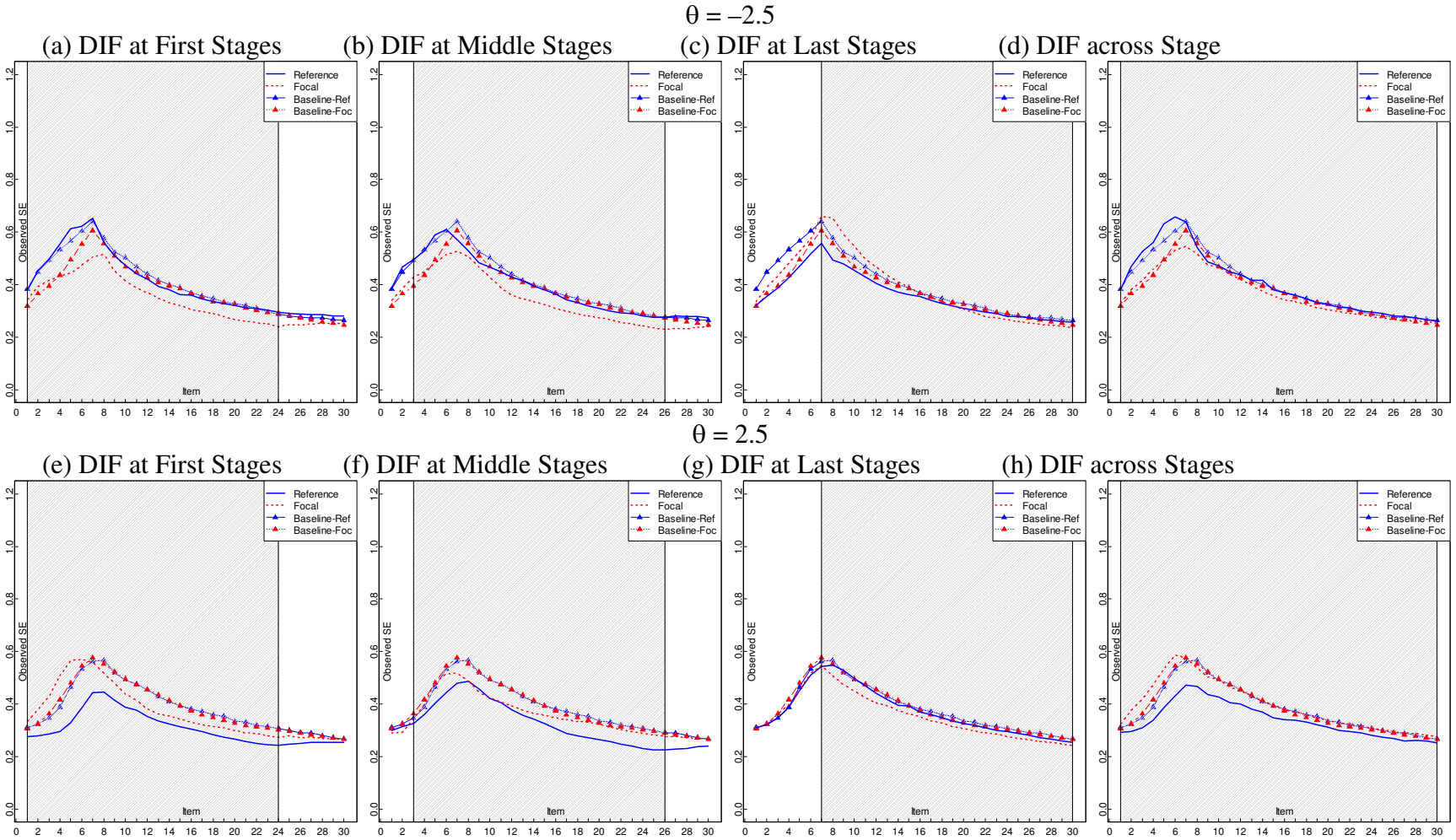
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 36. Observed SE for θ = ±2.5 when 6 items showed uniform DIF with a magnitude of 1.6.



Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 37. Observed SE for θ = ±2.5 when 24 items showed uniform DIF with a magnitude of 1.6.



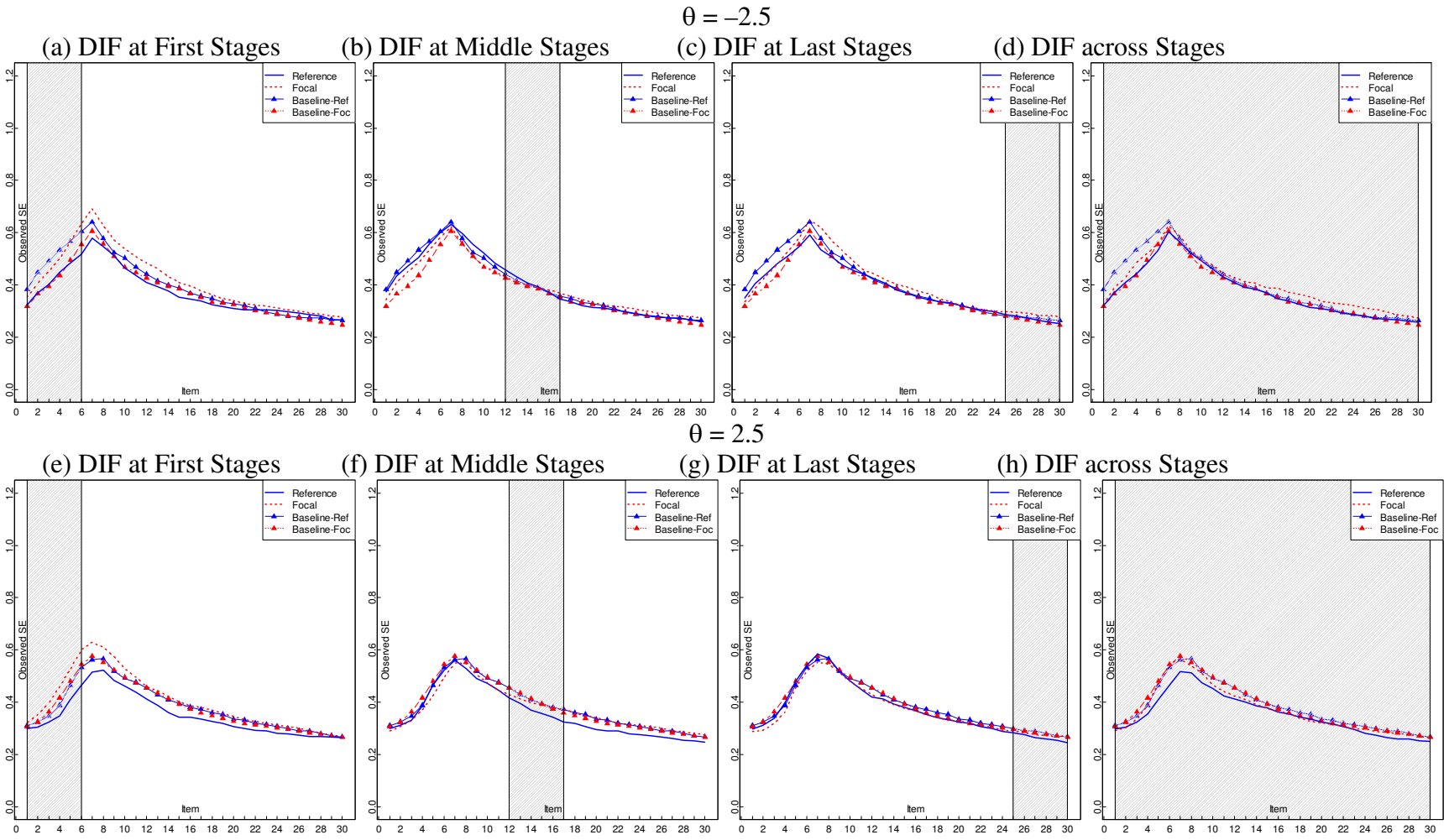
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

3.4.2 SE when DIF items exhibited only nonuniform DIF

The empirical SE values obtained from CAT with some operational items showing only nonuniform DIF are presented in Figures 38-41. With 6 items showing nonuniform DIF with the magnitude of .4 (Figure 38), the SE for $\theta = \pm 2.5$ successfully converged to the baseline SE in most conditions of DIF occurrence. When the number of DIF items with such a magnitude increased from 6 to 24 items (Figure 39), the SE for the reference and focal groups seemed to approach the baseline SE if more items were administered. However, the SE for the focal group was still larger than the baseline SE about .02 – .1 in most levels of DIF occurrence. On the other hand, the SE for the reference group was smaller than the baseline SE about .02 – .1 across DIF occurrence levels.

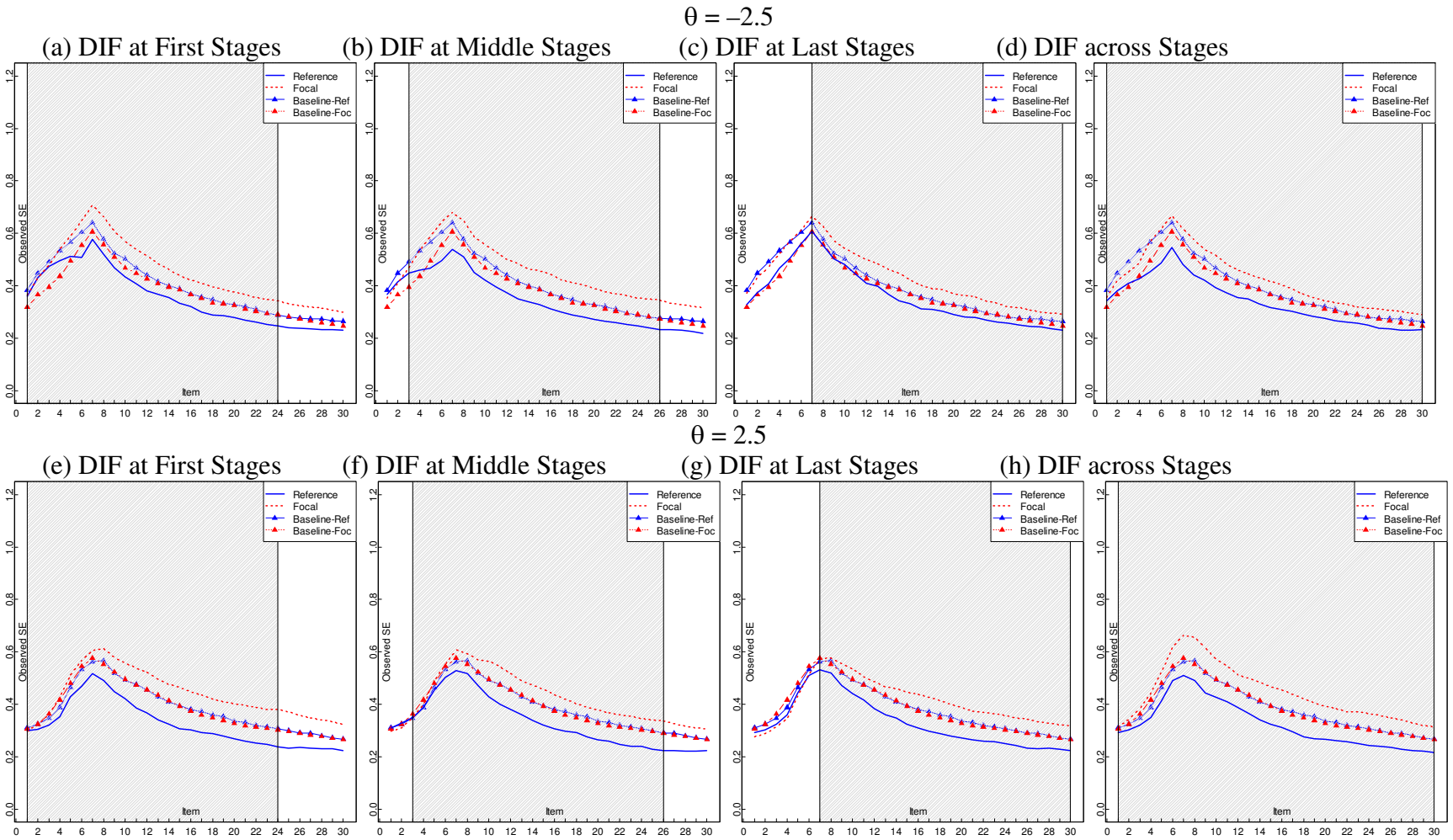
The SE obtained from CAT with some items displaying nonuniform DIF with the magnitude of 1.6 are presented in Figures 40-41. Overall, when nonuniform DIF with the magnitude of 1.6 items were administered, CAT yielded substantially larger SEs for the focal group, but smaller SEs for the reference group. As seen in Figure 40, the SE for the focal group was approximately .02-.4 larger than the baseline SE when DIF items occurred during the first, middle, and across stages of CAT. The SE for the reference group, on the other hand, was slightly smaller than the baseline, around .02-.2, in most of the simulation conditions. However, after 30 items were delivered, the difference in SEs of the focal and reference groups was decreased to .05-.2. Finally, Figure 41 reveals that when 24 items showing nonuniform DIF with the magnitude of 1.6 were administered, the group difference in SE was slightly decreased but still large (approximately .4 to .6) across DIF occurrence levels.

Figure 38. Observed SE for θ = ±2.5 when 6 items showed nonuniform DIF with a magnitude of .4.



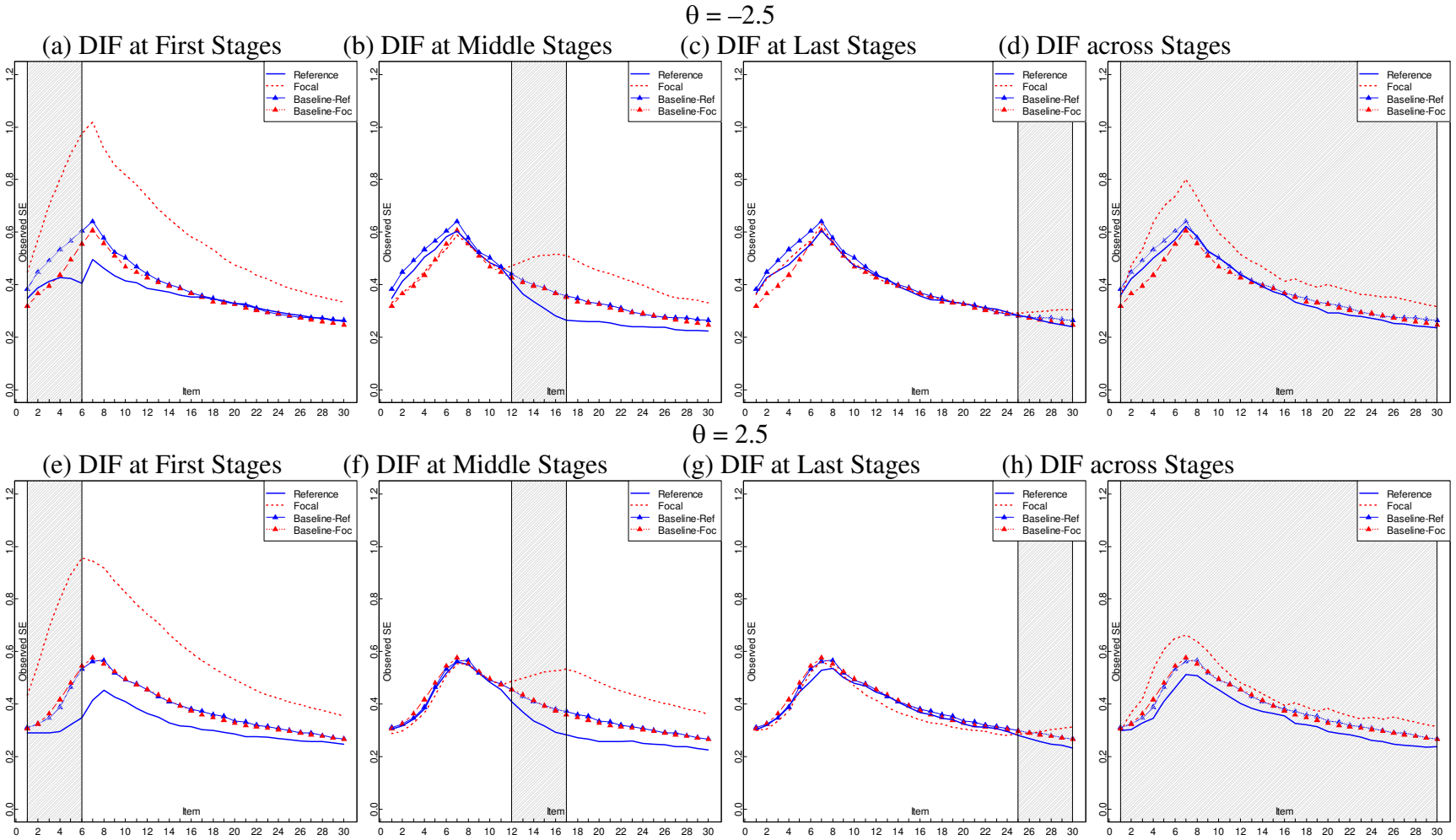
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 39. Observed SE for θ = ±2.5 when 24 items showed nonuniform DIF with a magnitude of .4.



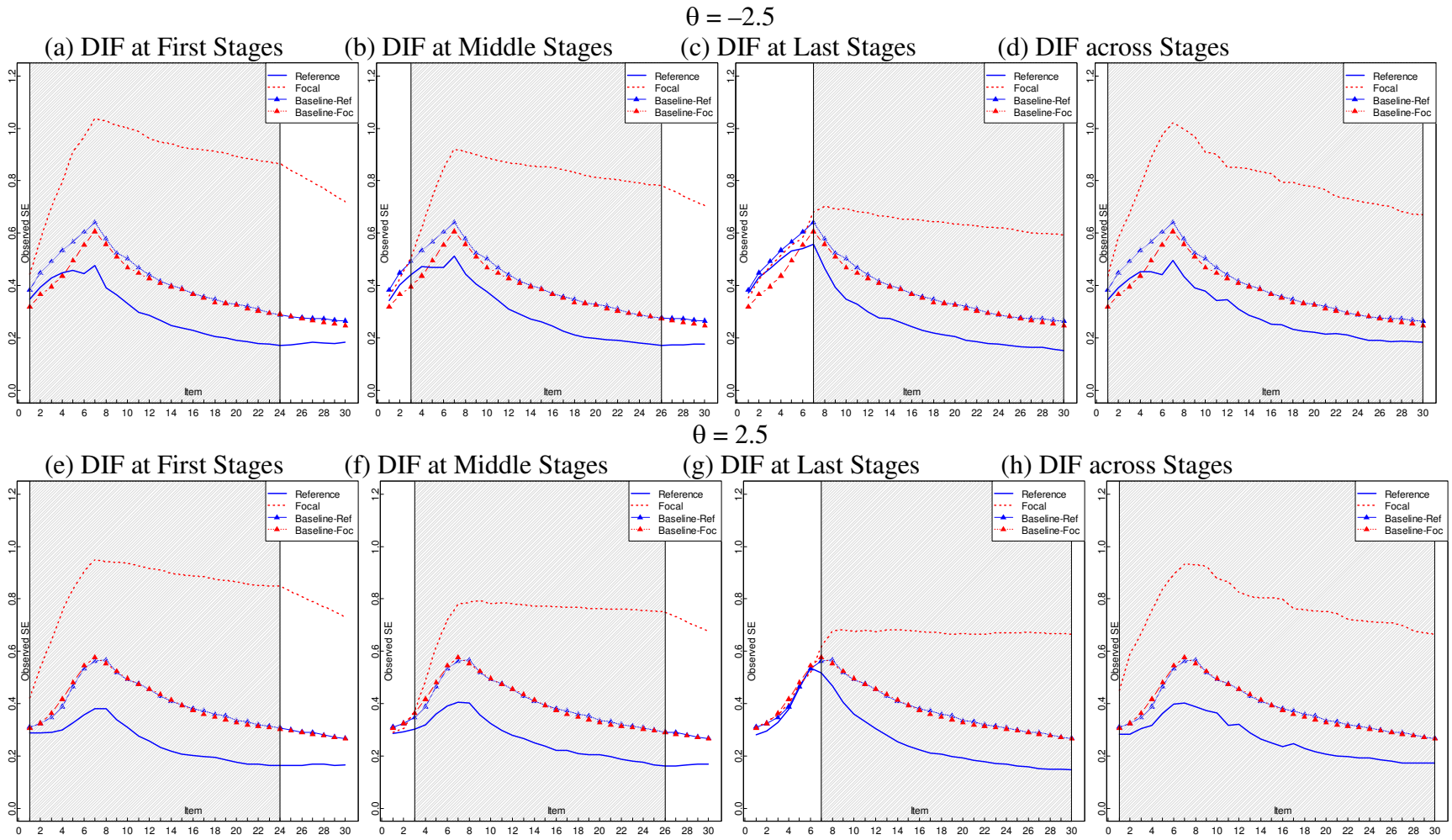
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 40. Observed SE for θ = ±2.5 when 6 items showed nonuniform DIF with a magnitude of 1.6.



Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 41. Observed SE for θ = ±2.5 when 24 items showed nonuniform DIF with a magnitude of 1.6.



Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

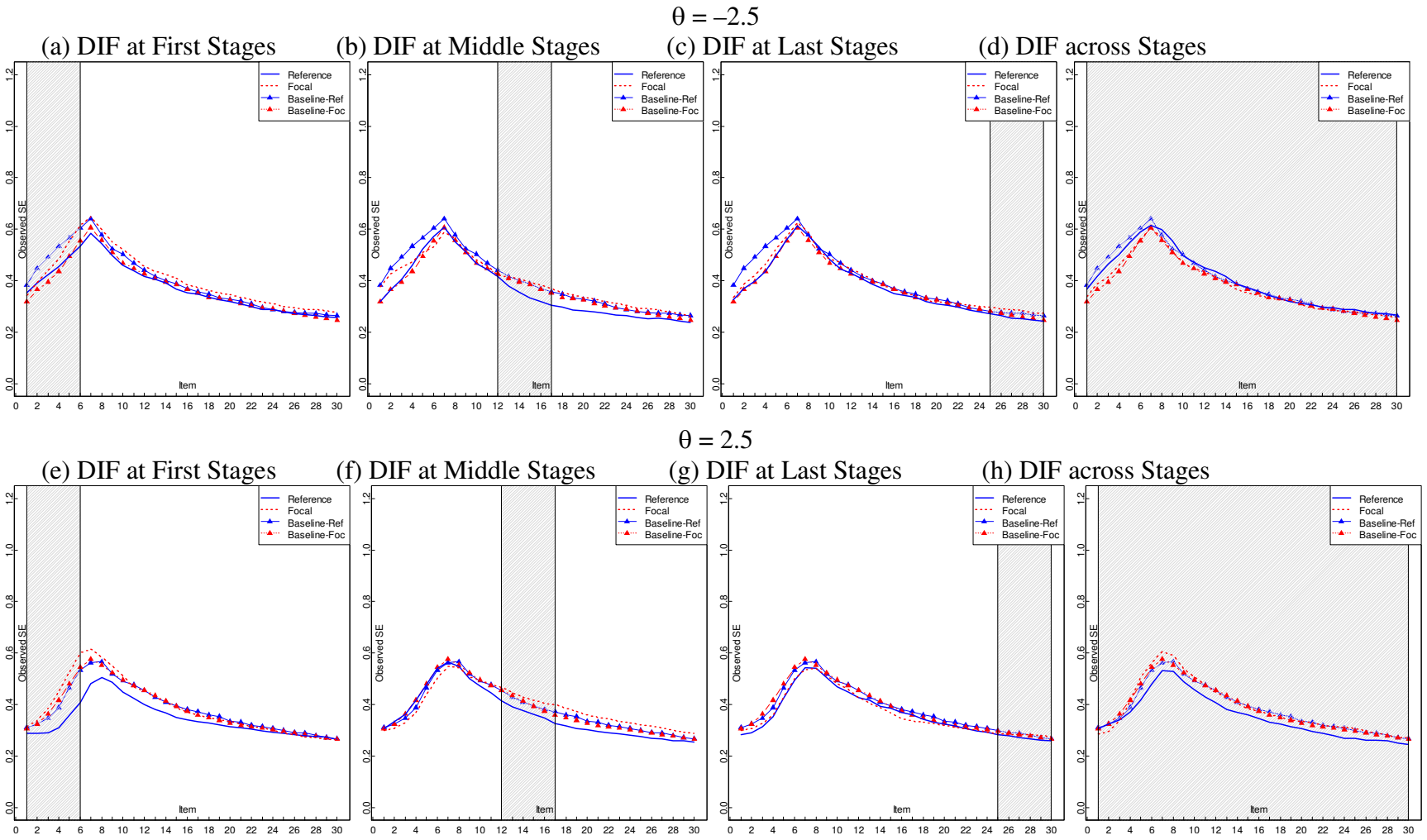
3.4.3 SE when DIF Items Exhibited Both Nonuniform and Uniform DIF with the Same Magnitude

The effect of DIF items on the observed SE seemed to be ignorable even when the magnitude of DIF was moderate (.4), if CAT delivered only a small number of DIF items. In Figure 42, the SE successfully converged to the baseline trend regardless of the generating θ and DIF occurrence. In such cases, the differences between the baseline SE and the SE from CAT with DIF items were almost 0 at the end of CAT. On the other hand, when 24 DIF items with a moderate magnitude were administered (Figure 43), the SE for examinees from both groups differed from the baseline SE about .1.

When DIF items showed both types of DIF in a very large magnitude (1.6), CAT yielded apparently larger SEs for the focal group. As seen in Figure 44, the SE for the focal group was approximately .02-.1 larger than the baseline SE when DIF items occurred during the first, middle, and across stages of CAT. The SE for the reference group, on the other hand, was about the same with the baseline in most of the simulation conditions. After 30 items were delivered, the SE lines of both groups became closer to the baseline. In fact, the SE lines of both groups were about the same at the end of CAT when DIF items occurred at the first and last stages of CAT for $\theta = -2.5$, and when DIF items occurred at the last stage of CAT for $\theta = 2.5$.

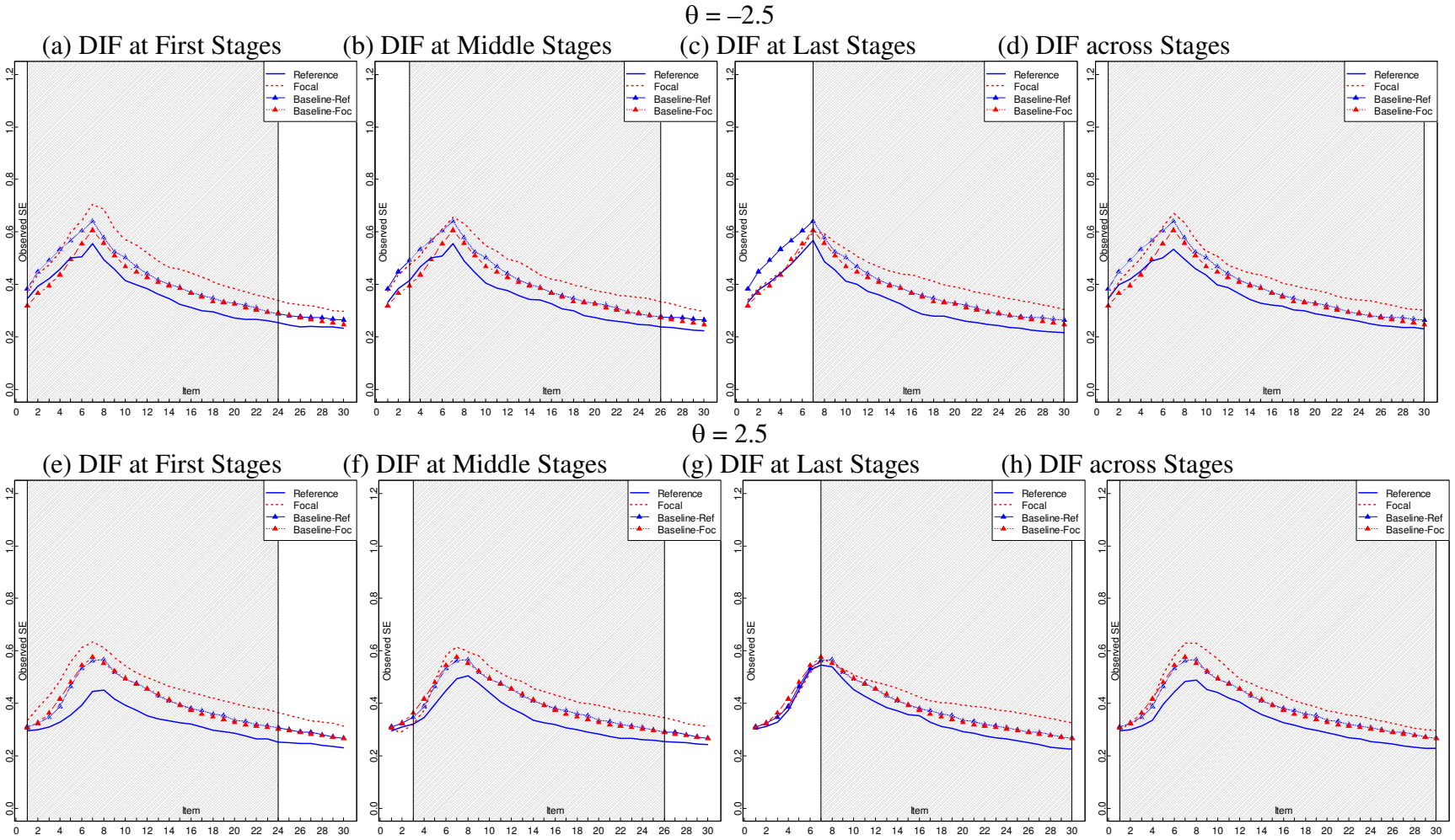
As shown in Figure 45, when 24 items showing both uniform and nonuniform DIF with the magnitude of 1.6 were administered, the group difference in SE was slightly decreased as items 7-30 were administered (with the exception of panel g), but still large (approximately .2 to .6) across DIF occurrence levels. Even after 30 CAT items were delivered, none of the SE lines could meet each other at the baseline SE.

Figure 42. Observed SE for θ = ±2.5 when 6 items showed nonuniform and uniform DIF with magnitudes of .4 and .4.



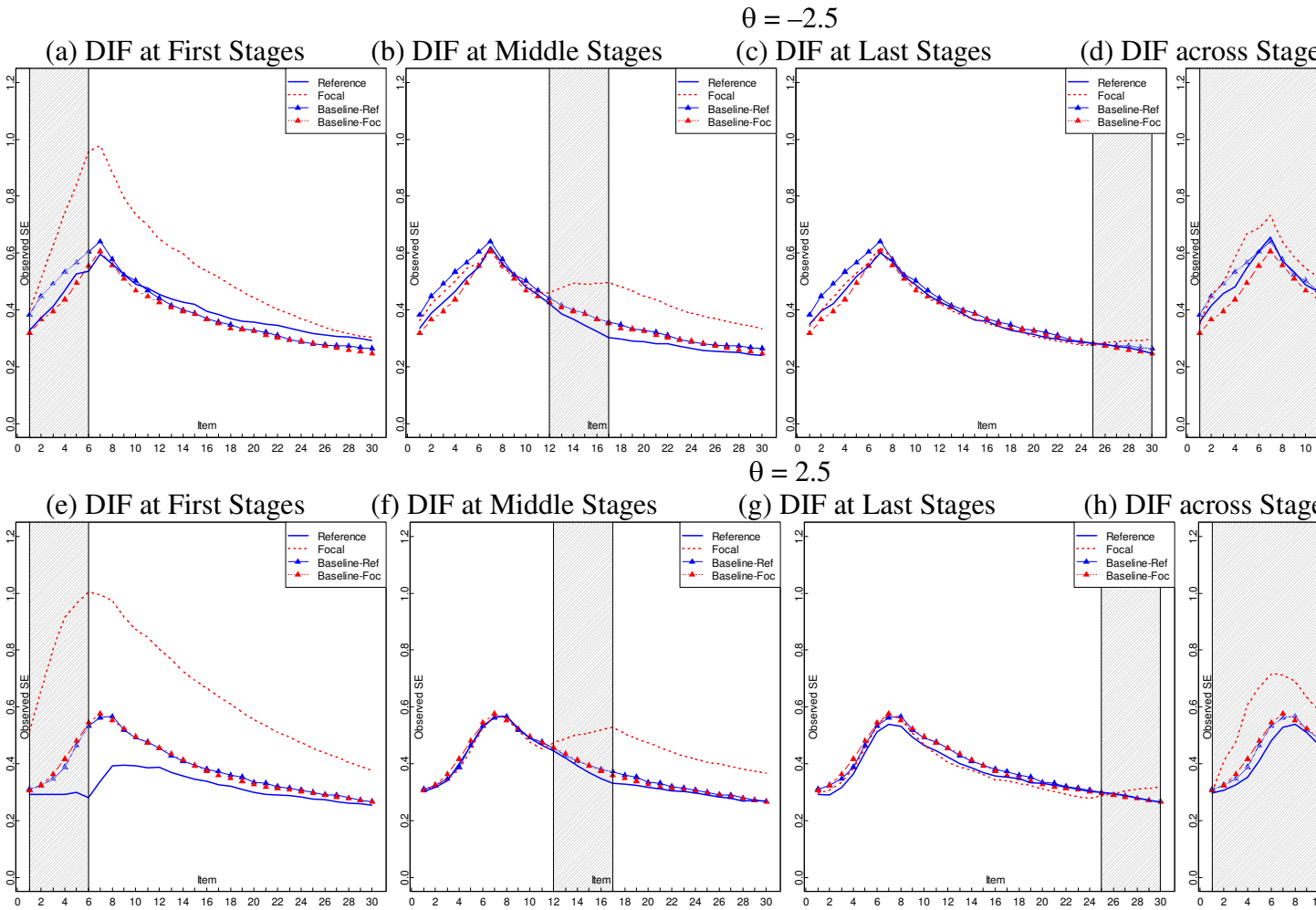
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 43. Observed SE for θ = ±2.5 when 24 items showed nonuniform and uniform DIF with magnitudes of .4 and .4.



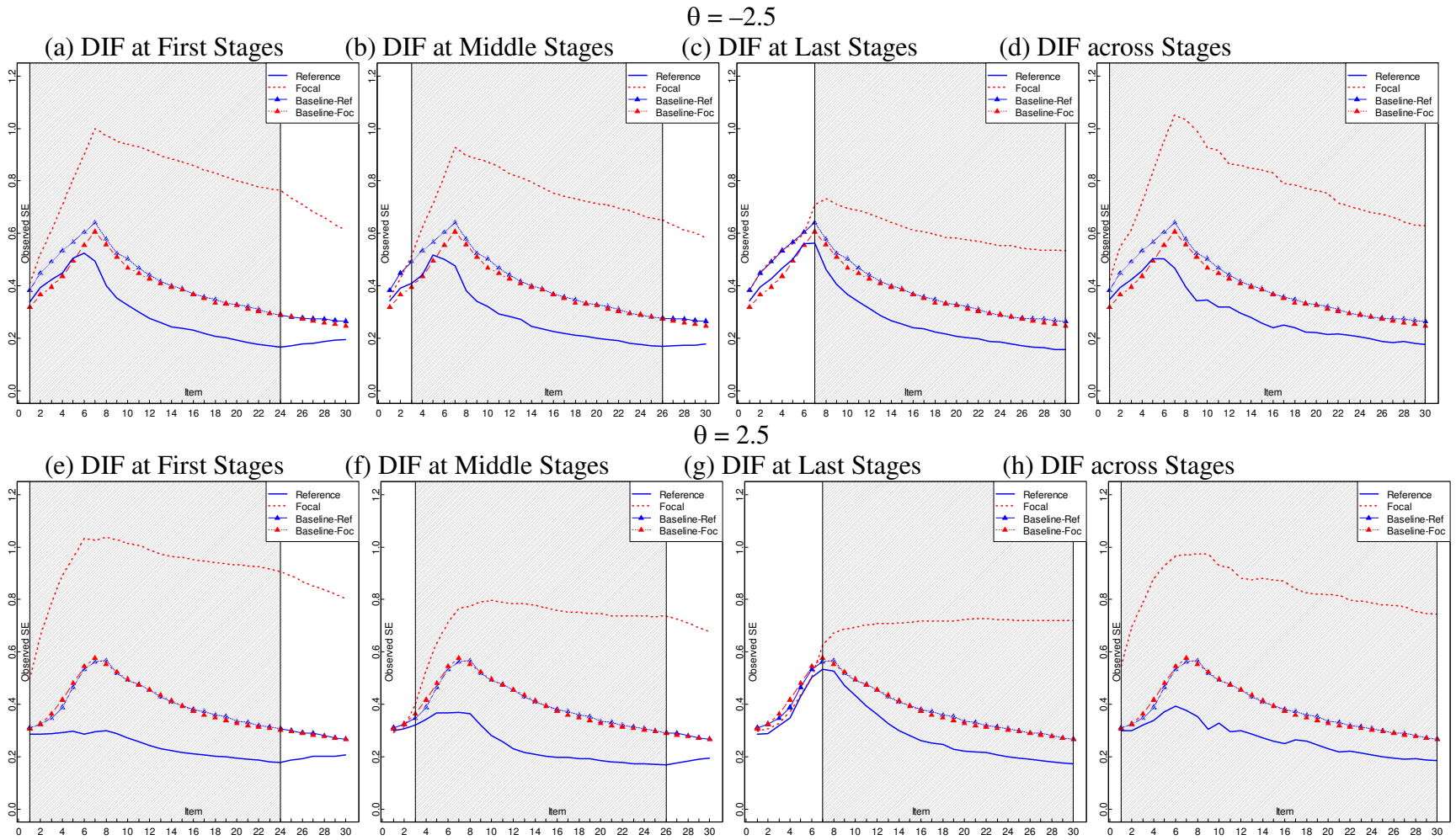
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 44. Observed SE for θ = ±2.5 when 6 items showed nonuniform and uniform DIF with magnitudes of 1.6



Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DI

Figure 45. Observed SE for θ = ±2.5 when 24 items showed nonuniform and uniform DIF with magnitudes of 1.6 and 1.6.



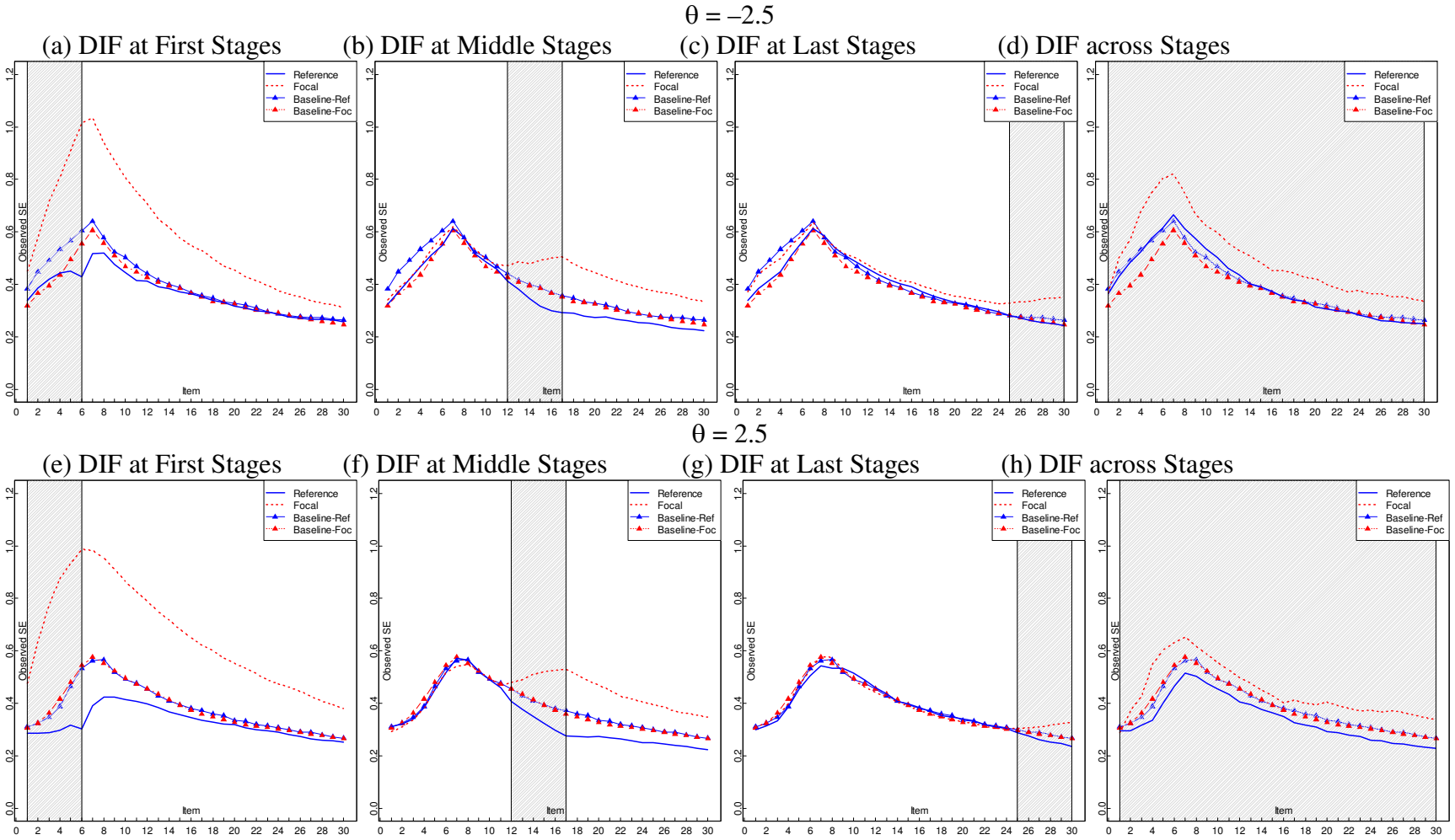
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

3.4.4 SE when DIF items exhibited both types of DIF with different magnitudes

Similar to BIAS and RMSE, the pattern of SE obtained from CAT with DIF items showing both types of DIF in different magnitudes was defined by the larger-magnitude DIF type. For example, Figures 46–47 present the patterns of SE obtained from items which displayed nonuniform DIF in a larger magnitude than uniform DIF. Thus, the SE from such conditions was influenced mostly by the effect of nonuniform DIF, similar to the patterns of SE in Section 3.4.2. Specifically, the SE of $\hat{\theta}_{CAT}$ for the focal group was larger than that for the reference group after DIF items were administered, despite the change in DIF occurrence, number of DIF items, and θ levels. In addition, the difference in SE between groups increased as the number of DIF items increased (Figure 46 vs. Figure 47).

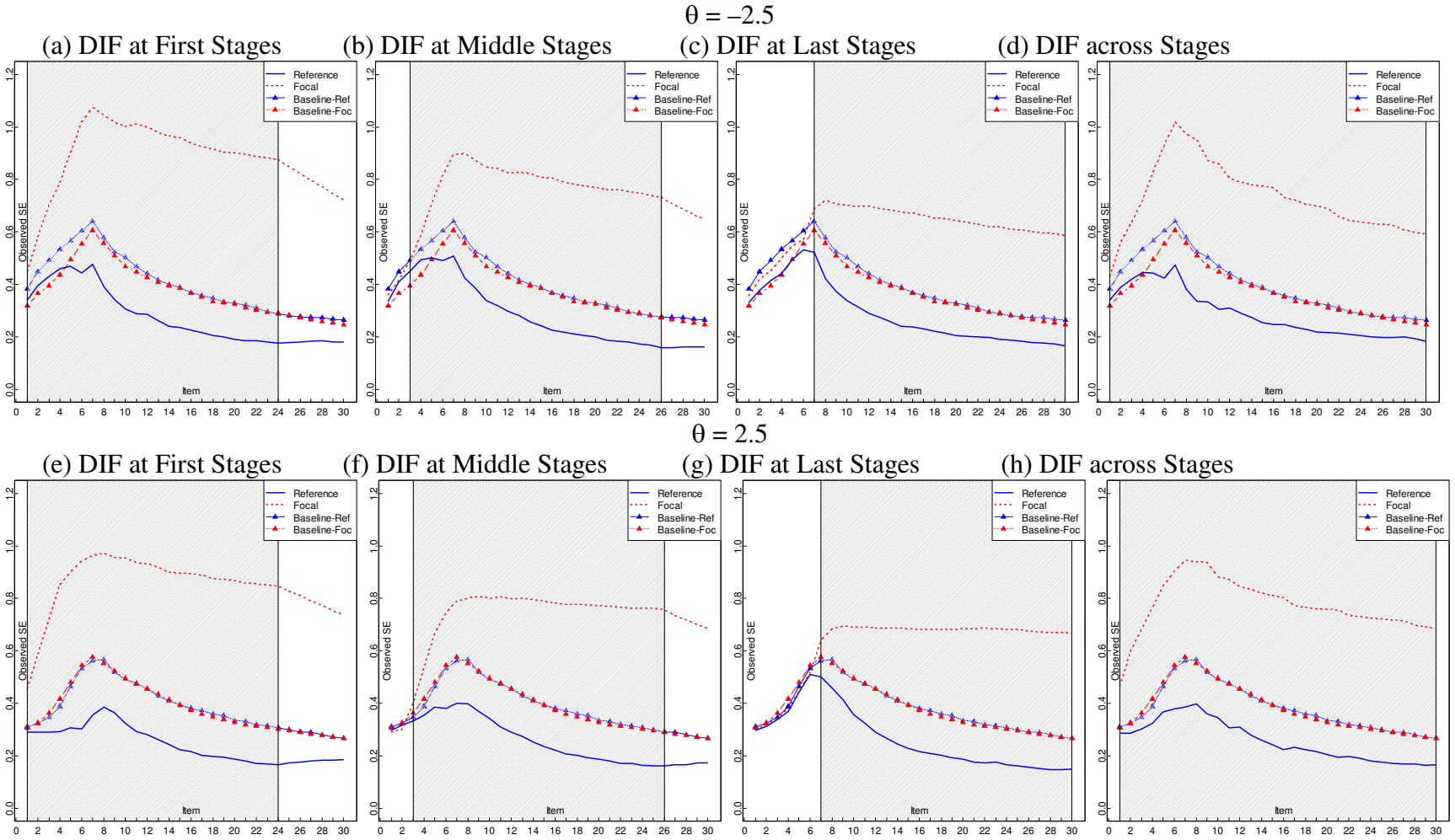
The patterns of SE in Figures 48–49 were influenced by the uniform DIF in operational items as the magnitude of uniform DIF was larger than the magnitude of nonuniform DIF. Thus, these SE patterns resembled those in Section 3.4.1. As seen in Figure 48, when only 6 items showed very large uniform DIF (magnitude = 1.6) and moderate nonuniform DIF (magnitude = .4) simultaneously, the SE for focal group examinees was generally similar to the SE for reference group examinees despite the stage of CAT that DIF items occurred. The difference in SE between reference and focal groups slightly increased (no more than .1) as the number of DIF items increased from 6 to 24 items (Figure 48 vs. Figure 49).

Figure 46. Observed SE for θ = ±2.5 when 6 items showed nonuniform and uniform DIF with magnitudes of 1.6 and .4.



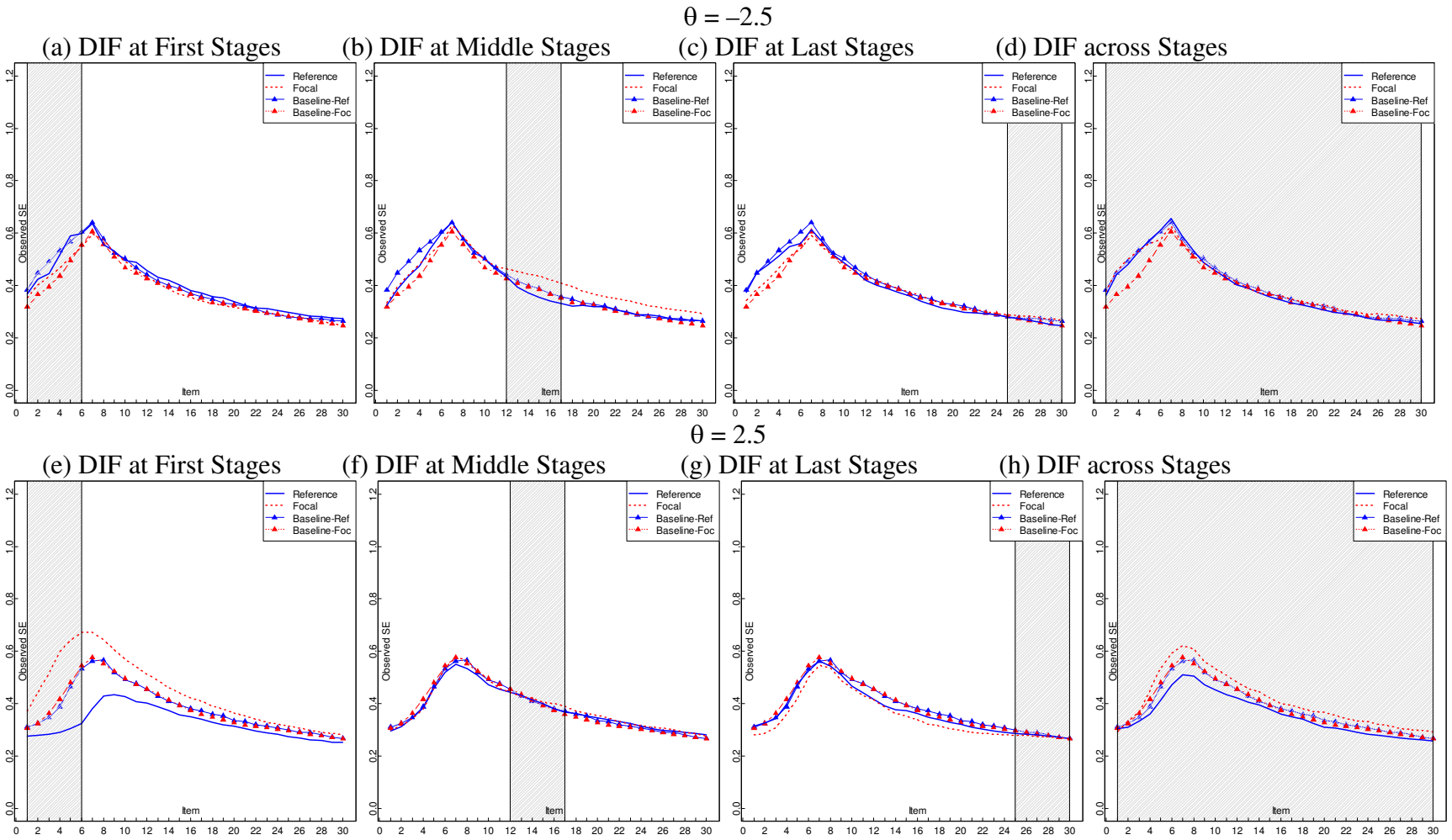
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 47. Observed SE for θ = ±2.5 when 24 items showed nonuniform and uniform DIF with magnitudes of 1.6 and .4.



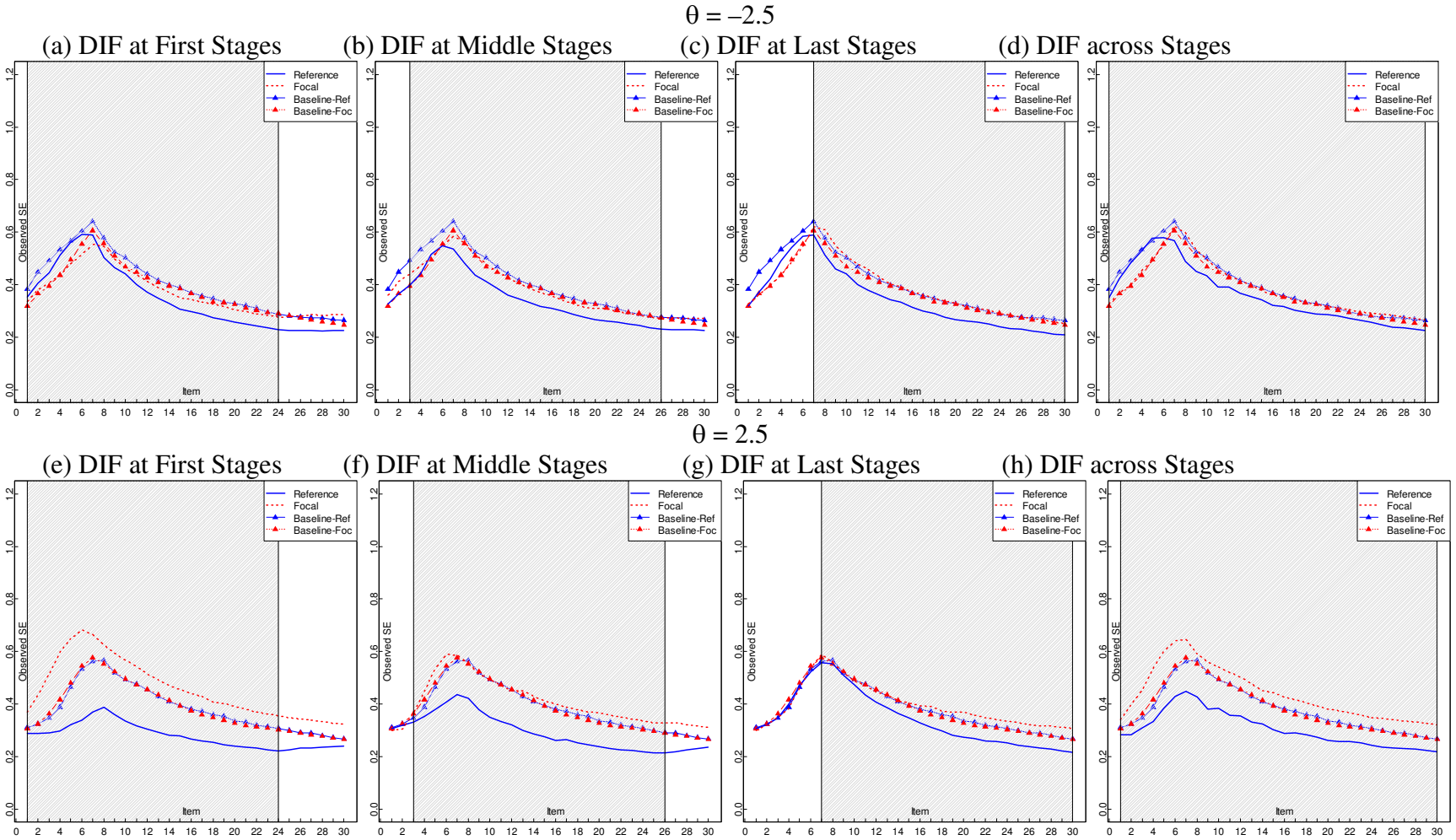
Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 48. Observed SE for θ = ±2.5 when 6 items showed nonuniform and uniform DIF with magnitudes of .4 and 1.6.



Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Figure 49. Observed SE for θ = ±2.5 when 24 items showed nonuniform and uniform DIF with magnitudes of .4 and 1.6.



Note: Baseline = DIF-free CAT condition or no DIF in operational items; The shaded area indicates operational items that showed DIF.

Chapter 4: Results of Study 2

Study 2 was designed to examine the accuracy of DIF detection in pretest items using three types of matching variables (number-correct score (NCS), ability estimate obtained from CBT ($\hat{\theta}_{CBT}$), and $\hat{\theta}_{CAT}$) under the CAT settings as simulated in Study 1. The question of interest was: to detect DIF in nonadaptive pretest items, does matching examinees on $\hat{\theta}_{CAT}$ provide more accurate results than NCS and $\hat{\theta}_{CBT}$? In this chapter, the results from some selected simulation conditions are presented, including when:

(1) operational items were DIF free

(2) 6 operational items of the 30-item test exhibited uniform DIF with a magnitude of .4 at the first stages of CAT

(3) 6 operational items of the 30-item test exhibited nonuniform DIF with a magnitude of .4 at the first stages of CAT

(4) 6 operational items of the 30-item test exhibited both types of DIF with magnitudes of .4 at the first stages of CAT

(5) 24 operational items of the 30-item test exhibited uniform DIF with a magnitude of 1.6 at the last stages of CAT

(6) 24 operational items of the 30-item test exhibited nonuniform DIF with a magnitude of 1.6 at the last stages of CAT

(7) 24 operational items of the 30-item test exhibited both types of DIF with magnitudes of 1.6 at the last stages of CAT

Based on the results of Study 1, the first four conditions here provided a small bias in $\hat{\theta}_{CAT}$. The last three conditions, in contrast, were likely to provide a large amount of bias. When the $\hat{\theta}_{CAT}$ obtained from these settings served as the matching variable for DIF detection in pretest items, it was expected that the results of such DIF detections would be less accurate. Similar results were also expected for the case of NCS and $\hat{\theta}_{CBT}$.

For each of the selected conditions above, Type I error (false-positive rates) and power (true-positive rates) of detecting uniform and nonuniform DIF in pretest items (using logistic regression and the Mantel-Haenszel (MH) statistics with NCS, $\hat{\theta}_{CBT}$, and $\hat{\theta}_{CAT}$ as the matching variable) are graphically presented in Figures 50-70. Appendix G

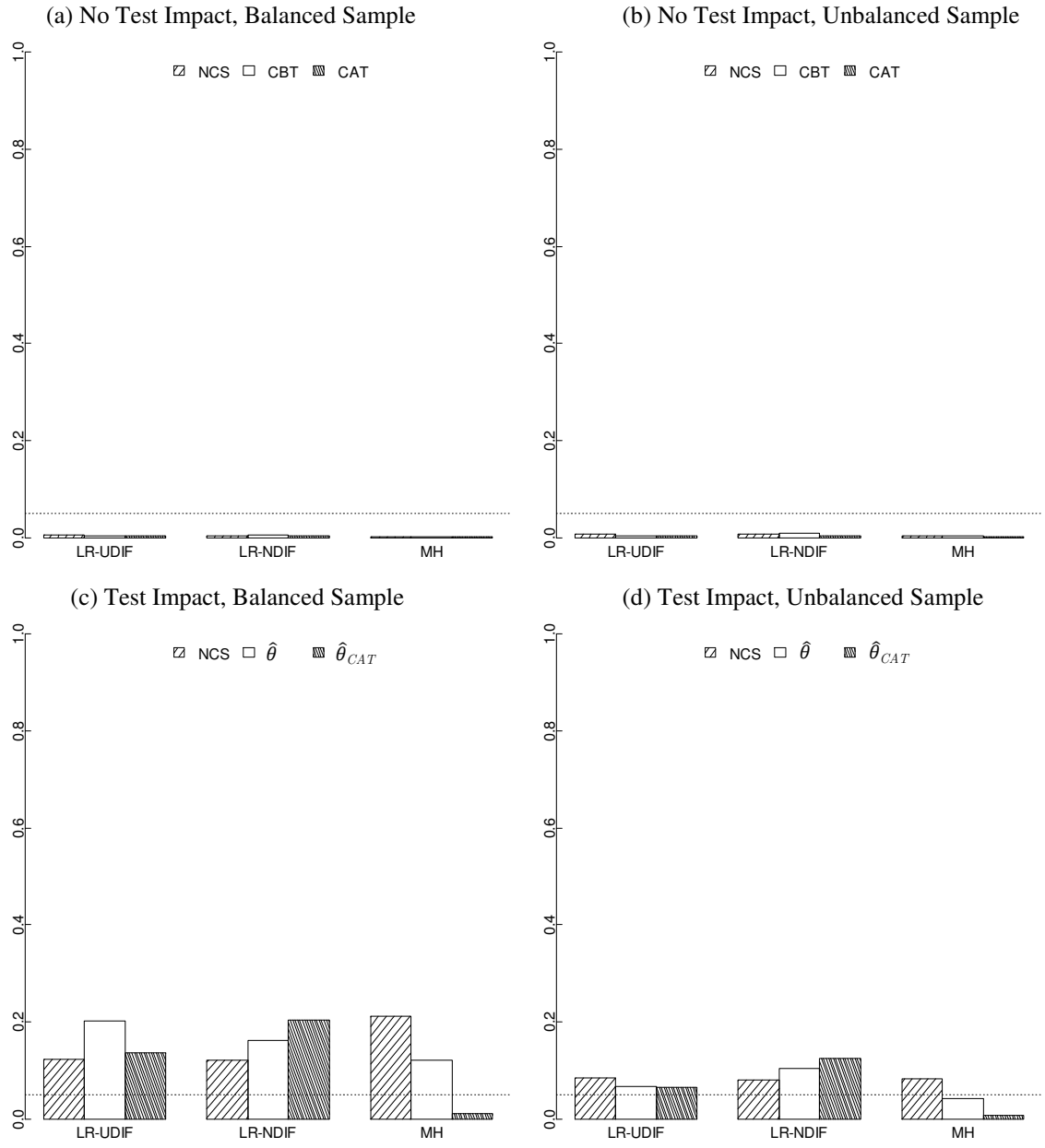
provides the detailed results of DIF detection. In each figure, there are four panels representing the results under the four combinations of test impact (no test impact vs. test impact) and sample size ratio (balanced vs. unbalanced). In addition, for logistic regression, the results from the significance test of the group main effect or uniform DIF effect (LR-UDIF) are presented along with the results from the significance test of group-matching interaction effect or nonuniform DIF effect (LR-NDIF).

4.1 Type I Error

When operational items had no DIF (Figure 50) or only the first six items of the operational test exhibited DIF (Figures 51-53), the patterns of Type I error were similar. Specifically, if test impact was absent, the average Type I error rates were much lower than .05 regardless of the sample size ratio, DIF type in operational items, DIF detection method, and type of matching (see panels a and b in the figures). However, when test impact existed, the error rates increased and exceeded the .05 cutoff for most conditions of the sample size ratio, DIF detection method, and matching variable. Only MH with $\hat{\theta}_{CAT}$ as the matching variable provided Type I error lower than .05 under the test impact condition (see the last column of panels c and d in the figures). It should be noted here that the observed Type I error rates that are too low are also not good.

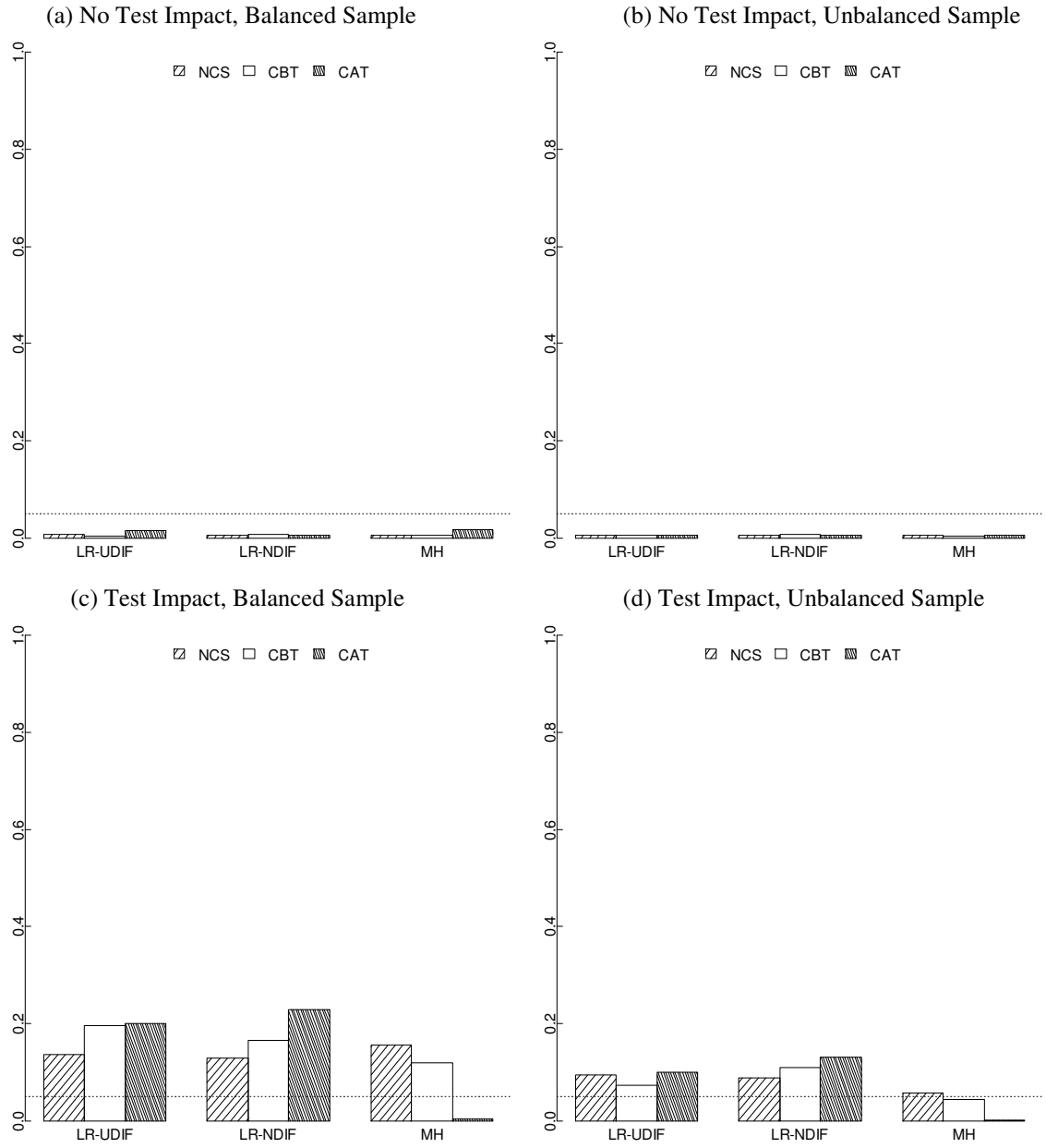
When DIF with the magnitude of 1.6 occurred at the end of the operational test, the average false-positive rates seemed to be dependent on the type of DIF in operational items, rather than the other simulation factors. When 24 items showed uniform DIF with the magnitude of 1.6 (Figures 54 and 56), Type I error rates for both LR and MH with $\hat{\theta}_{CAT}$ as the matching variable substantially increased (almost 1 in most conditions). In contrast, when operational items exhibited only nonuniform DIF at the end of the test, even with a large magnitude, Type I error rates were stable in most of the cases. For example, as seen in Figure 55, MH with $\hat{\theta}_{CAT}$ provided Type I error rates below .05, regardless of the test impact and sample size ratio. On the other hand, LR-UDIF and LR-NDIF could control Type I error only when there was no test impact. Finally, when most of the operational items exhibited uniform DIF with the magnitude of 1.6, MH with CAT yielded the largest Type I error rate (among the MH results) when test impact was absent and group sample sizes were equal (Figure 54a).

Figure 50. Type I Error of detecting DIF in DIF-free pretest items when the operational test was DIF-free.



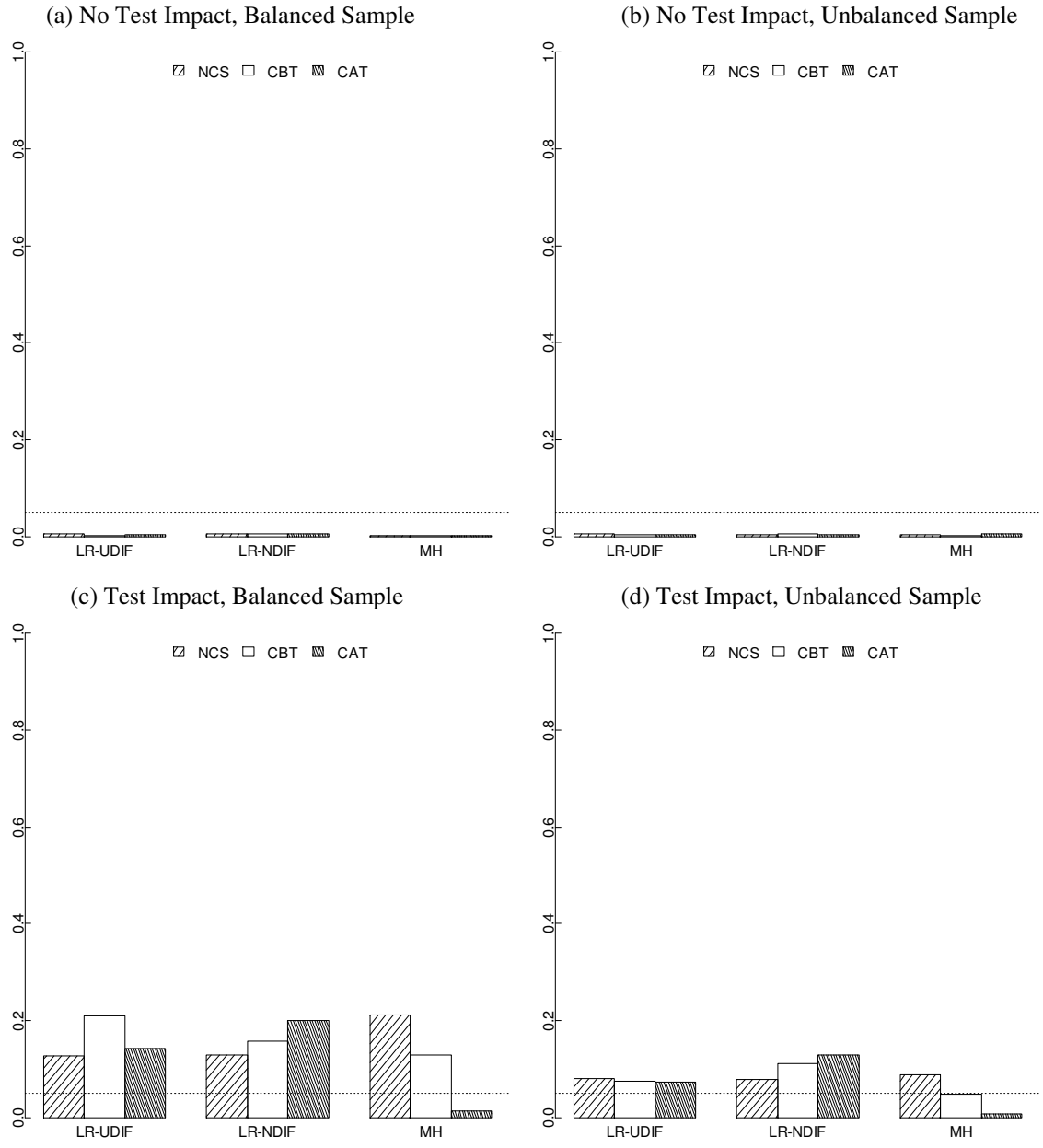
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

Figure 51. Type I Error of detecting DIF in DIF-free pretest items when the operational test consisted of 6 items exhibiting uniform DIF with a magnitude of .4 at the beginning of the test.



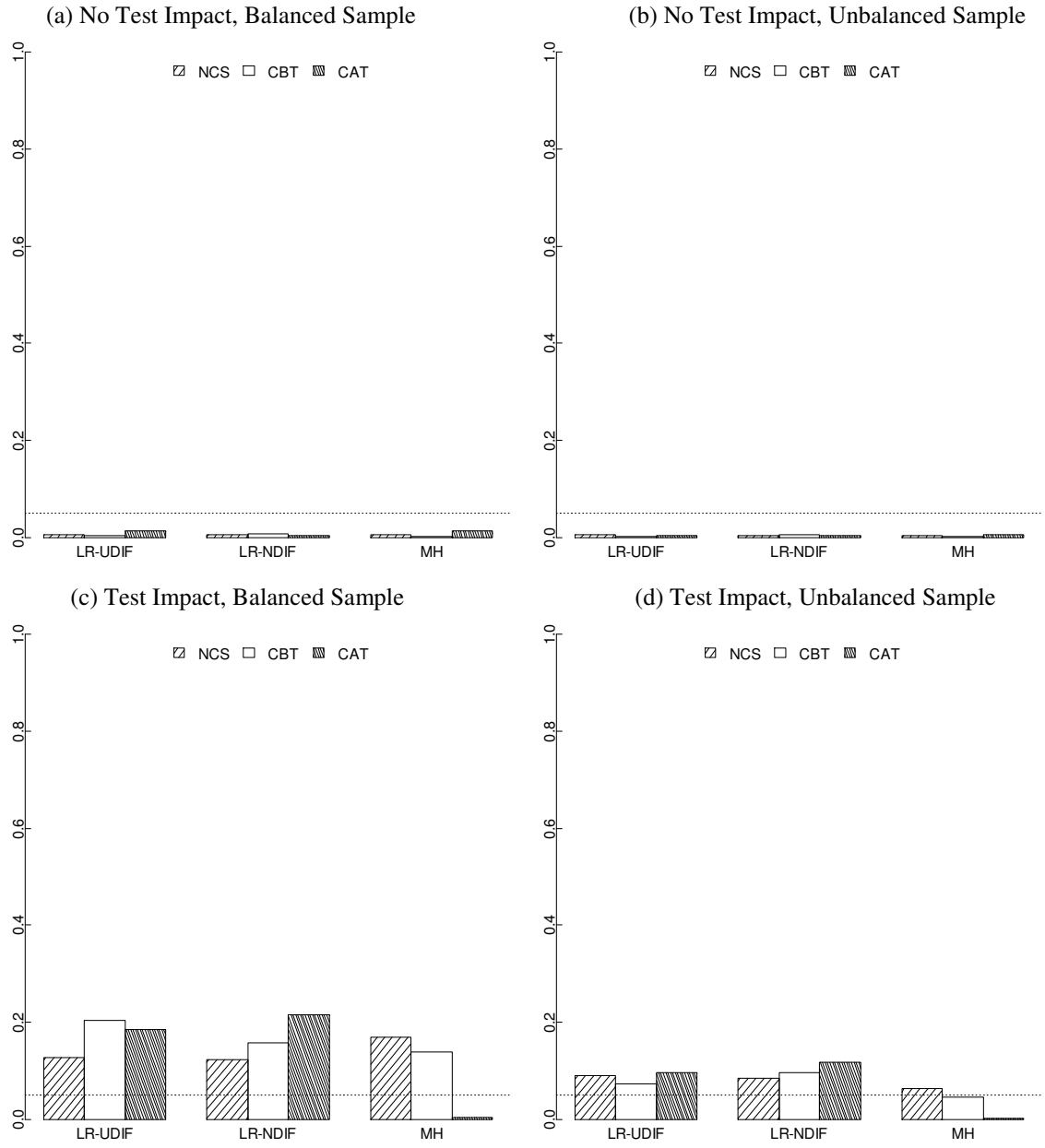
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

Figure 52. Type I Error of detecting DIF in DIF-free pretest items when the operational test consisted of 6 items exhibiting nonuniform DIF with a magnitude of .4 at the beginning of the test.



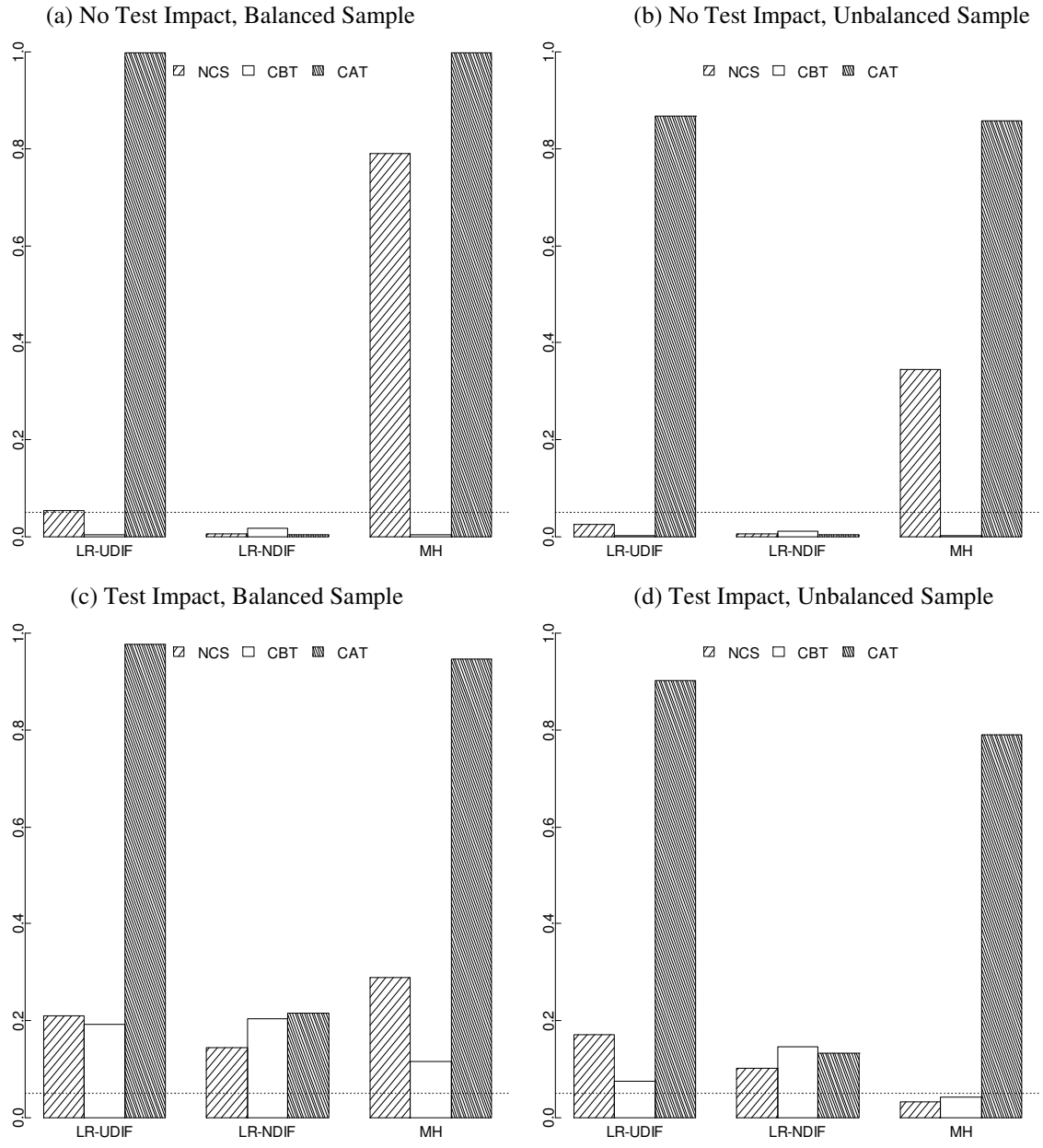
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

Figure 53. Type I Error of detecting DIF in DIF-free pretest items when the operational test consisted of 6 items exhibiting both nonuniform and uniform DIF with magnitudes of .4 at the beginning of the test.



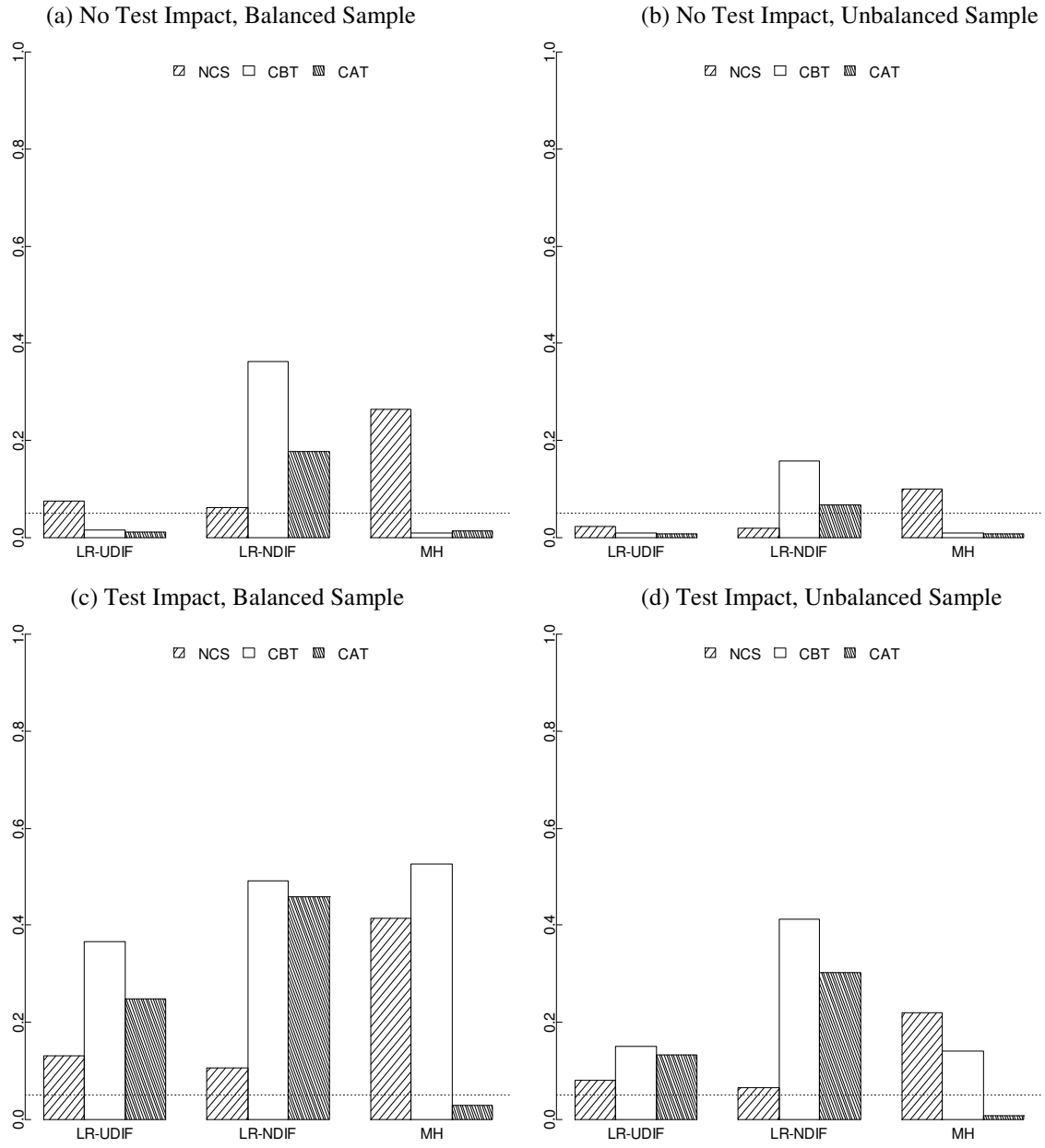
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

Figure 54. Type I Error of detecting DIF in DIF-free pretest items when the operational test consisted of 24 items exhibiting uniform DIF with a magnitude of 1.6 at the end of the test.



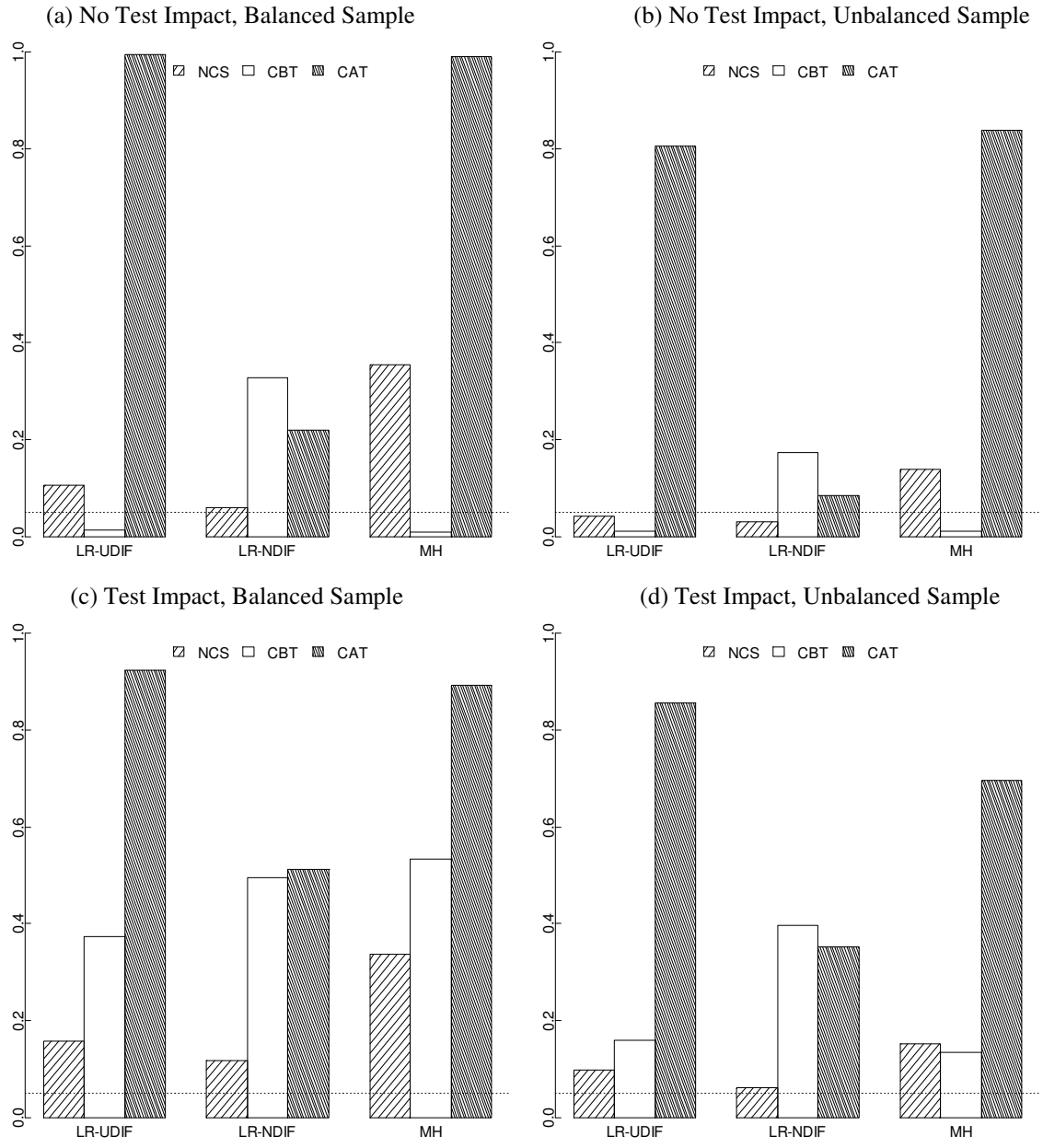
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

Figure 55. Type I Error of detecting DIF in DIF-free pretest items when the operational test consisted of 24 items exhibiting nonuniform DIF with a magnitude of 1.6 at the end of the test.



Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

Figure 56. Type I Error of detecting DIF in DIF-free pretest items when the operational test consisted of 24 items exhibiting both nonuniform and uniform DIF with magnitudes of 1.6 at the end of the test.



Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .05 or 5% level.

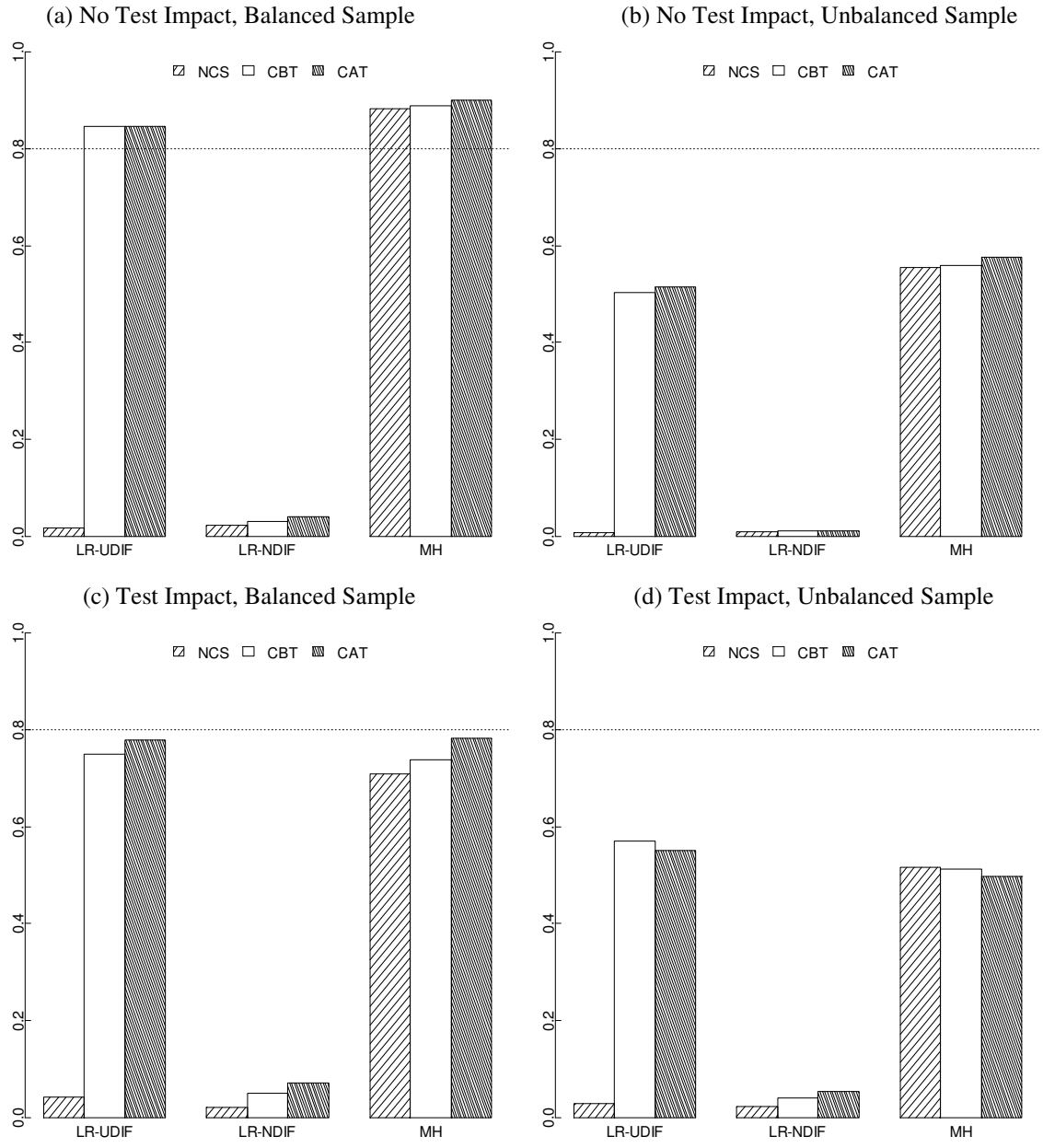
4.2 Power in detecting uniform DIF in pretest items

It was evident that using the ability estimate (either NCS, $\hat{\theta}_{CBT}$, or $\hat{\theta}_{CAT}$) obtained from a few items that exhibited small uniform DIF at the beginning of the operational test did not substantially reduce the power of the LR and MH in detecting uniform DIF in pretest items. As seen in Figures 57-60, the patterns of power were almost identical. The power of both LR and MH were largest when no test effect occurred and the reference and focal groups were balanced. The power decreased if test impact existed and/or group sample sizes were unbalanced. In such cases, the power of both LR and MH were lower than .8, regardless of the type of matching variable.

Specifically, MH with any choice of matching comparably provided the highest power (at least .8), given that there was no test impact and the sample sizes were identical (panel a). On the other hand, LR-UDIF with $\hat{\theta}_{CAT}$ had comparable power with, if not larger than, MH with $\hat{\theta}_{CAT}$ in most of the test impact and sample size ratio conditions. In contrast, LR-UDIF with NCS consistently yielded the lowest power across conditions of test impact and sample size ratio. More specifically, using $\hat{\theta}_{CAT}$ as the matching variable seemed to provide larger power (about 10%) than NCS and $\hat{\theta}_{CBT}$ when test impact existed in the equal sample cases (panel c). Regarding the power of LR-NDIF (i.e., LR could detect DIF but mistakenly flagged as nonuniform DIF), matching examinees on $\hat{\theta}_{CAT}$ generally provided the highest power for such detection, followed by $\hat{\theta}_{CBT}$ and NCS.

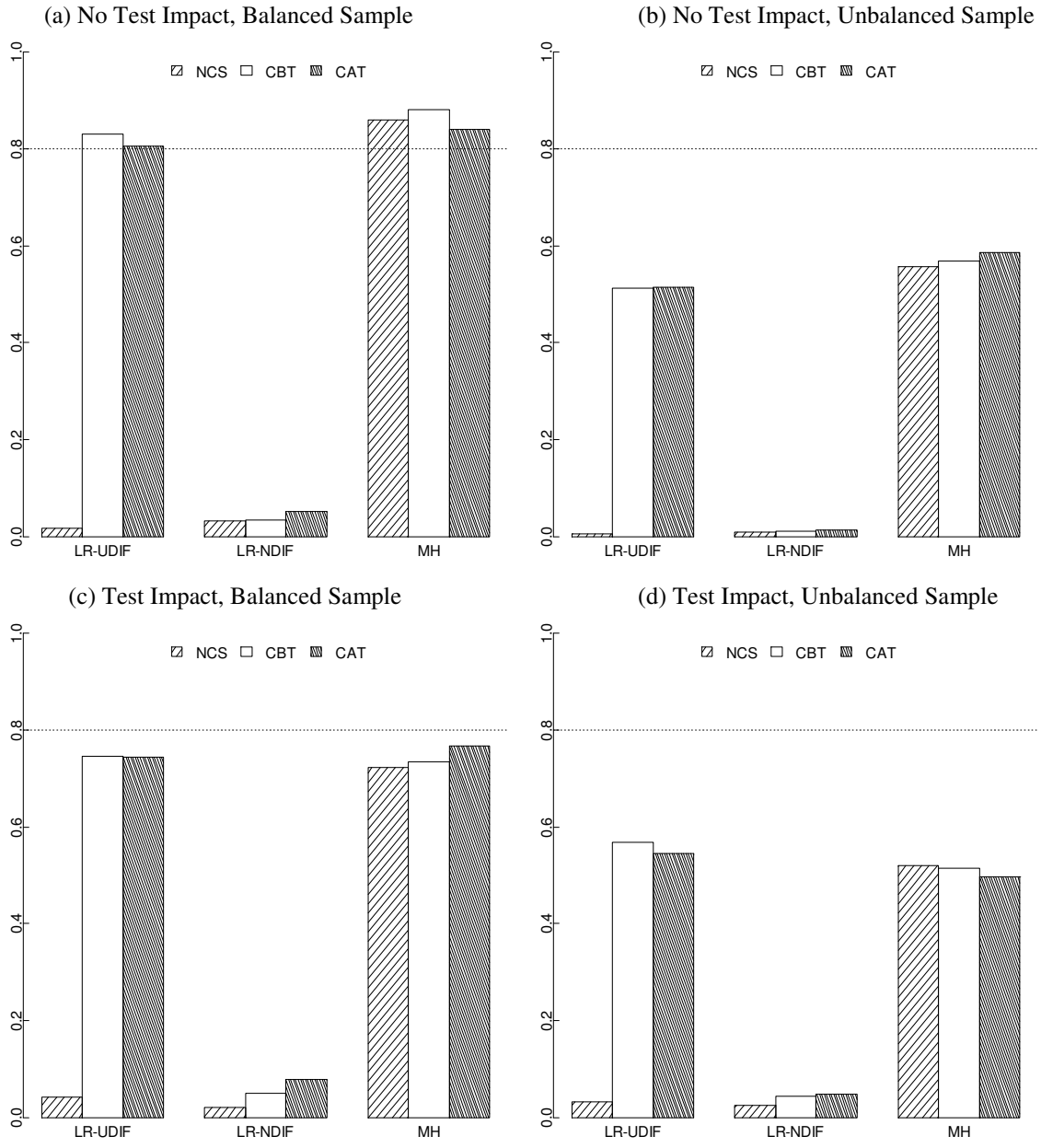
Figures 61-63 show that when the large-magnitude DIF was embedded into 24 items at the end of operational tests, the power rates of detecting uniform DIF in pretest items for all combinations of the detection method and matching variable essentially increased. Especially when 24 items with very large uniform DIF (magnitude = 1.6) were administered, the power of LR-UDIF using $\hat{\theta}_{CAT}$ as the matching variable increased from about .85 to almost 1 (Figure 61a). Also, many items in the operational test were contaminated by uniform DIF, the power of LR-UDIF and MH using $\hat{\theta}_{CAT}$ as the matching variable were higher than using NCS and $\hat{\theta}_{CBT}$ in most conditions. Only the case of LR-NDIF when 24 items exhibited nonuniform DIF (middle columns in Figure 62) that $\hat{\theta}_{CAT}$ provided lower power. However, it should be emphasized that such a power rate came from mistakenly detecting uniform DIF items as nonuniform DIF.

Figure 57. Power in detecting uniform DIF in pretest items when the operational test was DIF-free.



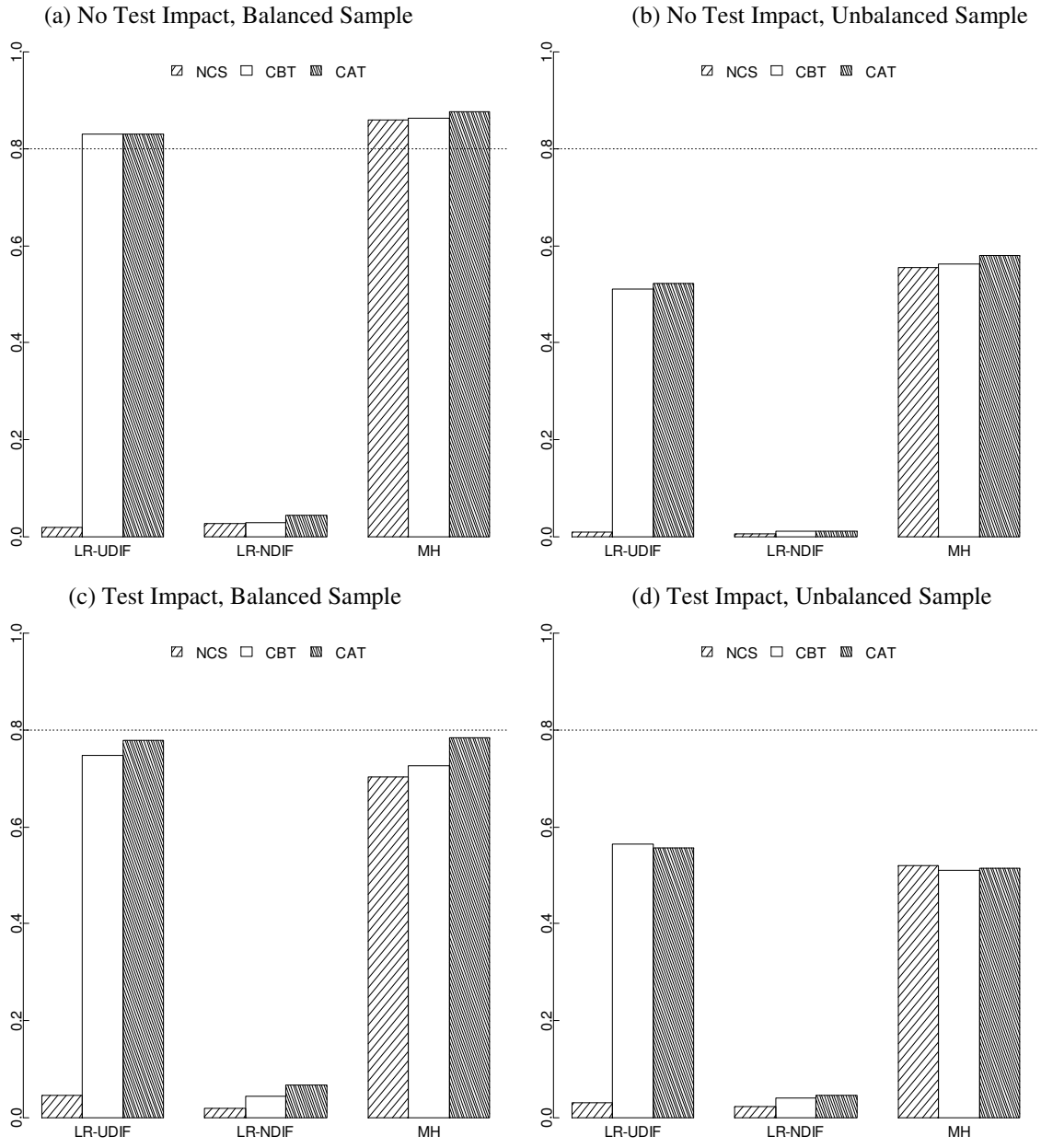
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 58. Power in detecting uniform DIF in pretest items when the operational test consisted of 6 items exhibiting uniform DIF with a magnitude of .4 at the beginning of the test.



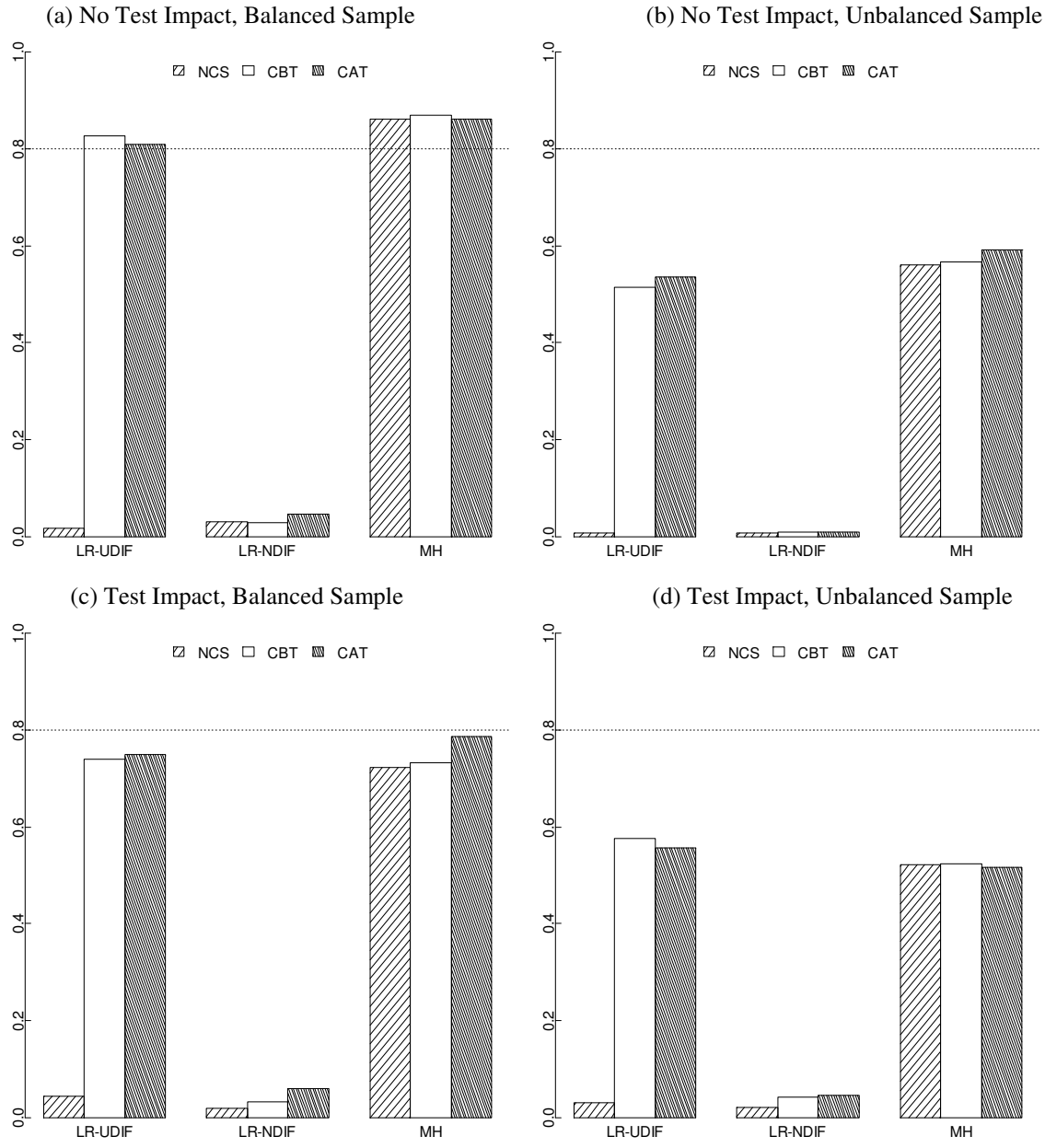
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 59. Power in detecting uniform DIF in pretest items when the operational test consisted of 6 items exhibiting nonuniform DIF with a magnitude of .4 at the beginning of the test.



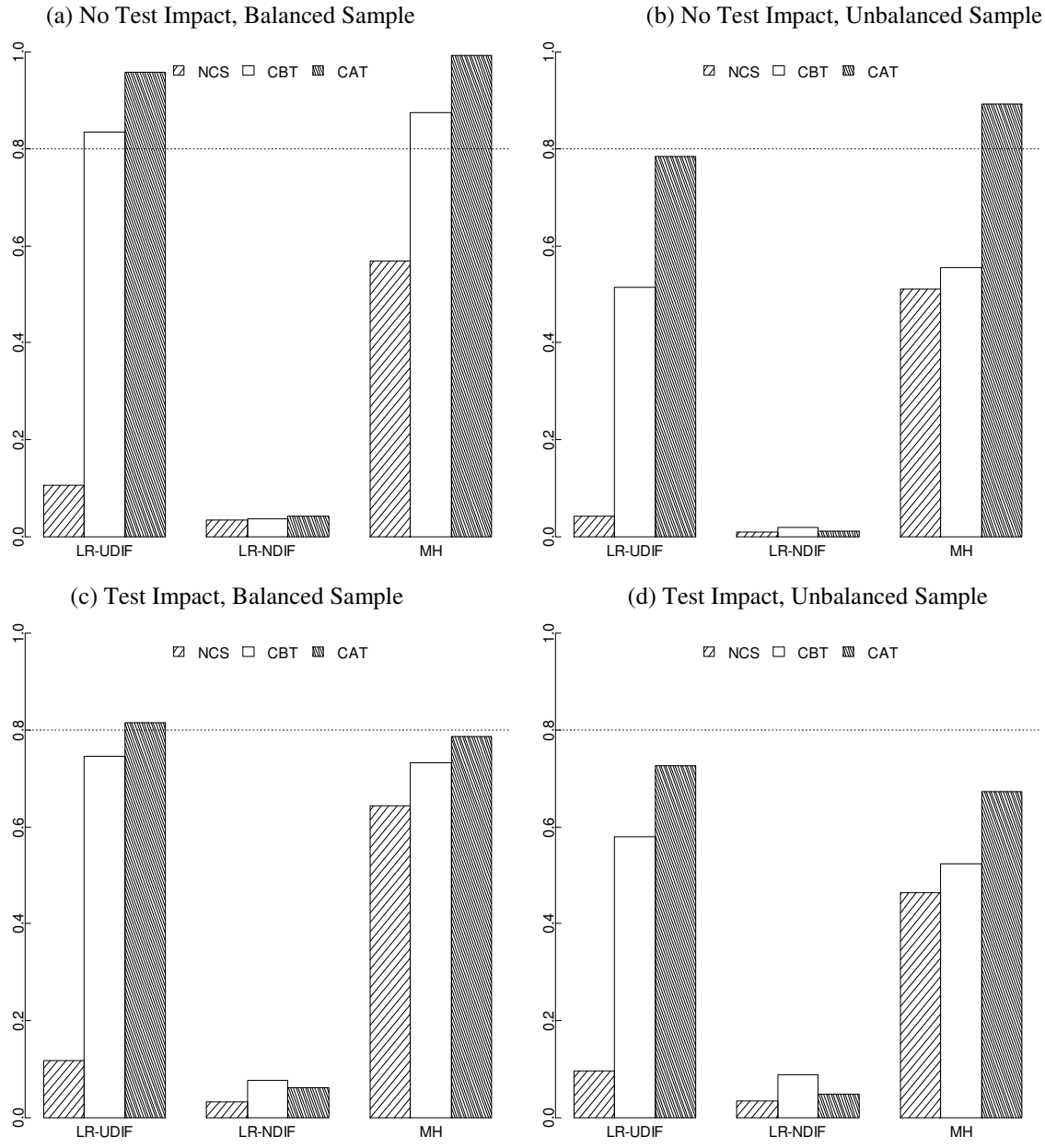
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 60. Power in detecting uniform DIF in pretest items when the operational test consisted of 6 items exhibiting both nonuniform and uniform DIF with magnitudes of .4 at the beginning of the test.



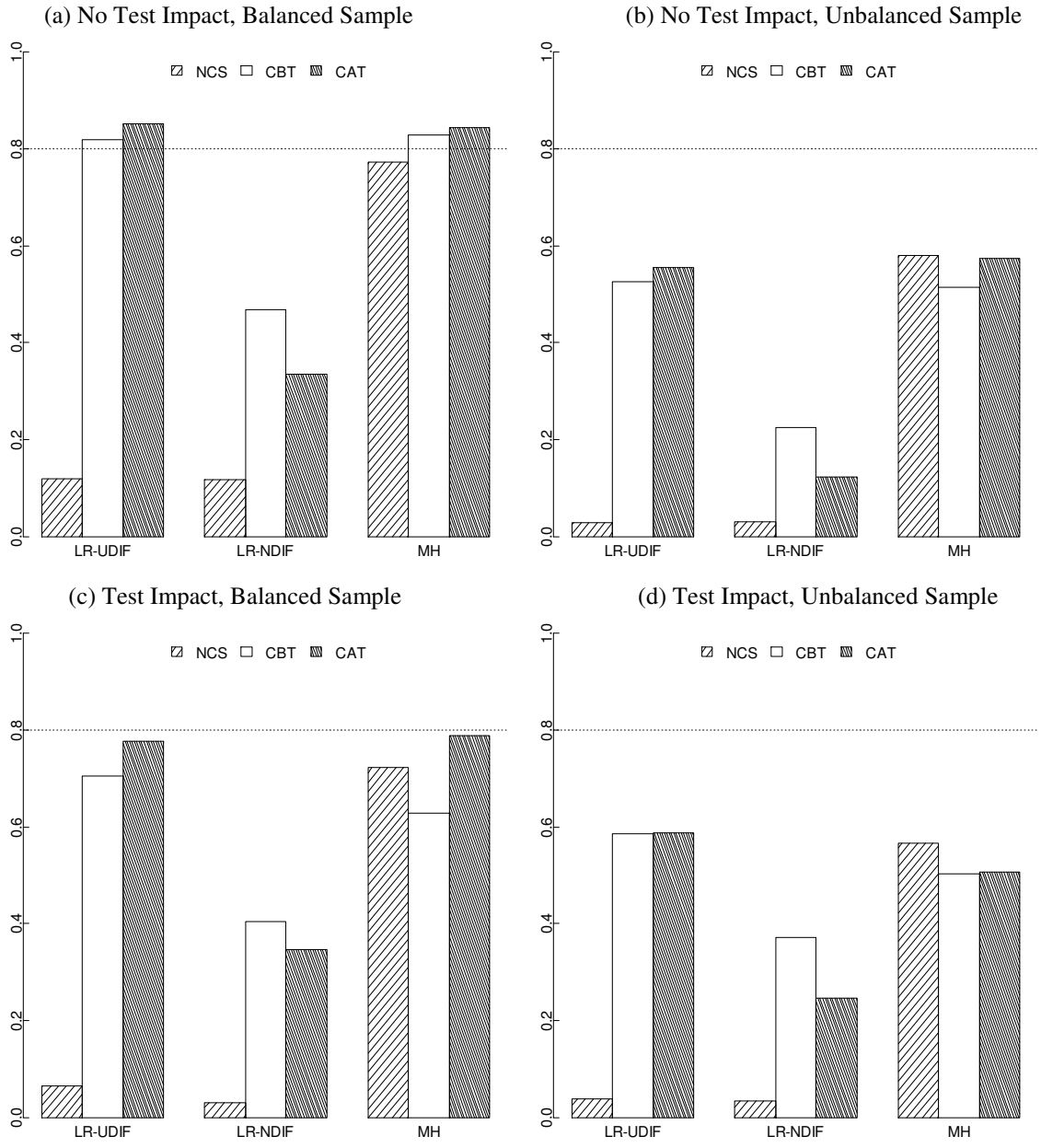
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 61. Power in detecting uniform DIF in pretest items when the operational test consisted of 24 items exhibiting uniform DIF with a magnitude of 1.6 at the end of the test.



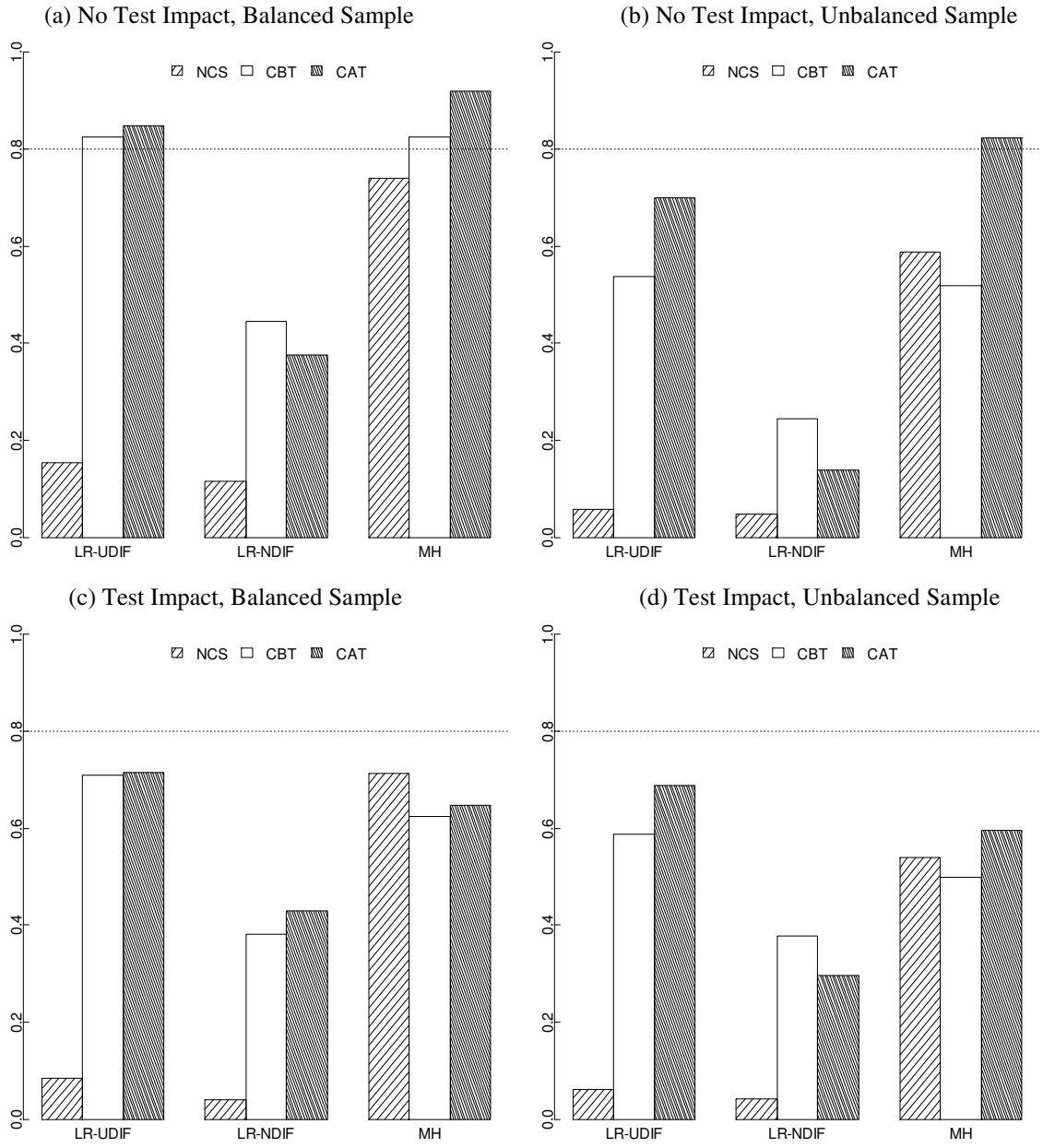
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 62. Power in detecting uniform DIF in pretest items when the operational test consisted of 24 items exhibiting nonuniform DIF with a magnitude of 1.6 at the end of the test.



Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 63. Power in detecting uniform DIF in pretest items when the operational test consisted of 24 items exhibiting both nonuniform and uniform DIF with magnitudes of 1.6 at the end of the test.



Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

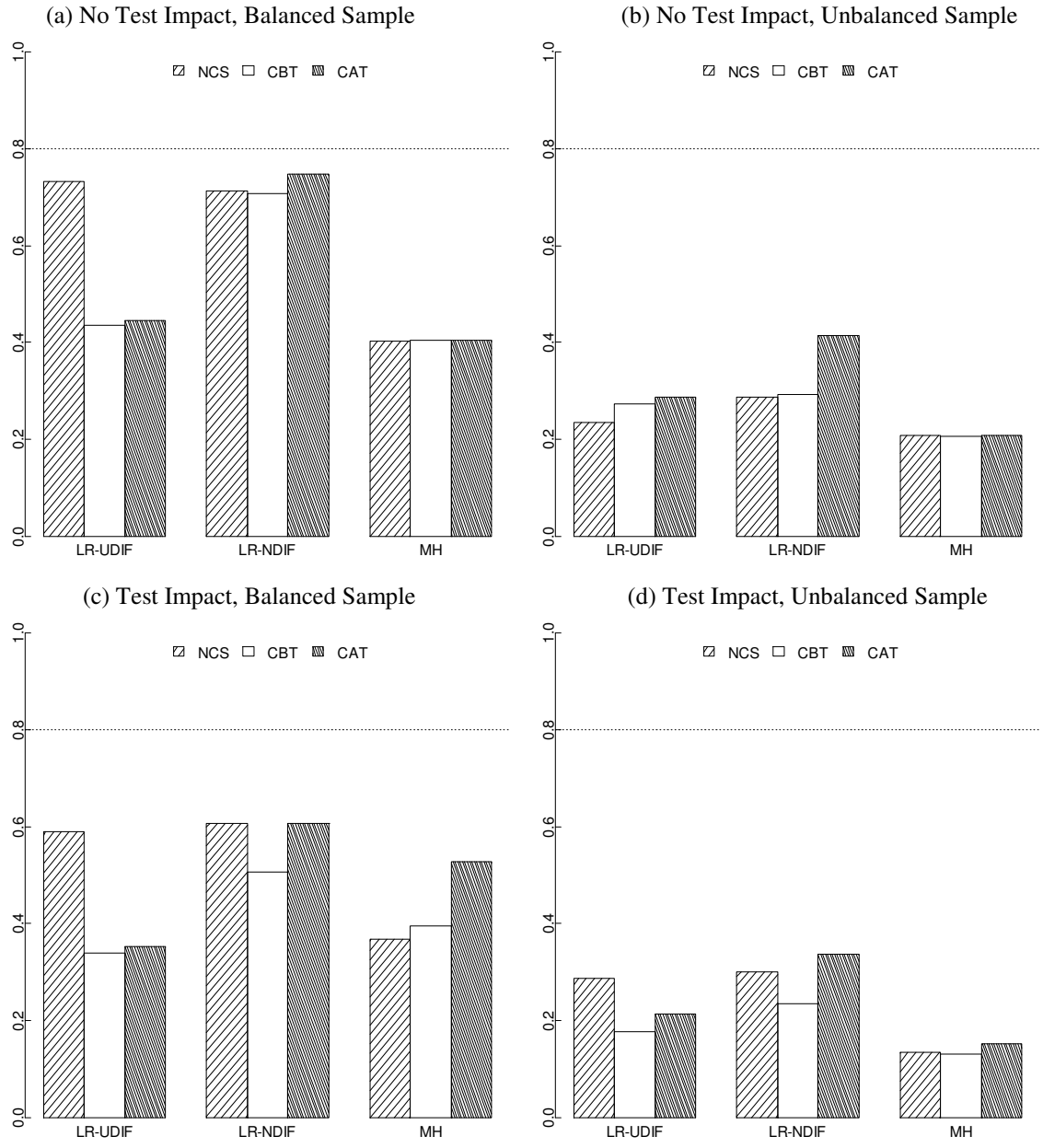
4.3 Power in detecting nonuniform DIF in pretest items

The patterns of power in detecting nonuniform DIF in pretest items were similar to the results of power in detecting uniform DIF in pretest items. That is, when 6 operational items with a small DIF magnitude were administered at the beginning of the test (Figures 65-67), results were almost identical to those observed when the operational test was DIF free (Figure 64). In fact, neither LR nor MH provided power larger than .8, regardless of the choice of matching variable, test impact and sample size ratio.

As seen in the figures, LR-NDIF with $\hat{\theta}_{CAT}$ provided the highest power in most of the simulation conditions. The power of LR-NDIF with $\hat{\theta}_{CAT}$ ranged between .60-.75 for balanced samples and .30-.40 for unbalanced samples. MH with $\hat{\theta}_{CAT}$, however, yielded the highest power for detecting nonuniform DIF in pretest items when 6 operational items showing uniform DIF with the magnitude of .4, test impact existed, and sample sizes were balanced (Figure 65c). LR-NDIF with NCS, on the other hand, was more powerful than LR-NDIF with $\hat{\theta}_{CAT}$ when detecting nonuniform DIF in pretest items that 6 operational items showing uniform DIF with the magnitude of .4, test impact existed, and sample sizes were balanced (Figures 66c and 67c). Interestingly, using $\hat{\theta}_{CAT}$ as the matching variable increased the power in detecting nonuniform DIF for MH in most conditions (with the exception of panels 66b and 67a). Using NCS as the matching variable for LR-UDIF led to mistakenly identifying nonuniform DIF as uniform DIF.

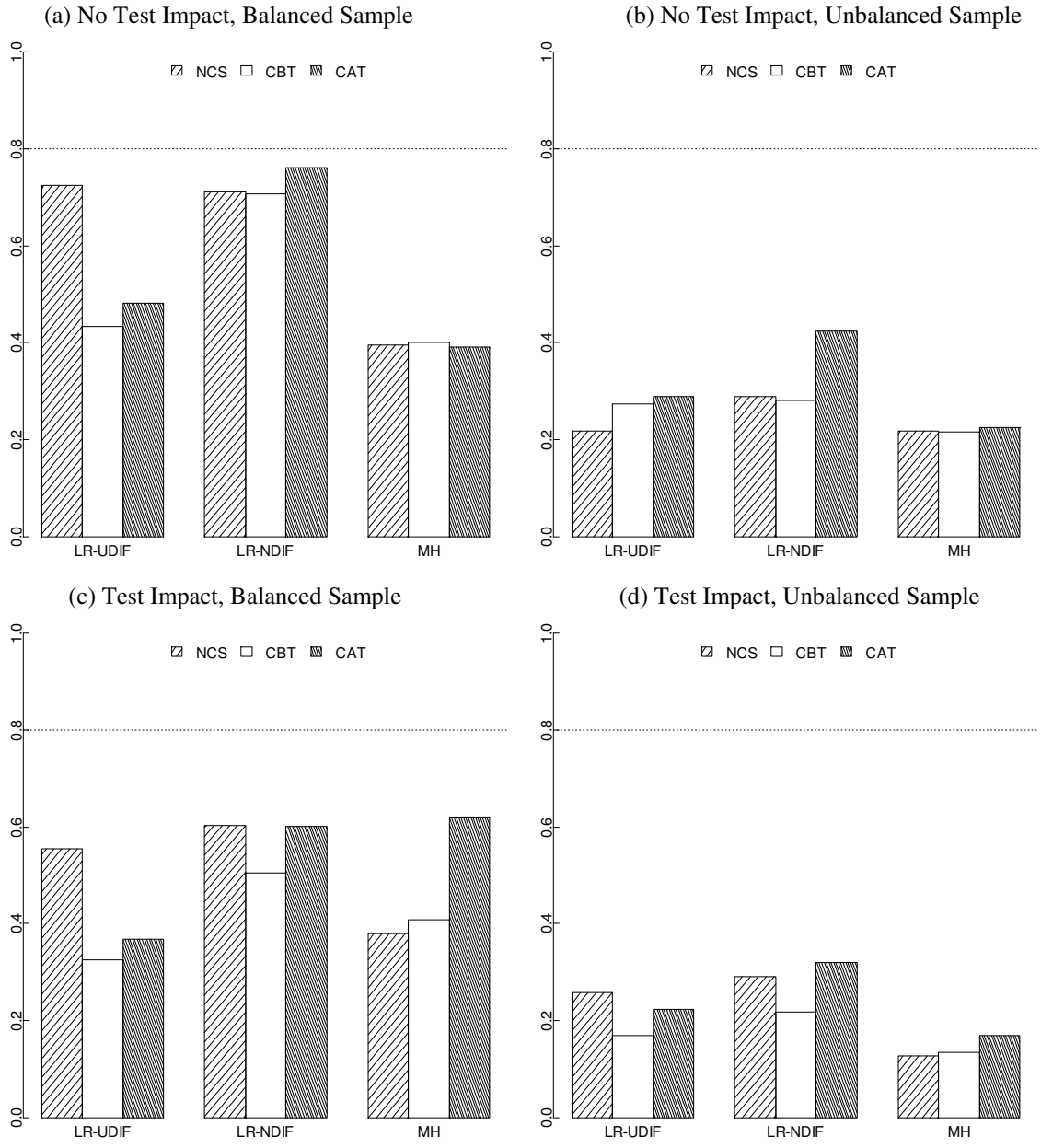
Figures 68-70 show the results of DIF detection when 24 items at the end of the operational test exhibiting DIF with the magnitude of 1.6. First, matching examinees on $\hat{\theta}_{CAT}$ increased the power of LR-UDIF from less than .50 to .85 when operational items showed uniform DIF with the magnitude of 1.6. However, such results suggest that the pretest items with nonuniform DIF were falsely identified as uniform DIF. Second, the existence of uniform DIF with the magnitude of 1.6 at the end of the operational test also increased the power of MH, especially MH with $\hat{\theta}_{CAT}$ as the matching variable. For example, the power of MH with $\hat{\theta}_{CAT}$ almost exceeded .8 and almost reached 1 when test impact existed, group sample sizes were balanced, and uniform DIF with the magnitude of 1.6 was embedded in operational items. Finally, nonuniform DIF with the magnitude of 1.6 contaminated in operational items reduced the power of all detection methods.

Figure 64. Power in detecting nonuniform DIF in pretest items when the operational test was DIF-free.



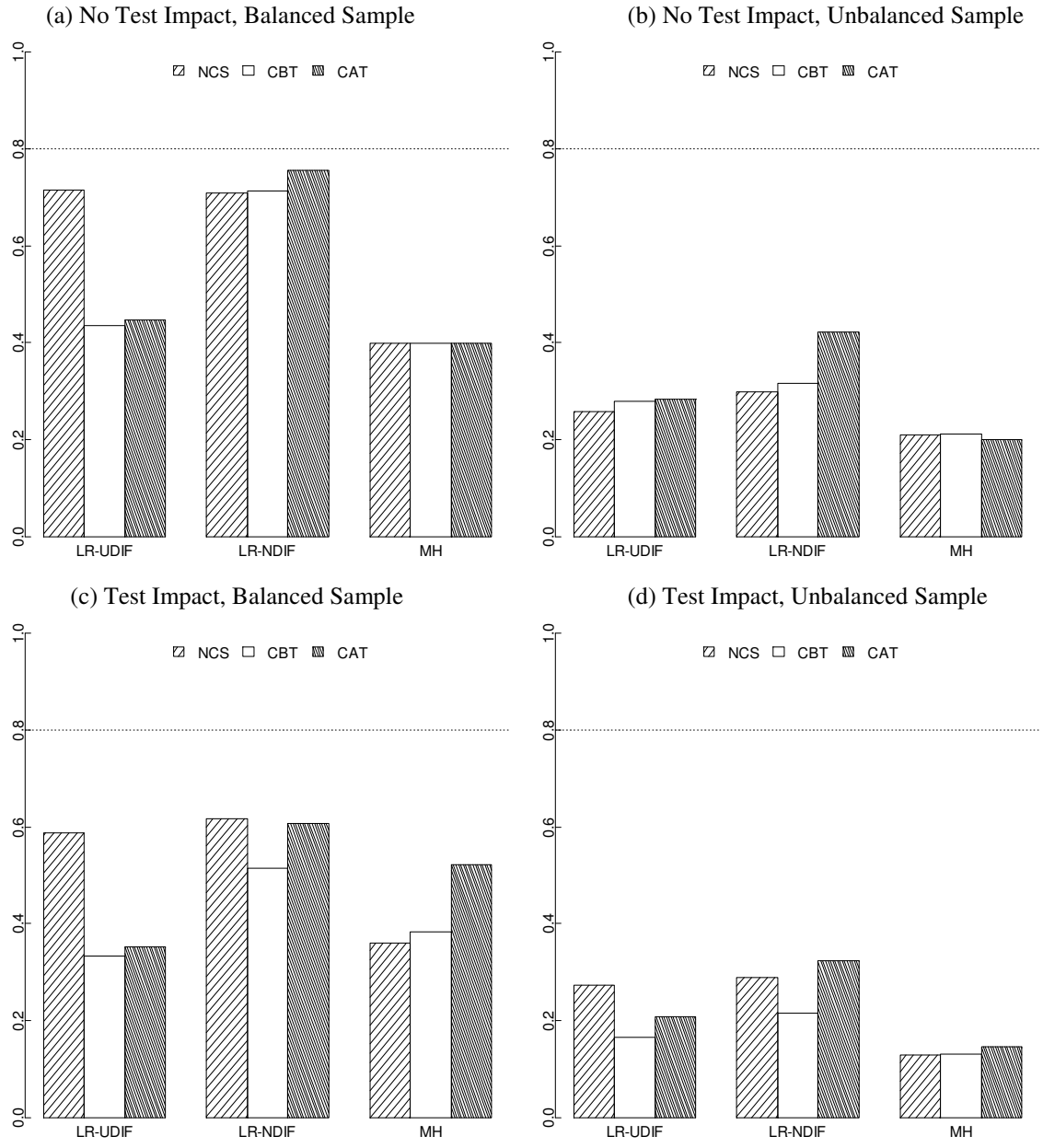
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 65. Power in detecting nonuniform DIF in pretest items when the operational test consisted of 6 items exhibiting uniform DIF with a magnitude of .4 at the beginning of the test.



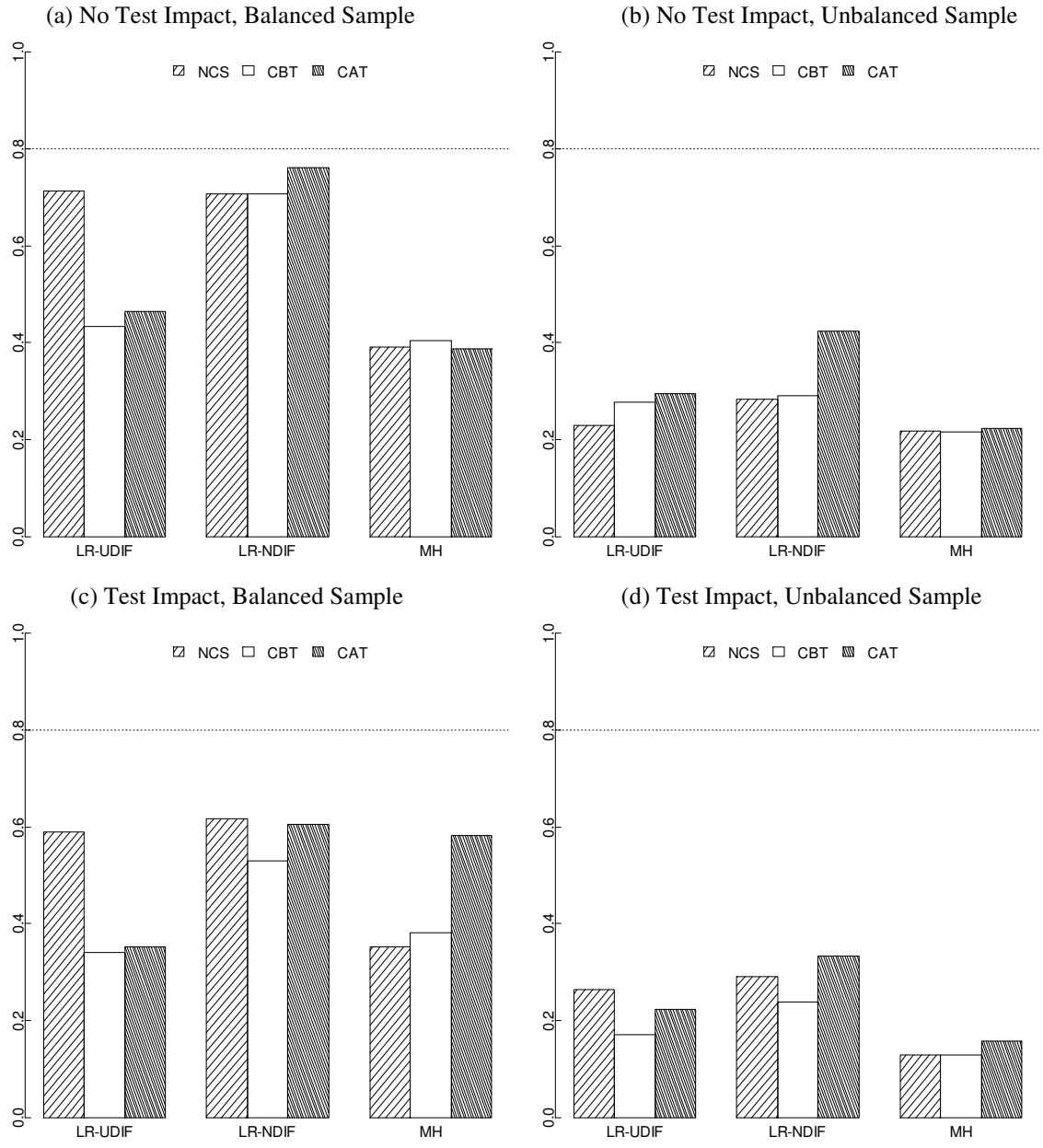
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 66. Power in detecting nonuniform DIF in pretest items when the operational test consisted of 6 items exhibiting nonuniform DIF with a magnitude of .4 at the beginning of the test.



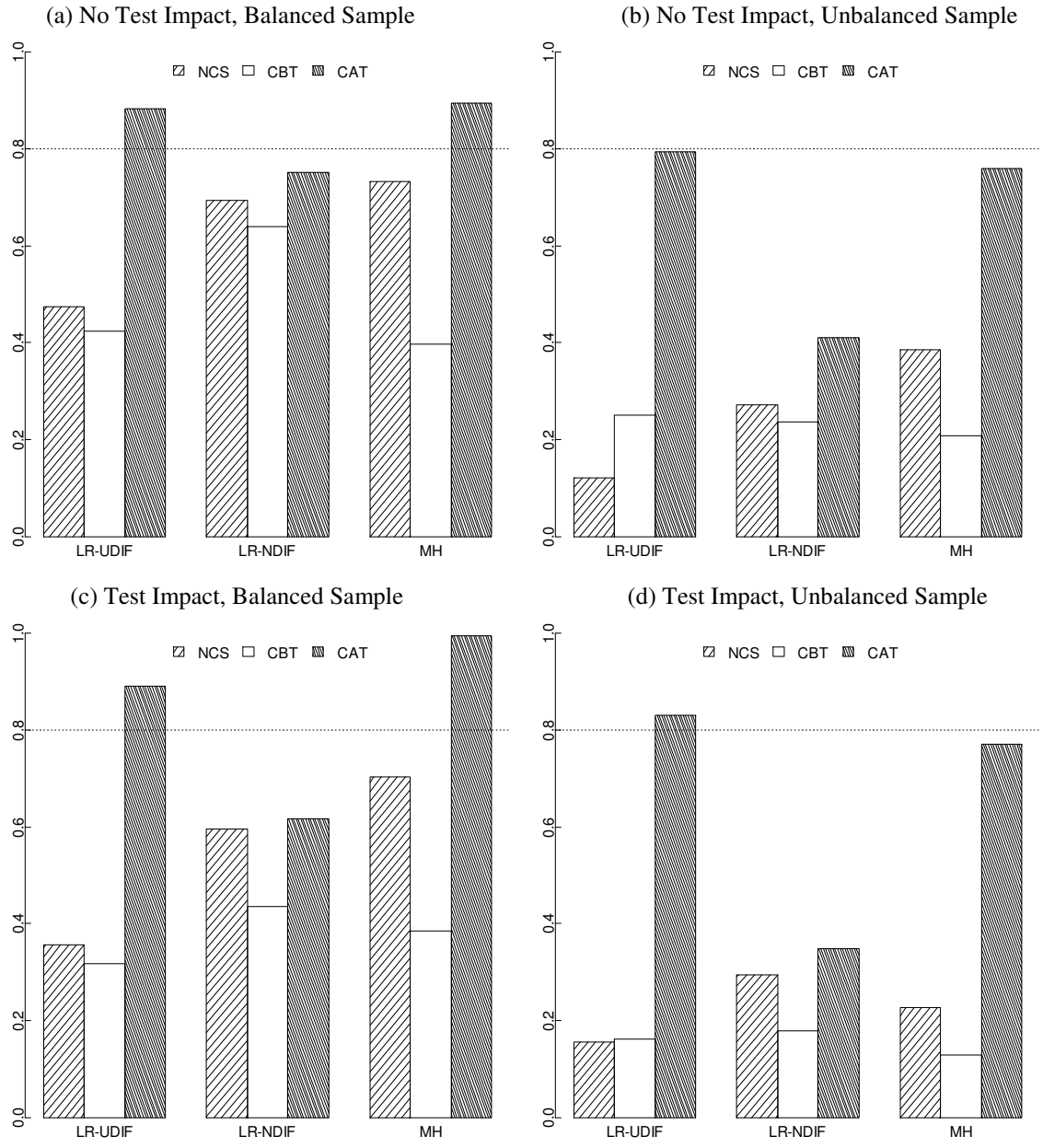
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 67. Power in detecting nonuniform DIF in pretest items when the operational test consisted of 6 items exhibiting both nonuniform and uniform DIF with magnitudes of .4 at the beginning of the test.



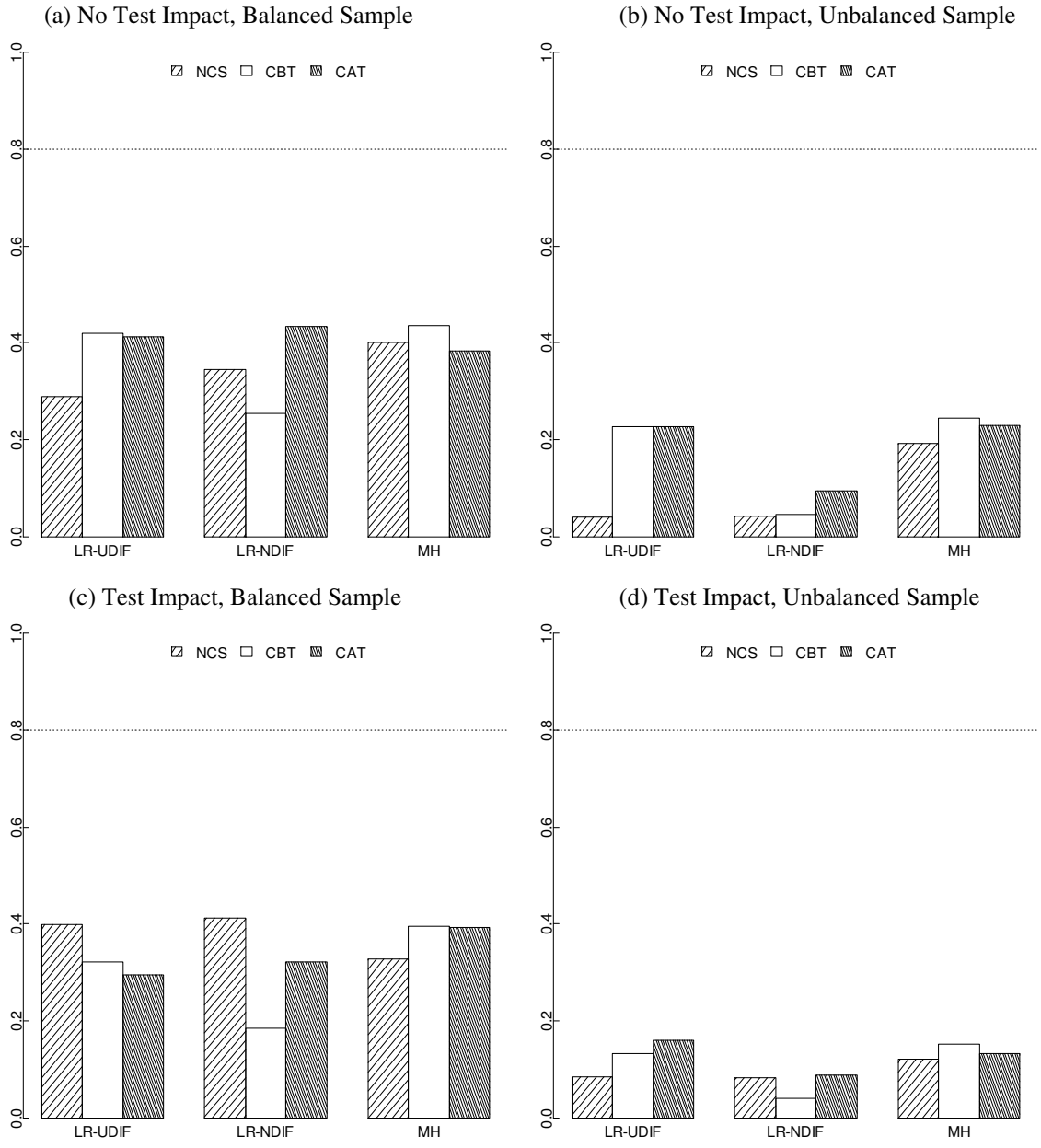
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 68. Power in detecting nonuniform DIF in pretest items when the operational test consisted of 24 items exhibiting uniform DIF with a magnitude of 1.6 at the end of the test.



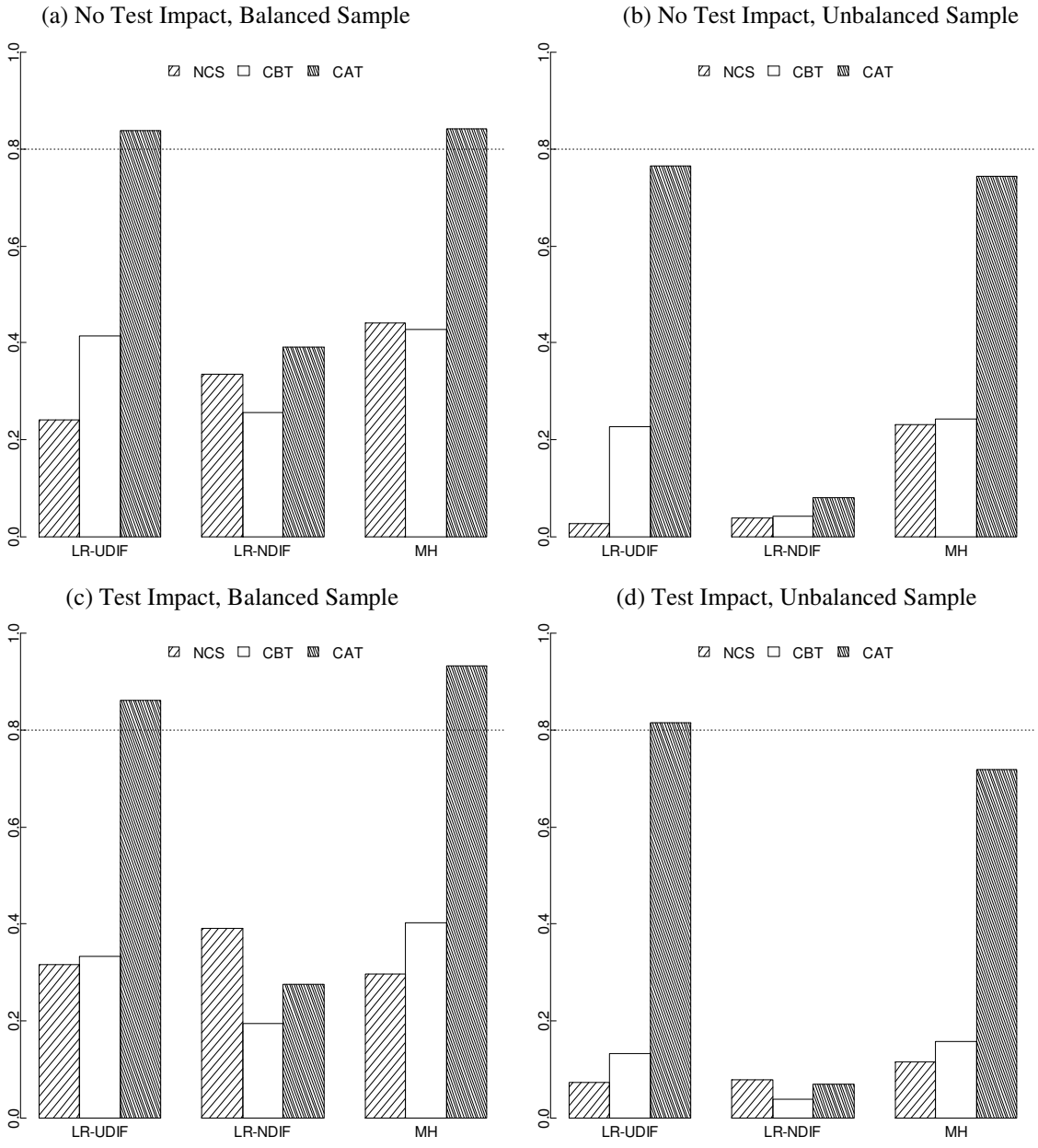
Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 69. Power in detecting nonuniform DIF in pretest items when the operational test consisted of 24 items exhibiting nonuniform DIF with a magnitude of 1.6 at the end of the test.



Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Figure 70. Power in detecting nonuniform DIF in pretest items when the operational test consisted of 24 items exhibiting both nonuniform and uniform DIF with a magnitude of 1.6 at the end of the test.



Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics
Horizontal dotted line indicates the .80 or 80% level.

Chapter 5: Conclusions

This study addressed two issues related to DIF in the context of CAT: 1) Can CAT adjust for the effect of DIF in operational items on the accuracy of the ability estimate ($\hat{\theta}_{CAT}$) and 2) Does matching examinees on $\hat{\theta}_{CAT}$ provide more accurate results of detecting DIF in nonadaptive pretest items than matching on the number-correct score (NCS) and the ability estimate obtained from nonadaptive computer-based testing ($\hat{\theta}_{CBT}$)? To answer the first question, a series of simulations were conducted by varying the level of DIF magnitude (0, .4, 1, and 1.6); DIF type (uniform and nonuniform); DIF contamination or the number of DIF items in a 30-item operational CAT (6, 15, and 24 items out of the 30-item test); and DIF occurrence or the stages of CAT that DIF items were delivered (first, middle, last, and randomly across stages). For the latter research question, test impact ($\mu_R - \mu_F$ = 0 and 1) and sample size ratio (N_R:N_F = 1:1 and 9:1) were also added to the simulation. Interestingly, these simulations revealed both expected and surprising findings to both DIF and CAT.

5.1 Can CAT adjust for the effect of DIF?

As discussed earlier, it was predicted that CAT could adjust for the effect of DIF in operational items if DIF occurred in the first stages of CAT. The results of Study 1 seemed to support such a hypothesis, though with some restrictions. Specifically, CAT successfully adjusted for the effect of DIF at the earlier stages if the number of DIF items and the magnitude of DIF were moderate (.4). In other situations, CAT seemed to reduce the effect of DIF as seen in the trend of SEs which increased when DIF items were delivered and decreased after CAT administered DIF-free subsequent items. However, the self-adjustment mechanism of CAT only was not enough to recover $\hat{\theta}_{CAT}$ from DIF effects.

It was revealed that even a moderate magnitude of uniform DIF at the first stages of CAT could lead to a large gap between $\hat{\theta}_{CAT}$ for the reference and focal groups. As shown in Figure 4a, for example, 24 DIF items with the magnitude of .4 yielded a difference about .5 in $\hat{\theta}_{CAT}$ for examinees from different groups who, in fact, had the

same true ability level. Based on such findings, it can be argued that uniform DIF items with a magnitude of .4 might not be negligible (as classified by most previous studies, see, e.g., Holland and Thayer, 1988) because such DIF items could lead to a moderate to large biased difference in the ability estimate between groups. Any group comparison based on $\hat{\theta}_{CAT}$ obtained from such items will also be invalid.

Another interesting finding was the effect of nonuniform DIF in operational CAT items. It was found that nonuniform DIF items had no effect on $\hat{\theta}_{CAT}$ for the reference group, regardless of DIF contamination, DIF occurrence, and true ability level. As seen in Figures 7-9 and 22-24, the values of bias and RMSE were impressively small (less than .02). The focal group examinees, in contrast, were severely affected by nonuniform DIF, even when a few nonuniform DIF items with the magnitude of .4 were administered at the first stages of CAT. It turned out that the lower-ability examinees in the focal group received higher positive biased $\hat{\theta}_{CAT}$, while the higher-ability examinees in the same group received a lower negative bias. In other words, even within the same group, nonuniform DIF had different effects for examinees with different ability levels.

It should also be noted that when nonuniform DIF items were administered at the first stages of CAT, bias in ability estimate and SE for the focal group were larger than when such items were administered at other stages of CAT. This is because nonuniform DIF in this study was manipulated by decreasing item discrimination parameters for the focal group. As compared to the item difficulty, the item discrimination has a larger influence on item information (Embretson & Reise, 2000), which in turn affects the precision of the ability estimate. If nonuniform DIF occurred at the first stages, CAT would have a poorer start of $\hat{\theta}_{CAT}$ for the focal group. Consequently, the precision of ability estimation would be worse as more nonuniform DIF items were consecutively administered because a less precise $\hat{\theta}_{CAT}$ in a proceeding stage was used to select an item for the next stage. That is, the precision of $\hat{\theta}_{CAT}$ was consecutively lessened by the consecutive nonuniform DIF items. The case of uniform DIF was different because only the difficulty level was affected. CAT could select an easier item in the next stage in which an examinee would have a higher chance to correctly answer the item and recover from the effect of uniform DIF. Thus, bias in $\hat{\theta}_{CAT}$ from uniform DIF items at the first

stages of CAT never exceeded the level of bias from uniform DIF items observed in other stages.

5.2 Did matching examinees on $\widehat{\theta}_{CAT}$ provide more accurate results of detecting DIF in nonadaptive pretest items than NCS and $\widehat{\theta}$?

The results from Study 2 suggested that matching examinees on $\hat{\theta}_{CAT}$ did not provide impressive advantages over the NCS and $\hat{\theta}$ in most of the simulation conditions. Overall, when operational items were contaminated with small DIF magnitude, the three matching variables yielded comparable results of DIF detection in pretest items. However, when the level of DIF contamination in operational items increased, matching examinees on $\hat{\theta}_{CAT}$ led to the worst situation of detecting DIF in pretest items, especially when large-uniform DIF items were used in the operational test. The reason is that, based on Study 1, even small magnitude-uniform DIF yielded a moderate to large difference in ability estimates between groups. Thus, using $\hat{\theta}_{CAT}$ as the matching variable would lead to incomparably matched examinees because the matching variable itself functioned differently across groups. Consequently, DIF detection matching examinees on such scores resulted in a large Type I error rate. This finding fulfilled the gap from the simulation by Lei, Chen, and Yu (2006) who also studied the performance of DIF detection using $\hat{\theta}_{CAT}$ as the matching variable, but did not manipulate DIF in operational CAT items.

It was also evident that DIF in operational items, especially CAT items, led to false identification of DIF type. For instance, when operational items were contaminated by nonuniform DIF (Figures 62), the proportion of statistical significance for the interaction term in logistic regression substantially increased, regardless of matching variables. This meant that logistic regression had a higher chance to mistakenly identify uniform DIF items as nonuniform DIF if the matching variable was affected by nonuniform DIF items in the operational test. On the other hand, when operational items were contaminated by uniform DIF, the main effect in logistic regression and the Mantel-Haenszel statistic tended to be statistically significant more often even if the studied items exhibited only nonuniform DIF (Figures 68 and 70). That is, pretest items exhibiting uniform DIF were

mistakenly identified as nonuniform DIF if the matching variable was obtained from nonuniform-DIF operational items.

5.3 Limitations, implications, and suggestions for future research

Unlike previous studies (e.g., Zwick, Thayer, & Wingersky, 1994a; Zwick, Thayer, & Lewis, 1997; Nandakumar & Roussos, 2001; Lei, Chen, & Yu, 2006) in which DIF was only embedded in pretest items, the present study manipulated DIF in operational CAT items. However, this study assumed that all DIF items were delivered to examinees at the same time, depending on the condition of DIF contamination and DIF occurrence. Specifically, under the condition of 6 DIF items at the first stages of CAT, all examinees would receive six DIF items at the stages 1-6. Such a simulation design was chosen because it was expected to maximize the effect of DIF and provide a better observation of the effect. Nevertheless, in reality, DIF may not always occur at the same stage for all examinees. Thus, it may be worth investigating the effect of DIF when different examinees receive DIF items at different stages.

Another limitation in this study was that no scale purification was implemented in Study 2. However, this was because of the lack of detection method specifically developed for detecting DIF in operational CAT items. In fact, according to the findings discussed throughout this study, DIF in CAT was serious because it not only affected the accuracy of $\hat{\theta}_{CAT}$, but also the results of statistical analyses based on such $\hat{\theta}_{CAT}$. Therefore, a detection method specifically designed for detecting DIF in operational items during the CAT process is needed.

Given the observation of the standard error of estimation (SE) provided in Section 3.4, it is worth considering the use of the SE as an index of DIF in operational items during the CAT process. As discussed in the section, the trend of SE changed when DIF items were administered. Unlike the BIAS and RMSE, SE can be computed solely from $\hat{\theta}_{CAT}$ (i.e., no need of the knowledge of the true ability level) at each stage of CAT. By monitoring the SE values during the CAT process, it should be expected that an abnormal change (e.g., a dramatic increase of the SE value after an item is answered) in the trend of SEs may signal DIF in operational CAT items.

References

Bao, H., Dayton, C. M., & Hendrickson, A. B. (2009). Differential item functioning amplification and cancellation in a reading test. *Practical Assessment, Research & Evaluation, 14(19).* Retrieved from http://pareonline.net/getvn.asp?v=14&n=19.

Bringsjord, E. L. (2001). *Computerized-adaptive versus paper-and-pencil testing environments: An experimental analysis of examinee experience* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3014369).

Budgell, G. R., Raju, N. S., & Quartetti, D. A. (1995). Analysis of differential item functioning in translated assessment instruments. *Applied Psychological Measurement, 19,* 309–321.

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

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

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.

Doran, H. C. (2010). *MiscPsycho: An R package for miscellaneous psychometric analyses.* Retrieved from http://cran.r-project.org/web/packages/MiscPsycho/vignettes/MP.pdf

Dorans, N. & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT.* Lincolnwood, IL: Scientific Software International.

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

Feng, X. (2003). *Statistical detection and estimation of differential item functioning in computerized adaptive testing* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3095579)

Fidalgo, A. M., Mellenbergh, G. J., & Muñiz, J. (2000). Effects of amount of DIF, test length, and purification type on robustness and power of Mantel-Haenszel procedures. *Methods of Psychological Research Online, 5,* 43-53.

Gratia, M. B. (1997). *Gender and ethnicity-based differential item functioning on the Myers-Briggs type indicator* (Master thesis). Retrieved from http://scholar.lib.vt.edu/theses/available/etd-235115949751281/ unrestricted/etd.pdf

Guyer, R. D. (2008). *Effect of early misfit in computerized adaptive testing on the recovery of theta* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3338946)

Guyer, R. D. (2010). *Manual for ScoreAll 4.0: IRT scoring for conventionally administered tests.* St. Paul MN: Assessment Systems Corporation.

Guyer, R., & Thompson, N.A., (2011). *User's manual for Xcalibre 4.1.* St. Paul MN: Assessment Systems Corporation.

Han, K. T. (2010). *SimulCAT: Windows application that simulates computerized adaptive test administration.* Retrieved from http://www.hantest.net/ simulcat.

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

Herrera, A.-N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel–Haenszel and logistic regression techniques. *Quality & Quantity, 42,* 739-755.

Hidalgo, M. D., & López-Pina, J. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement, 64,* 903-915.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer, & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kennedy, M. (1994). *The influence of sample size, effect size, and percentage of DIF*

items on the performance of the Mantel-Haenszel and logistic regression DIF identification procedures (Doctoral dissertation, University of Ottawa, Canada). Retrieved from http://www.ruor.uottawa.ca/en/ bitstream/handle/10393/6884/MM00610.PDF?sequence=1

Lei, P.-W., Chen, S.-Y., & Yu, L. (2006). Comparing methods of assessing differential item functioning in a computerized adaptive testing environment. *Journal of Educational Measurement,* 43, 254-264.

Li, Z. (2009). *Impact of differential item functioning on statistical conclusions* (Doctoral dissertation, University of British Columbia, Canada). Retrieved from https://circle.ubc.ca/bitstream/handle/2429/14680/ ubc_2010_spring_li_zhen.pdf?sequence=1

Li, Z., & Zumbo, B. D. (2009). Impact of differential item functioning on subsequent statistical conclusions based on observed test score data. *Psicologica, 30,* 343-370.

Linacre, J. M. (2000). Computer-adaptive testing: A methodology whose time has come. *MESA Memorandum No. 69.* Retrieved from *www.rasch.org/memo69.pdf*

Magis, D., & Raîche, G. (in press). *catR: an R package for computerized adaptive testing.* Applied Psychological Measurement. Retrieved from http://ppw.kuleuven.be/okp/_pdf/Magis2011CARPF.pdf

Nandakumar, R., & Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (LSAC Research Report 97-11). Retrieved from Law School Admission Council website: http://www.lsac.org/LSACResources/Research/CT/CT-97-11.pdf

Nandakumar, R., & Roussos, L. (2004). Evaluation of the CATSIB DIF procedure in a pretest setting. *Journal of Educational and Behavioral Statistics, 29,* 177-199.

Nandakumar, R., Banks, C. J., & Roussos, L. A. (2006). *Kernel-smoothed DIF detection procedure for computerized adaptive tests* (LSAC Research Report 00-08). Retrieved from Law School Admission Council website: http://www.lsac.org/LSACResources/Research/CT/CT-00-08.pdf

Partchev, I. (2009). *irtoys: Simple interface to the estimation and plotting of IRT models.* Retrieved from http://cran.r-project.org/web/packages/irtoys/ irtoys.pdf.

R Development Core Team (2007). *R: A language and environment for statistical computing.* Retrieved from http://www.R-project.org/.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197-207.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory Analyses. *Journal of Statistical Software, 17,* 1–25. Retrieved from http://www.jstatsoft.org/v17i05/.

Roznowski, M. & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: Do biased items result in poor measurement? *Educational and Psychological Measurement, 59,* 248-271.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group difference in trace lines. In H. Wainer & H. Braun (Eds.) *Test validity* (pp. 147–169). Hillsdale, NJ: Erlbaum.

Thompson, N. A., & Weiss, D. A. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16(1).* Retrieved from http://pareonline.net/getvn.asp?v=16&n=1

van der Linden,W. J. & Pashley, P.J. (2010). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.) *Elements of adaptive testing* (pp. 3–30). New York, NY: Springer.

Walker, C. M., Beretvas, S. N., & Ackerman, T. (2001). An examination of conditioning variables used in computer adaptive testing for DIF analyses. *Applied Measurement in Education, 14,* 3-16.

Walstad, W. B. & Robson, D. (1997). Differential item functioning and male-female

differences on multiple-choice tests in economics. *The Journal of Economic Education, 28,* 155-171.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2007). A meta-analysis of testing mode effects in Grade K–12 mathematics tests. *Educational and Psychological Measurement, 67,* 219-238.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K-12 reading assessment: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*, 5-4.

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

Weiss, D. J. & Guyer, R. (2010*). Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J. (2008). *Manual for the FastTEST professional testing system, version 2.* St. Paul, MN: Assessment Systems Corporation.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and Psychological Measurement, 57,* 412-421.

Zwick, R. (2010). The investigation of differential item functioning in adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 331-352). New York, NY: Springer.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to a computerized adaptive test. *Applied Psychological Measurement, 26,* 57-76.

Zwick, R., Thayer, D. T., & Lewis, C. (1997). *An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis* (ETS Research Report RR-97-21). Retrieved from Educational Testing Service website: http://www.ets.org/Media/Research/pdf/RR-97-21.pdf

Zwick, R., Thayer, D. T., & Wingersky, M. (1994a). A simulation study of methods for assessing differential item functioning in computerized adaptive tests. *Applied*

Psychological Measurement, 18, 121-140.

Zwick, R., Thayer, D. T., & Wingersky, M. (1994b). *DIF analyses for pretest items in computer-adaptive testing* (ETS Research Report RR-94-33). Retrieved from Education Resources Information Center website: http://www.eric.ed.gov/ERICWebPortal/recordDetail?accno=ED382660

Zwick, R., Thayer, D. T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32,* 341-363.

Appendix A: Item parameters in the generated bank for Study 1.

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
1	0.892	-3.572	0.203	38	0.857	-3.042	0.088	75	0.966	-2.528	0.164
2	0.902	-3.560	0.236	39	0.898	-2.989	0.018	76	1.056	-2.508	0.270
3	1.046	-3.534	0.131	40	0.808	-2.963	0.036	77	0.968	-2.466	0.218
4	0.902	-3.531	0.202	41	0.961	-2.955	0.025	78	0.967	-2.435	0.272
5	0.894	-3.520	0.102	42	0.843	-2.937	0.081	79	0.783	-2.420	0.259
6	0.854	-3.512	0.139	43	1.003	-2.935	0.261	80	0.839	-2.418	0.232
7	0.708	-3.503	0.280	44	0.990	-2.924	0.262	81	0.943	-2.416	0.075
8	1.008	-3.501	0.271	45	1.116	-2.884	0.262	82	0.872	-2.413	0.034
9	0.840	-3.493	0.050	46	0.951	-2.884	0.012	83	0.999	-2.396	0.037
10	0.794	-3.474	0.069	47	0.833	-2.868	0.141	84	0.938	-2.390	0.085
11	0.891	-3.470	0.144	48	1.127	-2.854	0.085	85	1.029	-2.385	0.142
12	0.884	-3.448	0.227	49	0.955	-2.839	0.060	86	0.829	-2.373	0.272
13	0.898	-3.427	0.157	50	0.841	-2.799	0.079	87	0.960	-2.353	0.269
14	0.996	-3.414	0.300	51	1.269	-2.779	0.128	88	0.854	-2.345	0.021
15	0.934	-3.402	0.248	52	0.870	-2.775	0.292	89	0.938	-2.345	0.142
16	0.816	-3.394	0.033	53	0.872	-2.768	0.097	90	0.859	-2.343	0.242
17	0.892	-3.334	0.027	54	0.914	-2.766	0.021	91	1.070	-2.311	0.098
18	0.963	-3.271	0.287	55	0.804	-2.765	0.143	92	0.830	-2.306	0.235
19	0.866	-3.270	0.086	56	0.907	-2.723	0.160	93	0.975	-2.301	0.163
20	0.861	-3.261	0.226	57	0.932	-2.719	0.194	94	0.850	-2.283	0.283
21	0.978	-3.259	0.063	58	0.927	-2.712	0.091	95	0.853	-2.272	0.127
22	0.856	-3.243	0.066	59	0.842	-2.684	0.105	96	1.110	-2.222	0.279
23	0.982	-3.215	0.007	60	1.002	-2.652	0.180	97	0.790	-2.192	0.286
24	1.011	-3.203	0.245	61	0.895	-2.639	0.050	98	1.076	-2.185	0.236
25	1.018	-3.201	0.152	62	0.969	-2.635	0.205	99	0.921	-2.127	0.229
26	1.074	-3.197	0.071	63	0.980	-2.627	0.252	100	0.883	-2.098	0.204
27	0.888	-3.187	0.185	64	0.851	-2.615	0.241	101	0.990	-2.094	0.252
28	0.803	-3.182	0.288	65	0.861	-2.614	0.264	102	0.841	-2.079	0.283
29	0.980	-3.164	0.085	66	0.855	-2.602	0.170	103	0.953	-2.075	0.292
30	0.981	-3.159	0.201	67	1.064	-2.588	0.200	104	0.795	-2.064	0.271
31	0.970	-3.141	0.238	68	0.696	-2.586	0.009	105	0.911	-2.061	0.029
32	0.989	-3.138	0.009	69	0.851	-2.571	0.041	106	0.852	-2.060	0.005
33	0.921	-3.094	0.010	70	0.939	-2.569	0.019	107	1.063	-2.056	0.272
34	0.853	-3.083	0.096	71	0.904	-2.565	0.253	108	0.783	-2.046	0.245
35	0.873	-3.081	0.046	72	1.059	-2.545	0.097	109	1.126	-2.032	0.266
36	1.055	-3.072	0.003	73	0.807	-2.529	0.206	110	0.801	-2.017	0.127
37	1.026	-3.043	0.155	74	0.922	-2.528	0.081	111	0.913	-2.006	0.002

Appendix A (cont.): Item parameters in the generated bank for Study 1.

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
112	0.823	-1.996	0.192	149	0.838	-1.465	0.014	186	0.956	-1.063	0.239
113	0.946	-1.978	0.227	150	0.876	-1.462	0.162	187	0.894	-1.063	0.258
114	0.845	-1.976	0.270	151	1.143	-1.458	0.036	188	0.904	-1.057	0.112
115	0.955	-1.969	0.281	152	0.812	-1.428	0.167	189	0.795	-1.032	0.141
116	0.794	-1.955	0.153	153	0.976	-1.415	0.141	190	0.921	-1.031	0.130
117	0.820	-1.952	0.001	154	0.969	-1.407	0.264	191	1.047	-1.012	0.104
118	1.037	-1.930	0.047	155	1.071	-1.406	0.141	192	1.046	-1.012	0.149
119	1.043	-1.929	0.092	156	1.078	-1.387	0.099	193	0.808	-1.010	0.030
120	0.929	-1.927	0.179	157	1.021	-1.375	0.216	194	0.951	-0.994	0.111
121	0.855	-1.924	0.258	158	0.946	-1.373	0.239	195	0.803	-0.988	0.092
122	0.914	-1.924	0.193	159	1.103	-1.372	0.112	196	1.049	-0.983	0.091
123	0.877	-1.921	0.164	160	0.764	-1.333	0.037	197	0.847	-0.973	0.130
124	0.800	-1.887	0.095	161	0.833	-1.331	0.060	198	0.932	-0.972	0.293
125	1.043	-1.862	0.034	162	0.881	-1.328	0.106	199	0.931	-0.942	0.182
126	0.918	-1.861	0.242	163	1.052	-1.311	0.231	200	1.021	-0.940	0.051
127	0.815	-1.822	0.140	164	0.933	-1.268	0.134	201	0.952	-0.920	0.272
128	0.968	-1.821	0.218	165	0.773	-1.266	0.178	202	0.866	-0.907	0.216
129	0.789	-1.791	0.183	166	0.877	-1.258	0.192	203	0.880	-0.867	0.278
130	0.829	-1.783	0.160	167	0.955	-1.249	0.075	204	1.008	-0.863	0.129
131	0.807	-1.770	0.085	168	0.894	-1.249	0.137	205	0.918	-0.848	0.122
132	0.806	-1.770	0.176	169	0.894	-1.248	0.091	206	0.959	-0.838	0.293
133	0.821	-1.765	0.084	170	0.893	-1.232	0.118	207	0.810	-0.832	0.103
134	0.666	-1.712	0.091	171	1.011	-1.231	0.144	208	0.970	-0.823	0.189
135	0.933	-1.634	0.060	172	0.896	-1.205	0.163	209	1.006	-0.816	0.109
136	0.994	-1.628	0.177	173	0.816	-1.204	0.166	210	0.954	-0.795	0.025
137	0.744	-1.608	0.300	174	1.038	-1.197	0.264	211	1.058	-0.786	0.267
138	0.973	-1.596	0.274	175	0.934	-1.185	0.100	212	0.891	-0.779	0.028
139	1.040	-1.586	0.184	176	0.877	-1.177	0.244	213	1.081	-0.777	0.165
140	0.937	-1.579	0.087	177	0.887	-1.169	0.131	214	0.900	-0.755	0.003
141	0.884	-1.574	0.172	178	0.940	-1.168	0.043	215	0.947	-0.754	0.053
142	0.909	-1.574	0.251	179	1.038	-1.153	0.207	216	0.932	-0.701	0.096
143	0.847	-1.567	0.222	180	0.885	-1.136	0.196	217	0.888	-0.689	0.109
144	1.014	-1.536	0.020	181	0.848	-1.131	0.134	218	0.902	-0.677	0.250
145	0.996	-1.529	0.123	182	0.904	-1.125	0.118	219	0.787	-0.631	0.278
146	1.049	-1.498	0.192	183	0.996	-1.109	0.231	220	0.900	-0.613	0.079
147	0.918	-1.497	0.205	184	1.101	-1.105	0.299	221	1.002	-0.613	0.052
148	1.046	-1.492	0.015	185	0.996	-1.073	0.043	222	0.878	-0.571	0.266

Appendix A (cont.): Item parameters in the generated bank for Study 1.

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
223	0.941	-0.553	0.078	260	1.106	-0.061	0.009	297	0.926	0.511	0.229
224	0.790	-0.548	0.204	261	0.996	-0.047	0.067	298	0.901	0.516	0.202
225	0.815	-0.547	0.069	262	0.848	-0.032	0.155	299	0.881	0.541	0.093
226	0.784	-0.537	0.149	263	0.930	0.017	0.136	300	0.799	0.551	0.060
227	0.877	-0.529	0.172	264	0.931	0.040	0.217	301	0.938	0.551	0.021
228	0.891	-0.521	0.117	265	0.984	0.049	0.143	302	1.099	0.564	0.175
229	0.898	-0.505	0.075	266	0.966	0.076	0.069	303	1.060	0.585	0.142
230	0.904	-0.500	0.134	267	1.069	0.104	0.185	304	0.775	0.593	0.173
231	0.799	-0.489	0.008	268	0.775	0.113	0.281	305	0.961	0.642	0.206
232	0.970	-0.483	0.109	269	0.774	0.129	0.043	306	0.948	0.656	0.176
233	1.015	-0.442	0.153	270	0.942	0.138	0.078	307	0.961	0.660	0.142
234	0.880	-0.422	0.077	271	0.916	0.139	0.034	308	0.843	0.663	0.162
235	0.862	-0.403	0.189	272	1.018	0.149	0.118	309	1.004	0.677	0.059
236	0.856	-0.386	0.087	273	0.852	0.161	0.058	310	0.902	0.686	0.201
237	0.867	-0.374	0.010	274	0.942	0.163	0.249	311	0.803	0.697	0.219
238	1.003	-0.363	0.272	275	0.812	0.189	0.278	312	0.947	0.711	0.026
239	0.913	-0.355	0.242	276	0.915	0.191	0.101	313	0.762	0.725	0.216
240	1.005	-0.339	0.132	277	0.908	0.194	0.201	314	0.828	0.727	0.091
241	1.174	-0.339	0.027	278	1.017	0.196	0.212	315	0.854	0.756	0.043
242	0.897	-0.321	0.257	279	0.869	0.198	0.137	316	0.851	0.760	0.033
243	0.843	-0.306	0.290	280	1.005	0.198	0.119	317	0.988	0.767	0.051
244	0.762	-0.305	0.153	281	0.869	0.213	0.063	318	0.751	0.772	0.128
245	0.766	-0.305	0.052	282	0.909	0.233	0.050	319	0.984	0.773	0.217
246	0.819	-0.296	0.235	283	0.936	0.241	0.165	320	0.858	0.781	0.264
247	0.874	-0.260	0.009	284	0.972	0.258	0.108	321	0.834	0.793	0.101
248	0.822	-0.253	0.174	285	0.789	0.258	0.254	322	0.772	0.800	0.227
249	0.926	-0.243	0.151	286	1.275	0.303	0.039	323	0.866	0.802	0.086
250	1.023	-0.239	0.237	287	1.077	0.317	0.240	324	0.990	0.820	0.003
251	0.868	-0.216	0.290	288	0.887	0.324	0.238	325	1.082	0.846	0.165
252	0.967	-0.205	0.284	289	0.841	0.374	0.032	326	0.972	0.852	0.067
253	0.942	-0.165	0.236	290	0.861	0.384	0.130	327	0.872	0.859	0.167
254	0.792	-0.137	0.202	291	0.749	0.393	0.233	328	0.943	0.864	0.030
255	1.103	-0.127	0.096	292	0.908	0.405	0.132	329	0.938	0.872	0.091
256	1.187	-0.122	0.242	293	0.795	0.410	0.195	330	0.923	0.878	0.217
257	1.002	-0.119	0.130	294	1.032	0.439	0.208	331	1.122	0.889	0.167
258	1.057	-0.108	0.173	295	0.971	0.459	0.129	332	0.846	0.904	0.144
259	0.946	-0.077	0.036	296	0.871	0.499	0.007	333	1.126	0.926	0.111

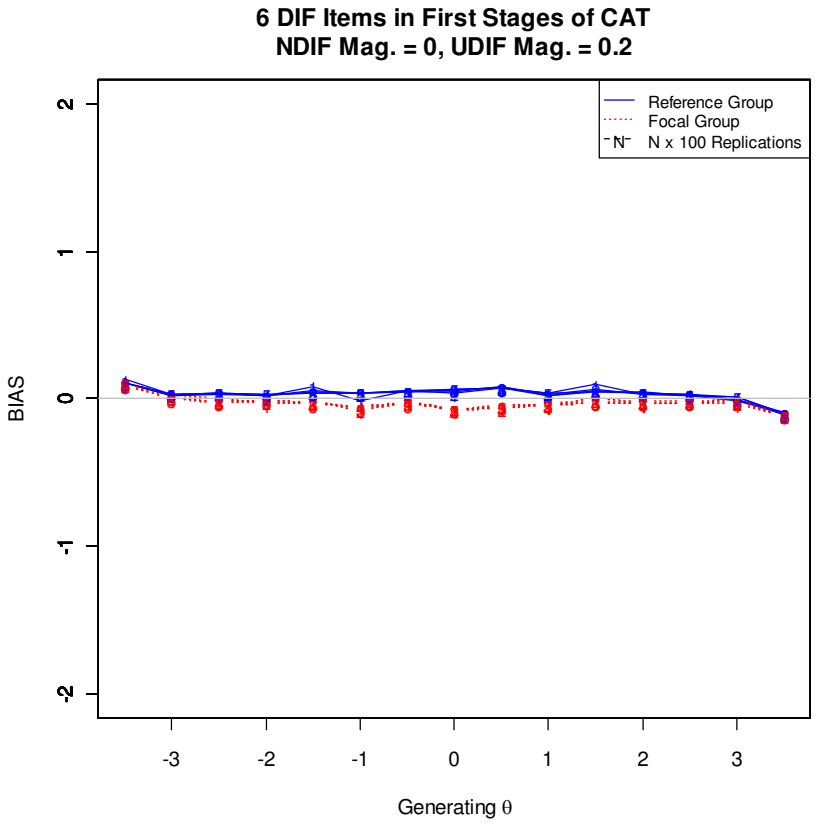
Appendix A (cont.): Item parameters in the generated bank for Study 1.

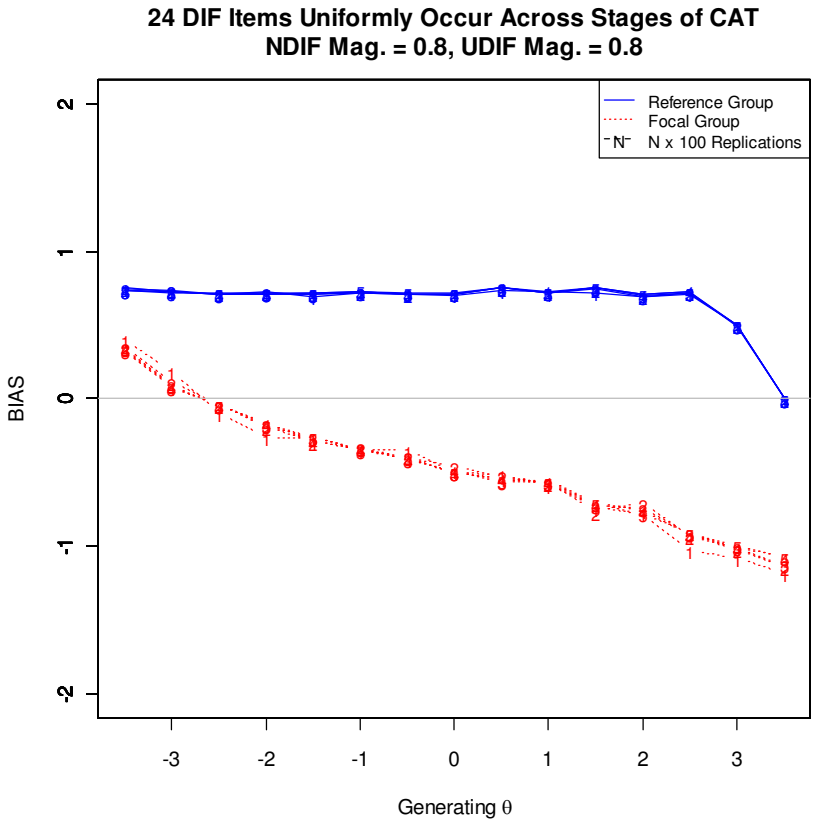
Item	*a*	*b*	*c*	Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
334	0.980	0.931	0.050	371	0.891	1.518	0.255	408	0.858	2.126	0.206
335	0.962	0.942	0.295	372	0.867	1.540	0.170	409	0.885	2.134	0.051
336	0.716	0.987	0.245	373	0.938	1.579	0.235	410	0.924	2.136	0.015
337	1.096	1.009	0.080	374	0.974	1.593	0.045	411	0.955	2.143	0.056
338	1.018	1.012	0.048	375	0.944	1.617	0.259	412	1.017	2.144	0.255
339	0.871	1.015	0.138	376	1.026	1.622	0.261	413	1.008	2.145	0.022
340	1.055	1.028	0.119	377	0.836	1.660	0.013	414	0.845	2.201	0.284
341	1.091	1.046	0.246	378	1.036	1.673	0.142	415	0.784	2.203	0.085
342	1.037	1.047	0.173	379	0.800	1.682	0.112	416	0.838	2.222	0.043
343	0.807	1.051	0.037	380	1.022	1.699	0.110	417	0.761	2.238	0.028
344	0.981	1.061	0.274	381	0.938	1.720	0.175	418	0.788	2.240	0.211
345	0.955	1.069	0.197	382	0.773	1.737	0.119	419	0.964	2.260	0.098
346	0.898	1.071	0.006	383	1.032	1.776	0.192	420	0.730	2.284	0.181
347	0.773	1.095	0.135	384	0.864	1.784	0.242	421	0.927	2.310	0.133
348	1.007	1.114	0.108	385	0.910	1.785	0.154	422	0.929	2.311	0.140
349	0.947	1.128	0.111	386	0.924	1.799	0.183	423	0.924	2.344	0.095
350	0.898	1.131	0.074	387	1.019	1.815	0.140	424	0.929	2.346	0.067
351	0.961	1.142	0.148	388	1.170	1.815	0.244	425	0.917	2.350	0.102
352	1.072	1.162	0.012	389	1.003	1.831	0.101	426	0.958	2.400	0.050
353	0.906	1.164	0.249	390	0.957	1.843	0.023	427	0.986	2.424	0.172
354	0.868	1.204	0.248	391	1.102	1.869	0.082	428	0.932	2.428	0.213
355	0.670	1.211	0.260	392	0.882	1.883	0.037	429	0.890	2.429	0.188
356	0.911	1.244	0.048	393	0.951	1.884	0.220	430	0.982	2.455	0.178
357	0.854	1.273	0.074	394	0.919	1.891	0.116	431	0.972	2.459	0.245
358	0.804	1.295	0.070	395	0.754	1.896	0.093	432	0.949	2.461	0.105
359	0.873	1.324	0.057	396	0.914	1.921	0.166	433	0.877	2.470	0.111
360	1.038	1.329	0.045	397	1.020	1.922	0.076	434	0.825	2.479	0.044
361	1.051	1.342	0.100	398	1.058	1.942	0.191	435	0.967	2.507	0.184
362	0.890	1.352	0.202	399	0.978	1.949	0.198	436	0.901	2.518	0.023
363	0.846	1.356	0.279	400	1.036	1.970	0.183	437	0.915	2.548	0.012
364	0.835	1.376	0.104	401	0.913	1.971	0.135	438	0.989	2.556	0.153
365	0.783	1.399	0.002	402	0.917	1.976	0.075	439	0.917	2.568	0.185
366	0.956	1.404	0.260	403	0.841	1.976	0.063	440	0.895	2.585	0.170
367	0.892	1.421	0.239	404	0.749	1.976	0.003	441	0.959	2.606	0.178
368	0.795	1.495	0.190	405	0.883	2.005	0.217	442	0.891	2.701	0.249
369	1.081	1.501	0.179	406	0.761	2.052	0.202	443	0.951	2.728	0.105
370	0.962	1.516	0.167	407	0.898	2.053	0.182	444	0.950	2.790	0.293

Appendix A (cont.): Item parameters in the generated bank for Study 1.

Item	*a*	*b*	*c*	Item	*a*	*b*	*c*	Item	*a*	*b*	*c*
445	0.866	2.791	0.138	464	0.912	3.157	0.043	483	1.003	3.408	0.062
446	0.899	2.793	0.169	465	1.113	3.173	0.027	484	0.877	3.429	0.230
447	0.924	2.873	0.074	466	0.982	3.177	0.167	485	0.851	3.446	0.107
448	0.962	2.888	0.200	467	0.970	3.199	0.059	486	0.852	3.450	0.230
449	0.838	2.893	0.040	468	0.969	3.208	0.010	487	0.953	3.462	0.001
450	0.935	2.902	0.071	469	0.904	3.231	0.166	488	0.823	3.465	0.248
451	0.894	2.903	0.269	470	0.861	3.239	0.195	489	0.839	3.467	0.250
452	0.853	2.936	0.016	471	0.836	3.252	0.261	490	0.960	3.478	0.091
453	0.786	2.951	0.075	472	0.918	3.263	0.085	491	0.841	3.482	0.238
454	1.000	2.984	0.046	473	0.972	3.272	0.021	492	0.947	3.483	0.113
455	1.023	3.024	0.061	474	0.881	3.273	0.283	493	1.070	3.489	0.074
456	0.847	3.034	0.199	475	0.898	3.275	0.034	494	0.826	3.528	0.213
457	0.963	3.041	0.003	476	1.004	3.289	0.028	495	0.951	3.536	0.270
458	0.854	3.048	0.174	477	0.837	3.300	0.093	496	0.944	3.557	0.167
459	0.855	3.092	0.243	478	0.902	3.313	0.204	497	0.866	3.564	0.274
460	1.005	3.115	0.187	479	0.872	3.332	0.148	498	1.014	3.574	0.184
461	0.838	3.120	0.213	480	0.841	3.337	0.069	499	0.989	3.585	0.024
462	1.091	3.126	0.256	481	0.956	3.359	0.243	500	0.958	3.589	0.047
463	0.910	3.134	0.207	482	0.883	3.405	0.158				

Appendix B: The average signed bias obtained from two extreme conditions after 100, 200, 300, 400, and 500 replications.



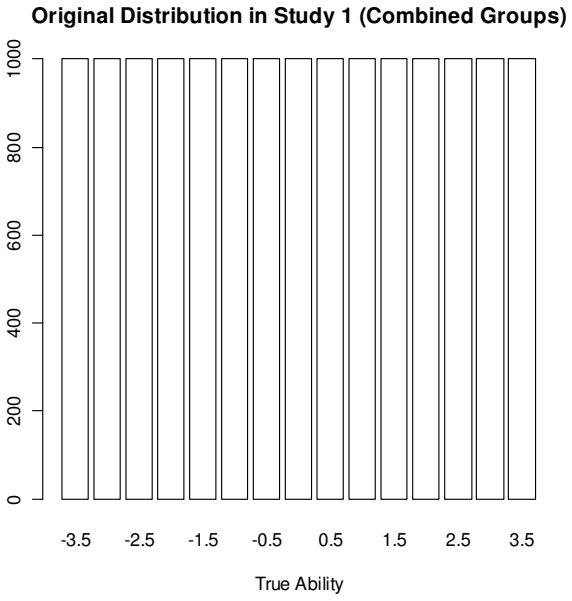


Appendix C: Item parameters of the pretest items in Study 2

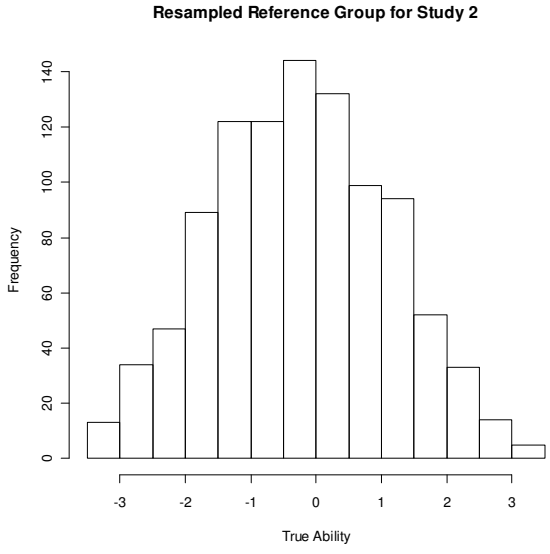
Item	No DIF		Uniform DIF				Nonuniform DIF			
	$a_R = a_F$	$b_R = b_F$	$a_R = a_F$	b_R	b_F	Mag.	a_R	a_R	$b_R = b_F$	Mag.
1	.74	–1.95	1.00	–1.30	–2.00	.60	2.01	.90	–1.50	.43
2	.74	–1.30	1.00	–1.30	–1.65	.30	1.97	.70	–1.50	.64
3	.74	–.65	1.00	–1.30	–.95	.30	1.79	.56	–1.50	.85
4	.74	.00	1.00	–1.30	–.60	.60	1.68	.47	–1.50	1.06
5	.74	.65	1.00	–.65	–1.35	.60	.72	.50	.00	.43
6	.74	1.30	1.00	–.65	–1.00	.30	.80	.46	.00	.64
7	.74	1.95	1.00	–.65	–.30	.30	.91	.43	.00	.85
8	1.00	–1.95	1.00	–.65	.05	.60	1.03	.40	.00	1.06
9	1.00	–1.30	1.00	.00	–.70	.60	2.01	.90	.00	.43
10	1.00	–.65	1.00	.00	–.35	.30	1.97	.70	.00	.64
11	1.00	.00	1.00	.00	.35	.30	1.79	.56	.00	.85
12	1.00	.65	1.00	.00	.70	.60	1.68	.47	.00	1.06
13	1.00	1.30	1.00	.65	–.05	.60	.72	.50	1.50	.43
14	1.00	1.95	1.00	.65	.30	.30	.80	.46	1.50	.64
15	1.50	.00	1.00	.65	1.00	.30	.91	.43	1.50	.85
16	1.50	1.95	1.00	.65	1.35	.60	1.03	.40	1.50	1.06

Note. The *c* parameter is fixed at .15 for all items. This table is adapted from Table 1 in Lei, Chen, & Yu (2006). Mag. = magnitude of DIF measured as the unsigned area between the item characteristic functions for the reference and focal groups.

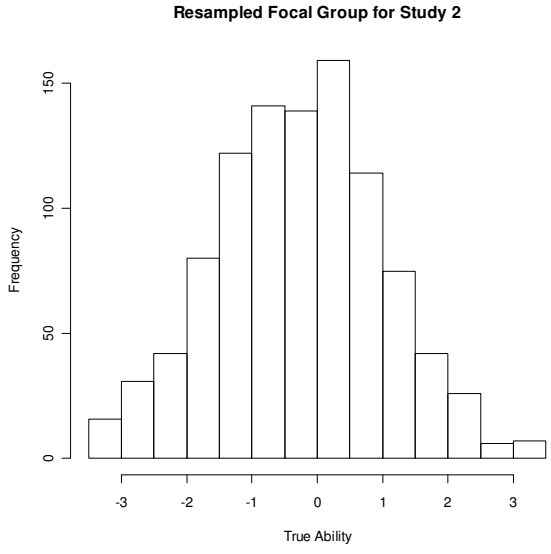
Appendix D: Histogram of true ability levels originally simulated for a condition in Study 1 (a) and histograms of true ability levels for the reference group (b) and the focal group (c) resampled from (a) for the corresponding condition in Study 2.



(a)



(b)



(c)

Appendix E: The R code for CAT simulations in Study 1 and resampling algorithm in Study 2

E.1 Subfunctions for various IRT-based computations

```
#Identify ID of operational items that exhibit DIF
get.dif.id <- function(con, occ){
 #1 = first, 2 = middle, 3 = last, 4 = uniformly across stages of CAT
 if (occ == 1){
    dif.id <- seq(1, con)
 } else if (occ == 2){
     dif.id <- seq(trunc((30 - con)/2), length.out = con)
   } else if (occ == 3){
       dif.id <- seq((31 - con), 30)
     } else if (occ == 4){
         dif.id <- sort(c(sample( 1:10, con/3, replace = FALSE),
                          sample(11:20, con/3, replace = FALSE),
                          sample(21:30, con/3, replace = FALSE)))
       } else if (occ == 0){}    ###This is for the baseline condition
return(dif.id)
}#get.dif.id

#Compute expected/theoretical item information function (iif)
f.iif.exp <- function(theta, abc){
 D <- 1.7
 out <- c()
 if (is.matrix(abc) == TRUE){      #if input abc as a matrix
     aa <- abc[,2]; bb <- abc[,3]; cc <- abc[,4]; j.id <- abc[,1]
     ni  <- length(theta)
     nj  <- length(aa)
     iif <- matrix(0, nrow = ni, ncol = nj)
     for (i in 1:ni){
      for (j in 1:nj){
       iif[i,j] <- (D*D*aa[j]*aa[j]*(1-cc[j]))/
                   ((cc[j]+exp( D*aa[j]*(theta[i]-bb[j]) ))*
                    (    1+exp(-D*aa[j]*(theta[i]-bb[j]) ))^2 )
       out <- rbind(out, cbind(theta[i], j.id[j], aa[j], bb[j], cc[j],
                               iif[i,j]))
      }#for j
     }#for i
 }#if
 if (is.matrix(abc) == FALSE){     #if input abc as a vector
     aa <- abc[1];  bb <- abc[2];  cc <- abc[3]; j.id <- 1
     iif <- (D*D*aa*aa*(1-cc))/
                 ((cc+exp( D*aa*(theta-bb) ))*
                 (  1+exp(-D*aa*(theta-bb) ))^2 )
       out <- rbind(out, cbind(theta, j.id, aa, bb, cc, iif))
 }
 colnames(out) <- c("theta", "j.id", "a", "b", "c", "iif")
 return(out)
}#f.iif.exp
```

```
#Compute item response function for 3PLM
irf.3pl <- function(theta, abc){
 D <- 1.7; a <- abc[1]; b <- abc[2]; c <- abc[3]
 irf <- c +(1-c)*(exp(D*a*(theta-b)))/(1+exp(D*a*(theta-b)))
 return(irf)
}#irf.3pl

#Check response pattern
resp.chk <- function(resp){
 if (all(c(0,1) %in% resp)){ resp.pat <- 10           #Mixed
 } else if (1 %in% resp)   { resp.pat <- 1   #All-1
   } else                  { resp.pat <- 0   #All-0
     }
 return(resp.pat)
}#resp.chk

#Update ability estimate and select the next item
update.theta <- function(resp, abc, theta.old){
 if (resp.chk(resp) == 10){
  x  <- unlist(resp);  nj <- length(abc[,1])
  pa <- list("3pl" = list(a = abc[,1], b = abc[,2], c = abc[,3]),
             "gpcm" = NULL)
  theta.tem <- irt.ability(x, pa, ind.dichot = 1:nj, method = "MLE",
       std.err = TRUE, control = list(D = 1.7, start_val = theta.old))
  sem.tem   <- attributes(theta.tem)$"std.err"
  if(abs(theta.tem) <= 3.5) {
     theta.new <- theta.tem
     sem.obs   <- sem.tem
  } else if (theta.tem < 0) {
     theta.new <- -3.5
     sem.obs   <- 3
    } else if (theta.tem > 0) {
       theta.new <- 3.5
       sem.obs   <- 3
      }
 } else if (resp.chk(resp) == 0){
    theta.new <- max(theta.old - .5, -3.5)
    sem.obs   <- 3
   } else if (resp.chk(resp) == 1){
      theta.new <- min(theta.old + .5, 3.5)
      sem.obs   <- 3
     }
 results <- cbind(theta.new, sem.obs)
 return(results)
}#update.theta

#Compute observed SEM for MLE, using the 2nd derivative of likelihood
get.sem.obs <- function(resp, abc, theta){
 D <- 1.7; aa <- abc[,1]; bb <- abc[,2]; cc <- abc[,3]
 prob.1  <- cc +(1-cc)/(1+exp(-D*aa*(theta-bb))); prob.0  <- 1-prob.1
 d2like  <- sum(D^2*aa^2*((prob.1-cc)/(1-cc)^2)*((prob.0/prob.1))*
              ((resp*cc-prob.1^2)/(prob.1)))
 sem.obs <- 1/sqrt(-1*d2like)
 return(sem.obs)}
```

E.2 Subfunction for implementing CAT as designed in Section 2.1

```
cat.r <- function(j,bank,theta.tru,mag.n,mag.u,i.con,i.occ,i.grp){
 theta.0 <- runif(1, -.5, .5); theta.est <- c(); j.aval <- 1:500
 j.used  <- c(); prob.1 <- c();   rand.u <- c(); resp   <- c()
 iif.exp <- c();sem.exp <- c();  sem.obs <- c()
 conds   <- c(mag.n, mag.u, i.con, i.occ)

 iif.temp   <- f.iif.exp(theta.0, bank)
 iif.exp[1] <- max(iif.temp[,"iif"])
 j.used[1]  <- iif.temp[iif.temp[,"iif"]==iif.exp[1], "j.id"]

 for (stage in 1:30){
  if (stage %in% j.dif.id){
   a.ini   <- bank[bank[,"j.id"]==j.used[stage], "a"]
   b.ini   <- bank[bank[,"j.id"]==j.used[stage], "b"]
   c.ini   <- bank[bank[,"j.id"]==j.used[stage], "c"]
   abc.use <- cbind(a.ini - ((-1)^i.grp)*mag.n,
                 b.ini + ((-1)^i.grp)*mag.u, c.ini)
  } else abc.use <- bank[bank[,"j.id"]==j.used[stage], c("a","b","c")]

  prob.1[stage] <- irf.3pl(theta.true[i], abc.use)
  rand.u[stage] <- runif(1)
  ifelse(prob.1[stage]>rand.u[stage], resp[stage]<-1, resp[stage]<-0)

  ifelse(stage==1, theta.old<-theta.0, theta.old<-theta.est[stage-1])
  update.temp      <- update.theta(resp, bank[j.used, 2:4], theta.old)
  theta.est[stage] <- update.temp[,1]
  sem.obs[stage]   <- update.temp[,2]
  sem.exp[stage]   <- 1/sqrt(sum(iif.exp[1:stage]))

  if (stage < 30){
   j.aval   <- setdiff(j.aval, j.used)
   iif.temp <- f.iif.exp(theta.est[stage],
                      subset(bank, bank[,"j.id"] %in% j.aval))
   iif.exp[stage+1] <- max(iif.temp[, "iif"])
   j.used[stage+1]<-iif.temp[iif.temp[,"iif"]==iif.exp[stage+1],"j.id"]
   }
 }#stage

 cat.out <- c(j, conds, i.grp, theta.true[i], theta.0,
             round(c(theta.est, iif.exp, sem.exp, sem.obs), digits=5),
             j.used, resp)
 return(cat.out)
}
```

E.3 Main function for generating CAT data when operational items exhibit DIF

```
dir.main   <- ""                        #Specify working directory
setwd(dir.main)
library(MiscPsycho)

require(doSMP)                          #Set up parallel computation
workers <- startWorkers(8)
registerDoSMP(workers)

source("0-subfunctions.r")  #Call subfunctions
source("0-cat-routine.r")               #Call CAT routine

theta.true <- seq(-3.5, 3.5, .5)
contam <- c(6, 15, 24)                  #Number of DIF items in operational CAT
labcon <- c(); labcon[c(6, 15, 24)] <- c(2, 5, 8)
occurr <- c(1, 2, 3, 4)          #1.first, 2.middle, 3.last, 4.uniform
mag.dif <- cbind(c(0, 0, 0,.2,.2,.2,.2,.5,.5,.5,.5,.8,.8,.8,.8), #mag.n
              c(.2,.5,.8, 0,.2,.5,.8, 0,.2,.5,.8, 0,.2,.5,.8)) #mag.u
bank <- as.matrix(read.table("500items.abc", header = TRUE)) #Item bank
colnames(bank) <- c("j.id", "a", "b", "c")

for (i.con in contam){
 for (i.occ in occurr){
  j.dif.id  <- get.dif.id(i.con, i.occ)

  for (i.mag in 1:15){ #15 pairs of mag.n & mag.u
   mag.n  <- mag.dif[i.mag, 1]
   mag.u  <- mag.dif[i.mag, 2]
   conds  <- c(mag.n,    mag.u,    i.con,   i.occ)
   theta.out <- c()

   for (i in 1:15){ #15 points of true theta
    for (i.grp in 1:2){#1 = reference, 2 = focal
     cat.out <- foreach(j=1:200, .packages = "MiscPsycho") %dopar%
              cat.r(j,bank,theta.tru[i],mag.n,mag.u,i.con,i.occ,i.grp)
     theta.out <- rbind(theta.out, do.call(rbind,cat.out))
    }#i.grp
   }#i.person

   fname <- paste(dir.main,"raw-with-na/", mag.n*10, mag.u*10,
                 labcon[i.con], i.occ, ".csv", sep = "")
   colnames(theta.out) <- c("rep", "n", "u", "con", "occ",
                            "grp", "t.tru", "t.est0",
                           paste("t.est", 1:30, sep = ""),
                           paste("iif", 1:30, sep = ""),
                           paste("se.e", 1:30, sep = ""),
                           paste("se.o", 1:30, sep = ""),
                           paste("j", 1:30, sep = ""),
                           paste("res", 1:30, sep = ""))
   write.csv(theta.out, file = fname, quote = FALSE, row.names = FALSE)
  }#i.mag
 }#i.occ
}#i.con
stopWorkers(workers)
```

E.4 Subfunction for resampling data for Study 2 (An example code for the condition of no test impact and balanced sample only)

```
fn.05 <- function(j, dif.r, dif.f, abc500, mag.n, mag.u, con, occ){
#j = #of replication; dif.r and dif.f = data generated in Study 1
#abc500 = item bank used in Study 1
#mag.n, mag.u, con, occ = conditions as manipulated in Study 1
 source("")  #Locate a file containing parameters of the pretest items

 #Assign prob. weights to examinees in each group, and resample data
 ru  <- runif(3750) #15000/4 = 3750
 wgh <- c(sort(dnorm(ru,0,.3)), sort(dnorm(ru,0,.3), decreasing = T))
 resam.r <- dif.r[sample(1:7500, 1000, prob = wgh),]
 resam.f <- dif.f[sample(1:7500, 1000, prob = wgh),]
 resam   <- rbind(resam.r, resam.f)

 #Randomly draw one set of 30 items from the bank for all examinees
 j30   <- 1:30; abc30 <- cbind(j30, abc500[sample(1:500, 30),])
 abc30.r <- abc30; colnames(abc30.r) <- c("j30", "j500", "a", "b", "c")
 abc30.f <- abc30; colnames(abc30.f) <- c("j30", "j500", "a", "b", "c")

 #Add DIF in a/b for F-group, depending on DIF Occ. & DIF Con.
 abc30.f[j30 %in% dif.id, "a"] <- abc30.f[j30 %in% dif.id, "a"] - mag.n
 abc30.f[j30 %in% dif.id, "b"] <- abc30.f[j30 %in% dif.id, "b"] + mag.u

 #Generate 0/1 on the 30-item test for the resampled examinees
 nat01.r <- irtgen(abc = abc30.r[, 3:5], theta = resam.r$t.tru)
 nat01.f <- irtgen(abc = abc30.f[, 3:5], theta = resam.f$t.tru)

 #Compute Number-Correct Score and CBT-based ability estimates
 all.r <- cbind(nat01.r, get.t.cbt(nat01.r, abc30.r), rowSums(nat01.r))
 all.f <- cbind(nat01.f, get.t.cbt(nat01.f, abc30.f), rowSums(nat01.f))
 all.nat <- rbind(all.r, all.f)
 colnames(all.nat) <- c(paste("nat", 1:30, sep=""), "t.cbt", "ncs")
   #Generate data for DIF-free pretest items
   pre.no.r <- irtgen(abc = abc.no, theta = resam.r$t.tru)
   pre.no.f <- irtgen(abc = abc.no, theta = resam.f$t.tru)
   pre.no   <- rbind(pre.no.r, pre.no.f)
   colnames(pre.no) <- seq(0.01, 0.16, .01)
   #Generate data for Uniform DIF pretest items
   pre.u.r <- irtgen(abc = abc.u.r, theta = resam.r$t.tru)
   pre.u.f <- irtgen(abc = abc.u.f, theta = resam.f$t.tru)
   pre.u   <- rbind(pre.u.r, pre.u.f)
   colnames(pre.u) <- seq(1.01, 1.16, .01)
   #Generate data for nonuniform DIF pretest items
   pre.n.r <- irtgen(abc = abc.n.r, theta = resam.r$t.tru)
   pre.n.f <- irtgen(abc = abc.n.f, theta = resam.f$t.tru)
   pre.n   <- rbind(pre.n.r, pre.n.f)
   colnames(pre.n) <- seq(2.01, 2.16, .01)

 #Combine all generated data
 out.all <- cbind(rep(j, 2000), resam, all.nat, pre.no, pre.u, pre.n)
 return(out.all)}
```

Appendix F: BIAS and RMSE as presented in Study 1 (only θ = ±3.5, ±2.5, ±1.5, and ±.5)

F.1 Magnitude of uniform DIF was .4 (Figures 4 and 19).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.104	.105	.139	.079	.141	.093	.123	.105	.183	.191	.221	.157	.230	.182	.206	.192
	15	.147	.074	.181	.064	.173	.063	.160	.061	.230	.163	.264	.137	.261	.134	.247	.135
	24	.194	.051	.220	.043	.228	.047	.211	.051	.285	.125	.305	.111	.308	.119	.303	.120
-2.5	6	.040	-.030	.044	-.034	.062	-.032	.030	-.008	.262	.250	.281	.256	.288	.262	.265	.263
	15	.086	-.086	.121	-.114	.113	-.096	.114	-.098	.287	.282	.283	.276	.274	.291	.306	.284
	24	.158	-.132	.184	-.128	.173	-.155	.180	-.150	.297	.296	.320	.293	.308	.318	.332	.305
-1.5	6	.042	-.034	.029	-.030	.039	-.019	.033	-.041	.260	.263	.253	.266	.255	.256	.269	.266
	15	.102	-.101	.133	-.114	.109	-.104	.090	-.112	.273	.262	.287	.271	.274	.267	.270	.291
	24	.172	-.174	.176	-.173	.151	-.159	.163	-.140	.301	.307	.318	.317	.324	.303	.305	.294
-.5	6	.048	-.028	.065	-.023	.023	-.027	.048	-.035	.258	.249	.248	.242	.237	.253	.249	.247
	15	.132	-.117	.122	-.098	.097	-.084	.124	-.090	.276	.272	.285	.265	.271	.255	.277	.283
	24	.161	-.156	.175	-.160	.163	-.155	.162	-.145	.300	.299	.304	.290	.297	.297	.299	.289
.5	6	.070	-.055	.044	-.038	.029	-.035	.052	-.023	.250	.243	.243	.245	.251	.257	.244	.238
	15	.116	-.092	.111	-.118	.101	-.083	.095	-.085	.278	.246	.271	.262	.272	.250	.269	.258
	24	.152	-.159	.151	-.169	.156	-.144	.165	-.166	.298	.287	.292	.297	.293	.292	.297	.294
1.5	6	.045	-.030	.070	-.018	.057	-.031	.068	-.036	.263	.249	.261	.246	.251	.245	.269	.259
	15	.139	-.111	.131	-.114	.112	-.094	.098	-.085	.296	.272	.286	.268	.260	.281	.283	.258
	24	.154	-.154	.177	-.150	.177	-.144	.176	-.159	.301	.303	.312	.296	.314	.281	.304	.292
2.5	6	.024	-.026	.061	-.029	.051	-.053	.036	-.038	.263	.281	.281	.267	.270	.242	.267	.265
	15	.084	-.083	.108	-.104	.102	-.105	.116	-.103	.297	.270	.299	.273	.277	.287	.306	.293
	24	.152	-.150	.149	-.145	.184	-.170	.153	-.165	.314	.290	.294	.292	.328	.312	.308	.318
3.5	6	-.104	-.113	-.093	-.143	-.094	-.132	-.092	-.110	.190	.196	.179	.229	.176	.221	.182	.188
	15	-.072	-.157	-.060	-.188	-.056	-.168	-.063	-.167	.155	.254	.136	.268	.132	.252	.133	.250
	24	-.044	-.212	-.041	-.227	-.034	-.211	-.038	-.209	.110	.295	.106	.314	.093	.289	.094	.289

F.2 Magnitude of uniform DIF was 1.0 (Figures 5 and 20).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
	6	.122	.092	.193	.070	.162	.068	.155	.072	.212	.172	.289	.154	.252	.150	.240	.143
-3.5	15	.242	.033	.346	.021	.292	.033	.276	.035	.320	.090	.421	.073	.374	.090	.354	.095
	24	.406	.012	.429	.013	.480	.010	.412	.015	.476	.052	.508	.058	.543	.051	.487	.059
	6	.076	-.046	.095	-.100	.095	-.103	.102	-.059	.278	.263	.285	.277	.269	.292	.292	.287
-2.5	15	.227	-.219	.284	-.304	.247	-.273	.237	-.235	.345	.346	.387	.399	.356	.390	.364	.350
	24	.401	-.395	.433	-.429	.443	-.425	.408	-.384	.482	.468	.507	.504	.503	.502	.485	.467
	6	.109	-.105	.093	-.114	.089	-.102	.090	-.121	.271	.269	.270	.279	.278	.286	.272	.272
-1.5	15	.246	-.239	.267	-.269	.239	-.229	.239	-.226	.355	.347	.382	.372	.351	.364	.355	.355
	24	.422	-.395	.396	-.422	.398	-.429	.399	-.383	.499	.477	.468	.501	.480	.504	.472	.468
	6	.154	-.108	.083	-.097	.102	-.085	.096	-.106	.292	.268	.257	.278	.277	.268	.261	.266
-.5	15	.277	-.239	.261	-.251	.227	-.241	.246	-.248	.388	.355	.361	.365	.326	.349	.349	.346
	24	.440	-.420	.408	-.410	.389	-.391	.404	-.396	.512	.486	.476	.488	.468	.465	.480	.460
	6	.106	-.122	.116	-.091	.111	-.072	.106	-.106	.255	.283	.263	.281	.274	.248	.277	.271
.5	15	.281	-.272	.266	-.239	.231	-.254	.271	-.251	.378	.373	.367	.354	.345	.363	.373	.356
	24	.405	-.424	.392	-.377	.389	-.385	.408	-.391	.477	.496	.470	.456	.468	.461	.480	.469
	6	.106	-.082	.117	-.107	.104	-.087	.071	-.102	.281	.282	.278	.275	.283	.258	.274	.279
1.5	15	.253	-.269	.270	-.269	.254	-.228	.267	-.248	.361	.368	.381	.371	.386	.340	.367	.363
	24	.406	-.416	.445	-.405	.430	-.394	.403	-.408	.492	.487	.522	.471	.501	.469	.485	.483
	6	.063	-.055	.122	-.103	.103	-.101	.096	-.111	.263	.270	.302	.286	.285	.279	.293	.300
2.5	15	.230	-.244	.298	-.256	.271	-.262	.247	-.216	.348	.359	.402	.372	.379	.376	.374	.346
	24	.380	-.383	.437	-.391	.438	-.415	.403	-.398	.463	.478	.507	.476	.518	.490	.485	.481
	6	-.101	-.121	-.061	-.192	-.058	-.181	-.083	-.145	.187	.200	.138	.280	.125	.269	.164	.233
3.5	15	-.030	-.251	-.016	-.337	-.020	-.291	-.031	-.279	.079	.335	.059	.410	.069	.366	.090	.362
	24	-.009	-.402	-.006	-.443	-.004	-.465	-.009	-.416	.042	.477	.039	.513	.028	.535	.044	.487

F.3 Magnitude of uniform DIF was 1.6 (Figures 6 and 21).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
	6	.154	.091	.239	.046	.197	.053	.193	.070	.245	.168	.324	.114	.291	.119	.289	.147
-3.5	15	.363	.021	.504	.007	.456	.010	.381	.018	.441	.076	.569	.037	.526	.063	.457	.071
	24	.636	.003	.694	.001	.739	.001	.648	.003	.686	.029	.741	.019	.782	.012	.701	.026
	6	.113	-.067	.171	-.162	.165	-.148	.153	-.108	.290	.271	.324	.312	.315	.319	.308	.279
-2.5	15	.394	-.358	.457	-.489	.406	-.455	.360	-.357	.490	.438	.520	.565	.481	.524	.448	.450
	24	.662	-.636	.681	-.680	.673	-.707	.643	-.645	.720	.684	.734	.721	.720	.745	.694	.692
	6	.157	-.132	.155	-.167	.141	-.151	.162	-.154	.301	.285	.305	.306	.295	.292	.318	.295
-1.5	15	.420	-.401	.415	-.435	.375	-.399	.402	-.364	.494	.482	.482	.508	.450	.493	.486	.455
	24	.673	-.647	.648	-.669	.646	-.676	.644	-.645	.720	.697	.693	.720	.694	.734	.696	.695
	6	.187	-.166	.137	-.167	.146	-.138	.153	-.154	.321	.314	.283	.313	.285	.291	.296	.282
-.5	15	.429	-.413	.374	-.430	.364	-.375	.431	-.388	.506	.488	.455	.502	.447	.458	.512	.469
	24	.673	-.658	.646	-.672	.612	-.629	.641	-.648	.719	.710	.694	.717	.663	.685	.690	.695
	6	.184	-.201	.134	-.150	.145	-.129	.159	-.155	.308	.317	.287	.291	.288	.284	.285	.292
.5	15	.415	-.437	.403	-.393	.387	-.373	.412	-.415	.482	.501	.485	.470	.464	.454	.484	.491
	24	.688	-.671	.649	-.630	.631	-.594	.650	-.654	.733	.722	.698	.678	.681	.646	.695	.699
	6	.144	-.162	.148	-.153	.141	-.160	.152	-.172	.297	.296	.297	.296	.299	.284	.299	.313
1.5	15	.411	-.414	.425	-.413	.407	-.394	.360	-.392	.492	.481	.512	.482	.498	.473	.444	.468
	24	.648	-.655	.699	-.668	.658	-.620	.659	-.640	.704	.697	.748	.712	.708	.668	.716	.692
	6	.087	-.109	.195	-.171	.154	-.150	.139	-.130	.273	.289	.339	.323	.326	.301	.301	.310
2.5	15	.361	-.401	.467	-.459	.422	-.401	.393	-.389	.447	.472	.553	.534	.511	.485	.482	.479
	24	.638	-.662	.671	-.684	.713	-.670	.649	-.642	.687	.713	.712	.735	.758	.712	.696	.699
	6	-.087	-.140	-.042	-.235	-.053	-.195	-.053	-.203	.168	.237	.107	.320	.120	.281	.116	.288
3.5	15	-.015	-.376	-.005	-.492	-.004	-.447	-.012	-.400	.057	.456	.037	.556	.031	.509	.053	.475
	24	-.002	-.640	-.001	-.676	.000	-.734	-.002	-.638	.017	.694	.014	.734	.005	.774	.016	.699

F.4 Magnitude of nonuniform DIF was .4 (Figures 7 and 22).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.119	.118	.120	.121	.104	.117	.106	.116	.210	.210	.202	.215	.189	.204	.194	.209
	15	.089	.134	.086	.117	.090	.137	.082	.134	.173	.229	.166	.206	.164	.238	.160	.221
	24	.076	.158	.084	.168	.073	.149	.088	.150	.152	.248	.160	.260	.148	.243	.167	.248
-2.5	6	-.002	.019	.007	.001	.004	.013	.032	.001	.266	.277	.261	.273	.253	.280	.261	.275
	15	.005	.034	.010	.005	.035	.010	.015	.017	.245	.285	.237	.301	.250	.279	.242	.279
	24	.005	.018	-.009	.033	.016	-.007	.027	.027	.231	.300	.220	.318	.233	.294	.234	.293
-1.5	6	-.001	.021	.006	-.011	.009	.000	-.012	.000	.252	.254	.243	.274	.256	.269	.263	.254
	15	-.019	-.009	.006	.014	-.005	-.022	-.005	-.021	.247	.282	.234	.294	.235	.275	.232	.282
	24	-.010	-.013	.000	-.010	-.010	-.018	.007	-.003	.222	.312	.213	.294	.224	.299	.236	.306
-.5	6	-.003	.007	.000	-.020	-.016	.012	.027	-.024	.242	.270	.240	.246	.247	.260	.250	.256
	15	.014	.002	.028	-.004	.009	-.009	.002	-.006	.226	.274	.226	.259	.236	.254	.220	.256
	24	.022	-.016	.007	-.002	.003	-.001	.034	.000	.213	.292	.213	.289	.218	.284	.231	.288
.5	6	.002	-.009	.012	.008	.002	.005	.013	.001	.241	.247	.232	.252	.227	.244	.234	.251
	15	.003	.011	-.002	.001	.009	-.012	.015	.012	.234	.272	.223	.274	.233	.268	.231	.267
	24	.000	-.020	.005	.008	.005	.003	.006	.012	.211	.293	.219	.290	.224	.275	.214	.282
1.5	6	.014	.006	.020	.002	.015	-.006	.016	.015	.241	.246	.245	.270	.240	.266	.247	.258
	15	.024	-.026	.012	.003	.008	-.007	.009	-.027	.227	.284	.232	.278	.225	.285	.227	.274
	24	.035	-.019	.011	.000	.023	-.003	.013	-.009	.223	.273	.224	.300	.218	.315	.227	.297
2.5	6	.017	-.016	.002	.008	.006	.000	.022	.001	.265	.271	.248	.277	.247	.271	.252	.265
	15	.018	-.002	.017	.004	.009	-.010	.013	-.017	.236	.305	.242	.307	.228	.296	.232	.303
	24	.029	-.006	.019	.002	.013	-.001	.026	-.015	.227	.324	.224	.307	.226	.319	.218	.314
3.5	6	-.103	-.114	-.086	-.114	-.088	-.107	-.090	-.102	.184	.196	.160	.195	.167	.196	.167	.183
	15	-.084	-.135	-.090	-.122	-.088	-.122	-.094	-.125	.159	.229	.168	.212	.166	.211	.169	.216
	24	-.083	-.149	-.072	-.145	-.079	-.126	-.075	-.137	.160	.244	.142	.231	.155	.218	.140	.227

F.5 Magnitude of nonuniform DIF was 1.0 (Figures 8 and 23).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.109	.147	.093	.153	.087	.126	.099	.141	.197	.241	.175	.243	.170	.218	.191	.234
	15	.071	.196	.064	.199	.075	.173	.079	.165	.144	.308	.136	.303	.147	.279	.156	.272
	24	.058	.266	.054	.298	.055	.255	.055	.250	.130	.394	.121	.416	.121	.364	.122	.381
-2.5	6	.002	.057	.012	-.012	.011	-.023	-.005	.011	.262	.277	.239	.299	.235	.289	.247	.293
	15	.005	.088	.028	-.020	.016	-.028	.033	.016	.225	.370	.222	.375	.213	.362	.229	.352
	24	.027	.112	.037	.029	.045	-.003	.019	.043	.204	.441	.197	.416	.199	.423	.199	.402
-1.5	6	.006	.036	.011	-.018	.017	-.014	-.014	.010	.258	.288	.243	.300	.232	.291	.233	.298
	15	-.019	.040	.005	-.026	-.003	-.057	.003	.032	.221	.353	.201	.355	.221	.360	.204	.347
	24	.014	.002	.010	-.025	.022	-.072	.007	.037	.206	.429	.187	.414	.193	.418	.189	.413
-.5	6	-.008	.024	.025	.000	.005	-.002	.020	-.003	.226	.284	.243	.279	.238	.282	.233	.281
	15	.009	-.002	.035	-.011	.009	-.011	.018	.017	.212	.348	.206	.367	.207	.335	.212	.333
	24	.045	-.007	.040	-.052	.041	-.034	.036	.002	.198	.415	.185	.423	.188	.389	.191	.409
.5	6	.028	.002	.000	-.002	-.009	-.018	.015	-.015	.222	.278	.224	.279	.234	.274	.229	.275
	15	-.002	-.007	.000	.003	.023	.001	.010	-.029	.205	.345	.194	.333	.204	.312	.204	.312
	24	.021	-.045	.007	-.020	-.005	-.027	.018	-.039	.191	.419	.179	.389	.182	.401	.188	.409
1.5	6	.036	-.036	.014	-.011	.022	.003	.035	-.026	.236	.290	.225	.288	.225	.289	.237	.279
	15	.036	-.082	.017	-.005	.024	.000	.037	-.061	.214	.331	.207	.342	.191	.341	.223	.333
	24	.035	-.113	.047	-.045	.032	-.024	.041	-.077	.188	.454	.191	.418	.172	.419	.196	.410
2.5	6	.018	-.060	.030	.035	.002	-.013	.026	-.017	.251	.274	.234	.312	.245	.308	.252	.304
	15	.023	-.132	.002	.011	.008	-.042	.008	-.058	.235	.394	.209	.383	.219	.387	.226	.356
	24	.039	-.159	.031	-.116	.010	.029	.026	-.108	.207	.453	.195	.463	.174	.463	.181	.448
3.5	6	-.102	-.161	-.082	-.125	-.094	-.129	-.085	-.131	.183	.260	.159	.221	.173	.217	.162	.229
	15	-.070	-.215	-.061	-.171	-.083	-.163	-.074	-.168	.144	.328	.133	.283	.161	.263	.144	.280
	24	-.060	-.298	-.053	-.305	-.055	-.213	-.059	-.257	.123	.442	.118	.461	.112	.336	.125	.395

F.6 Magnitude of nonuniform DIF was 1.6 (Figures 9 and 24).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.094	.179	.077	.168	.090	.149	.082	.154	.179	.296	.156	.274	.169	.241	.159	.252
	15	.059	.396	.037	.289	.052	.237	.056	.273	.127	.605	.091	.428	.110	.353	.119	.412
	24	.043	.694	.035	.608	.032	.417	.037	.535	.106	.942	.095	.807	.086	.582	.092	.739
-2.5	6	-.007	.147	-.003	.004	.011	.000	.025	.041	.262	.364	.223	.331	.240	.306	.238	.319
	15	-.007	.279	.057	-.032	.061	-.048	.044	.080	.227	.574	.203	.490	.214	.422	.210	.483
	24	.030	.425	.033	.241	.051	-.054	.037	.334	.187	.836	.181	.744	.162	.595	.187	.748
-1.5	6	-.019	.099	.002	-.025	.003	.020	.004	.040	.265	.322	.236	.333	.242	.306	.227	.312
	15	-.001	.217	.015	-.045	-.002	-.036	.002	.080	.203	.596	.192	.451	.193	.430	.199	.455
	24	-.002	.215	.008	-.023	.012	-.114	.006	.101	.181	.801	.182	.715	.177	.622	.175	.709
-.5	6	.004	.035	.039	-.014	.011	-.032	.026	.000	.242	.346	.223	.312	.216	.302	.220	.279
	15	.027	.067	.047	-.097	.035	-.041	.021	.042	.201	.537	.204	.452	.197	.410	.206	.393
	24	.031	.002	.035	-.059	.040	-.164	.021	-.003	.178	.693	.169	.713	.173	.622	.175	.607
.5	6	-.006	-.053	.019	-.009	.017	.005	.023	-.012	.229	.331	.213	.301	.212	.292	.232	.285
	15	.011	-.173	.010	-.010	.013	-.001	.012	-.058	.197	.570	.181	.432	.181	.362	.187	.418
	24	.010	-.154	-.009	-.112	.012	-.047	.015	-.149	.162	.746	.160	.695	.150	.552	.154	.664
1.5	6	.035	-.137	.017	-.014	.034	-.005	.036	-.049	.244	.356	.224	.321	.228	.309	.228	.297
	15	.037	-.325	.025	.000	.035	-.042	.031	-.168	.212	.652	.188	.515	.182	.416	.192	.480
	24	.040	-.405	.046	-.149	.051	-.017	.058	-.300	.163	.846	.161	.649	.156	.650	.171	.727
2.5	6	.052	-.195	.002	.003	.032	-.011	.046	-.082	.255	.405	.227	.362	.236	.313	.243	.326
	15	.021	-.434	.013	-.004	.027	-.006	.018	-.193	.218	.696	.185	.543	.184	.480	.193	.537
	24	.029	-.614	.035	-.434	.022	.006	.033	-.448	.169	.953	.173	.804	.150	.665	.178	.802
3.5	6	-.101	-.255	-.086	-.152	-.082	-.153	-.094	-.172	.186	.382	.165	.262	.161	.247	.180	.287
	15	-.064	-.638	-.060	-.291	-.064	-.218	-.057	-.390	.135	.849	.120	.463	.128	.343	.120	.600
	24	-.044	-.922	-.036	-.833	-.037	-.413	-.041	-.781	.104	1.161	.087	1.038	.097	.624	.098	1.015

F.7 Magnitudes of nonuniform and uniform DIF were .4 and .4 respectively (Figures 10 and 25).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.105	.112	.129	.097	.113	.113	.130	.104	.185	.196	.208	.179	.196	.200	.212	.184
	15	.143	.112	.172	.095	.155	.082	.150	.103	.231	.203	.267	.182	.235	.166	.236	.191
	24	.179	.078	.190	.087	.207	.083	.165	.094	.261	.156	.268	.175	.283	.168	.245	.185
-2.5	6	.039	.000	.049	-.029	.034	-.023	.033	-.022	.261	.277	.243	.267	.245	.275	.269	.258
	15	.102	-.082	.137	-.116	.137	-.109	.105	-.073	.256	.281	.282	.316	.273	.314	.268	.290
	24	.153	-.152	.177	-.155	.190	-.181	.175	-.129	.280	.334	.285	.336	.288	.355	.291	.330
-1.5	6	.030	-.033	.039	-.041	.048	-.044	.055	-.032	.251	.270	.244	.261	.252	.268	.257	.259
	15	.099	-.108	.124	-.103	.128	-.120	.125	-.074	.265	.302	.261	.299	.272	.330	.272	.296
	24	.155	-.168	.170	-.162	.190	-.176	.185	-.147	.271	.337	.287	.340	.297	.353	.296	.344
-.5	6	.071	-.027	.048	-.038	.037	-.035	.045	-.018	.248	.268	.236	.277	.240	.271	.242	.260
	15	.110	-.093	.108	-.082	.106	-.108	.128	-.105	.258	.286	.253	.301	.253	.303	.262	.305
	24	.168	-.159	.197	-.160	.187	-.164	.185	-.174	.270	.313	.289	.339	.286	.357	.278	.329
.5	6	.055	-.036	.057	-.042	.070	-.020	.055	-.028	.251	.262	.250	.255	.250	.259	.247	.242
	15	.098	-.119	.110	-.090	.106	-.082	.112	-.094	.252	.285	.253	.278	.256	.287	.255	.294
	24	.154	-.178	.165	-.134	.168	-.144	.178	-.159	.270	.332	.277	.308	.272	.312	.281	.330
1.5	6	.060	-.057	.059	-.041	.052	-.018	.059	-.050	.250	.260	.243	.266	.235	.252	.238	.264
	15	.108	-.110	.128	-.093	.121	-.095	.110	-.099	.271	.298	.273	.291	.271	.297	.249	.294
	24	.186	-.172	.195	-.179	.186	-.148	.179	-.161	.298	.345	.295	.345	.294	.329	.277	.334
2.5	6	.035	-.039	.041	-.026	.053	-.022	.043	-.029	.266	.268	.259	.289	.265	.277	.249	.274
	15	.097	-.116	.116	-.094	.111	-.086	.102	-.080	.270	.305	.265	.307	.268	.297	.261	.302
	24	.171	-.178	.197	-.207	.203	-.165	.178	-.177	.288	.362	.312	.378	.304	.366	.290	.347
3.5	6	-.087	-.116	-.074	-.149	-.075	-.128	-.076	-.128	.163	.196	.154	.242	.153	.210	.155	.217
	15	-.063	-.185	-.034	-.206	-.045	-.184	-.057	-.170	.134	.281	.098	.300	.105	.281	.129	.258
	24	-.032	-.234	-.023	-.256	-.021	-.236	-.026	-.253	.087	.334	.069	.356	.067	.332	.075	.352

F.8 Magnitudes of nonuniform and uniform DIF were 1.0 and 1.0 respectively (Figures 11 and 26).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.108	.120	.184	.094	.181	.106	.160	.122	.192	.207	.273	.176	.267	.202	.241	.216
	15	.240	.114	.331	.107	.314	.099	.273	.105	.314	.217	.400	.198	.380	.191	.350	.199
	24	.412	.109	.424	.109	.498	.094	.418	.099	.470	.213	.482	.226	.538	.191	.470	.197
-2.5	6	.054	.020	.157	-.072	.157	-.070	.095	-.059	.271	.283	.286	.312	.285	.305	.262	.293
	15	.244	-.107	.342	-.209	.329	-.208	.280	-.131	.351	.349	.411	.411	.398	.388	.360	.358
	24	.424	-.272	.463	-.303	.494	-.374	.455	-.303	.473	.451	.509	.499	.526	.538	.499	.482
-1.5	6	.113	-.030	.129	-.078	.126	-.070	.123	-.047	.269	.277	.275	.295	.281	.288	.282	.275
	15	.254	-.199	.298	-.233	.287	-.206	.286	-.152	.339	.397	.367	.439	.369	.413	.369	.385
	24	.418	-.367	.454	-.369	.463	-.400	.435	-.324	.462	.556	.491	.580	.501	.586	.480	.526
-.5	6	.137	-.080	.142	-.064	.111	-.058	.140	-.075	.293	.285	.287	.288	.272	.291	.293	.284
	15	.296	-.192	.281	-.209	.298	-.185	.295	-.184	.375	.401	.356	.409	.359	.370	.360	.378
	24	.447	-.397	.438	-.386	.436	-.368	.420	-.386	.485	.575	.476	.577	.471	.553	.460	.557
.5	6	.132	-.119	.148	-.046	.093	-.048	.134	-.071	.276	.304	.271	.291	.247	.257	.276	.299
	15	.266	-.303	.295	-.191	.299	-.151	.317	-.229	.340	.473	.371	.395	.370	.366	.387	.411
	24	.441	-.460	.437	-.383	.443	-.266	.441	-.390	.486	.631	.484	.549	.489	.461	.486	.561
1.5	6	.142	-.128	.136	-.074	.148	-.060	.131	-.090	.286	.292	.269	.295	.292	.293	.280	.305
	15	.287	-.322	.312	-.214	.289	-.180	.285	-.271	.371	.494	.382	.412	.356	.375	.361	.442
	24	.439	-.517	.448	-.378	.487	-.316	.454	-.394	.484	.665	.493	.564	.529	.515	.500	.552
2.5	6	.070	-.139	.150	-.048	.110	-.062	.124	-.090	.268	.338	.299	.307	.269	.297	.279	.311
	15	.230	-.326	.294	-.205	.304	-.216	.251	-.237	.336	.482	.367	.435	.378	.427	.341	.446
	24	.414	-.553	.448	-.500	.464	-.341	.416	-.472	.467	.716	.491	.668	.501	.563	.468	.639
3.5	6	-.087	-.189	-.043	-.139	-.043	-.158	-.049	-.157	.168	.295	.102	.228	.099	.258	.115	.254
	15	-.012	-.382	-.002	-.327	-.006	-.279	-.012	-.320	.050	.504	.012	.437	.037	.381	.050	.437
	24	-.001	-.595	-.001	-.628	-.001	-.471	-.001	-.541	.008	.727	.008	.754	.009	.603	.012	.666

F.9 Magnitudes of nonuniform and uniform DIF were 1.6 and 1.6 respectively (Figures 12 and 27).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.127	.147	.275	.138	.228	.135	.195	.137	.214	.250	.347	.233	.307	.226	.276	.233
	15	.380	.273	.555	.214	.507	.180	.488	.197	.438	.428	.593	.354	.550	.289	.533	.334
	24	.660	.402	.726	.346	.822	.283	.733	.327	.685	.621	.747	.544	.841	.430	.755	.513
-2.5	6	.114	.088	.225	-.057	.222	-.036	.190	-.004	.316	.317	.330	.339	.333	.300	.325	.294
	15	.397	.087	.488	-.186	.534	-.170	.450	-.100	.462	.471	.536	.532	.568	.441	.502	.399
	24	.660	-.028	.717	-.143	.746	-.328	.714	-.045	.689	.613	.739	.600	.762	.626	.736	.631
-1.5	6	.156	.029	.216	-.064	.218	-.048	.181	-.053	.309	.307	.324	.336	.321	.322	.317	.303
	15	.421	-.096	.532	-.285	.487	-.178	.467	-.099	.481	.544	.573	.588	.528	.495	.521	.420
	24	.687	-.258	.741	-.418	.759	-.513	.714	-.266	.714	.749	.759	.845	.779	.872	.737	.756
-.5	6	.215	-.073	.210	-.068	.196	-.044	.203	-.032	.332	.329	.323	.330	.321	.307	.320	.319
	15	.453	-.226	.476	-.261	.487	-.185	.494	-.158	.496	.575	.519	.577	.521	.460	.533	.456
	24	.673	-.455	.686	-.548	.706	-.401	.718	-.401	.694	.879	.708	.929	.724	.768	.739	.755
.5	6	.194	-.159	.208	-.041	.195	-.051	.202	-.092	.309	.377	.317	.303	.303	.290	.323	.306
	15	.461	-.410	.512	-.194	.517	-.144	.512	-.266	.505	.681	.550	.489	.558	.408	.557	.502
	24	.729	-.711	.720	-.424	.766	-.309	.750	-.550	.752	1.028	.742	.835	.785	.648	.770	.860
1.5	6	.191	-.240	.232	-.063	.209	-.045	.197	-.111	.314	.419	.345	.326	.331	.312	.328	.310
	15	.425	-.580	.515	-.180	.516	-.196	.470	-.365	.481	.805	.557	.500	.562	.468	.522	.606
	24	.688	-.890	.748	-.526	.775	-.318	.753	-.708	.716	1.149	.770	.817	.796	.660	.774	.979
2.5	6	.111	-.292	.245	-.034	.201	-.046	.171	-.148	.279	.478	.365	.368	.333	.322	.302	.357
	15	.411	-.782	.530	-.216	.521	-.173	.459	-.514	.485	1.017	.585	.592	.566	.512	.516	.774
	24	.670	-1.137	.735	-.833	.776	-.280	.722	-.912	.702	1.391	.760	1.073	.795	.771	.746	1.176
3.5	6	-.090	-.335	-.018	-.182	-.030	-.162	-.045	-.229	.174	.515	.063	.288	.081	.253	.112	.360
	15	-.004	-.997	.000	-.401	.000	-.305	.000	-.606	.029	1.212	.000	.567	.006	.434	.006	.839
	24	.000	-1.434	.000	-1.186	.000	-.609	.000	-1.075	.000	1.649	.000	1.384	.000	.811	.000	1.298

F.10 Magnitudes of nonuniform and uniform DIF were 1.0 and .4 respectively (Figures 13 and 28).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.093	.127	.138	.124	.113	.114	.119	.110	.177	.224	.228	.208	.198	.200	.207	.194
	15	.123	.153	.155	.155	.152	.139	.131	.151	.203	.256	.236	.260	.231	.241	.216	.270
	24	.143	.188	.157	.207	.176	.166	.152	.195	.225	.308	.240	.327	.252	.275	.225	.305
-2.5	6	.019	.031	.070	-.010	.094	-.049	.048	-.022	.248	.280	.237	.294	.263	.296	.247	.304
	15	.113	.021	.162	-.101	.156	-.107	.114	-.051	.265	.334	.271	.364	.266	.361	.249	.361
	24	.181	-.031	.205	-.129	.234	-.207	.205	-.109	.272	.441	.282	.449	.301	.445	.287	.419
-1.5	6	.022	.003	.053	-.062	.048	-.007	.009	.007	.247	.284	.236	.318	.236	.275	.246	.279
	15	.108	-.055	.137	-.119	.109	-.112	.106	-.047	.242	.339	.253	.400	.227	.382	.235	.337
	24	.179	-.089	.174	-.155	.180	-.191	.188	-.155	.273	.451	.264	.458	.266	.437	.277	.458
-.5	6	.062	-.001	.080	-.039	.059	-.035	.071	-.015	.253	.270	.259	.285	.242	.266	.251	.288
	15	.112	-.083	.127	-.126	.116	-.102	.134	-.070	.249	.347	.260	.357	.238	.349	.246	.331
	24	.179	-.176	.183	-.164	.201	-.190	.190	-.150	.260	.447	.255	.451	.267	.440	.264	.438
.5	6	.052	-.068	.046	-.037	.050	-.029	.079	-.004	.236	.293	.224	.258	.244	.290	.246	.280
	15	.104	-.125	.106	-.086	.117	-.070	.122	-.112	.241	.362	.242	.332	.229	.336	.238	.331
	24	.181	-.208	.162	-.146	.148	-.164	.162	-.176	.263	.460	.248	.431	.235	.401	.257	.442
1.5	6	.080	-.068	.063	-.016	.077	-.021	.071	-.052	.252	.306	.251	.292	.250	.297	.244	.277
	15	.134	-.156	.128	-.093	.122	-.109	.113	-.127	.248	.384	.248	.379	.248	.353	.245	.339
	24	.195	-.256	.200	-.215	.194	-.151	.206	-.216	.283	.479	.278	.455	.266	.440	.286	.447
2.5	6	.064	-.103	.037	-.024	.058	-.036	.075	-.024	.251	.290	.252	.295	.243	.286	.244	.302
	15	.118	-.243	.142	-.038	.140	-.104	.116	-.122	.252	.455	.259	.381	.252	.371	.247	.381
	24	.164	-.298	.183	-.239	.182	-.168	.192	-.246	.257	.524	.268	.525	.259	.485	.278	.496
3.5	6	-.098	-.147	-.059	-.146	-.054	-.141	-.070	-.137	.180	.244	.132	.242	.119	.232	.144	.230
	15	-.037	-.287	-.022	-.263	-.024	-.193	-.038	-.242	.091	.417	.068	.382	.071	.303	.098	.372
	24	-.011	-.411	-.012	-.419	-.009	-.345	-.016	-.347	.047	.561	.048	.565	.036	.492	.052	.492

F.11 Magnitudes of nonuniform and uniform DIF were 1.6 and .4 respectively (Figures 14 and 29).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.100	.176	.121	.158	.121	.142	.118	.162	.180	.287	.202	.250	.204	.238	.194	.260
	15	.110	.360	.132	.264	.122	.222	.129	.263	.190	.542	.207	.374	.196	.346	.209	.427
	24	.123	.594	.128	.578	.144	.364	.120	.483	.195	.857	.203	.812	.210	.536	.184	.704
-2.5	6	.014	.122	.067	-.027	.079	-.031	.047	.037	.258	.335	.235	.337	.255	.355	.254	.339
	15	.100	.195	.159	-.053	.175	-.097	.130	.030	.241	.560	.259	.456	.258	.420	.252	.476
	24	.171	.282	.207	.163	.239	-.125	.198	.148	.250	.774	.263	.668	.291	.599	.271	.610
-1.5	6	.032	.069	.057	-.032	.068	-.007	.048	.002	.234	.336	.248	.321	.236	.311	.244	.303
	15	.125	.111	.127	-.140	.125	-.110	.124	.054	.254	.528	.229	.508	.234	.475	.247	.483
	24	.177	.126	.198	-.055	.189	-.202	.195	.053	.258	.783	.263	.706	.258	.685	.266	.674
-.5	6	.051	.011	.056	-.063	.075	-.026	.069	.000	.242	.317	.235	.305	.238	.297	.242	.296
	15	.111	-.069	.143	-.079	.160	-.111	.125	-.015	.225	.496	.245	.479	.257	.438	.228	.432
	24	.190	-.102	.216	-.228	.199	-.137	.215	-.093	.253	.751	.265	.731	.255	.633	.272	.632
.5	6	.042	-.095	.044	-.017	.053	-.007	.058	-.018	.227	.326	.221	.314	.224	.287	.228	.307
	15	.112	-.205	.102	-.007	.122	-.034	.118	-.121	.236	.605	.226	.443	.225	.404	.226	.414
	24	.175	-.338	.160	-.172	.179	-.149	.190	-.225	.247	.822	.237	.728	.248	.609	.250	.714
1.5	6	.090	-.165	.075	.003	.080	-.029	.075	-.039	.262	.365	.249	.303	.237	.300	.243	.310
	15	.135	-.353	.132	-.058	.124	-.089	.152	-.259	.258	.665	.229	.503	.224	.437	.236	.523
	24	.200	-.505	.203	-.194	.205	-.075	.203	-.410	.268	.903	.259	.745	.259	.670	.258	.832
2.5	6	.068	-.244	.058	.003	.056	.000	.071	-.098	.262	.453	.233	.348	.244	.326	.241	.352
	15	.117	-.522	.119	-.040	.140	-.058	.135	-.257	.244	.764	.241	.512	.234	.489	.243	.564
	24	.178	-.768	.184	-.581	.177	-.121	.190	-.565	.257	1.063	.254	.898	.233	.679	.252	.886
3.5	6	-.089	-.281	-.060	-.155	-.068	-.143	-.068	-.185	.163	.429	.132	.262	.147	.234	.137	.306
	15	-.030	-.728	-.016	-.308	-.019	-.233	-.025	-.439	.088	.949	.056	.483	.063	.362	.077	.628
	24	-.009	-1.028	-.004	-.896	-.002	-.441	-.006	-.784	.043	1.269	.022	1.130	.015	.632	.029	1.046

F.12 Magnitudes of nonuniform and uniform DIF were 1.6 and 1.0 respectively (Figures 15 and 30).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.112	.147	.188	.163	.192	.135	.165	.167	.185	.276	.265	.266	.273	.228	.245	.268
	15	.236	.324	.347	.234	.318	.209	.300	.217	.315	.503	.405	.392	.378	.326	.367	.366
	24	.387	.518	.431	.446	.488	.328	.405	.422	.435	.761	.480	.662	.526	.481	.453	.636
-2.5	6	.057	.088	.149	-.048	.145	-.059	.101	.001	.263	.305	.280	.330	.280	.322	.272	.301
	15	.228	.145	.344	-.168	.366	-.151	.295	-.020	.321	.511	.404	.489	.415	.419	.360	.472
	24	.425	.142	.473	.049	.511	-.242	.468	-.002	.469	.691	.506	.679	.535	.610	.502	.644
-1.5	6	.109	.073	.149	-.047	.149	-.025	.110	-.024	.264	.327	.285	.327	.270	.322	.258	.307
	15	.281	.026	.333	-.226	.321	-.138	.283	-.011	.350	.519	.393	.550	.381	.472	.359	.433
	24	.426	-.101	.475	-.231	.470	-.328	.454	-.068	.464	.754	.508	.761	.504	.755	.489	.656
-.5	6	.131	-.001	.125	-.063	.134	-.047	.159	-.046	.275	.304	.265	.328	.271	.293	.282	.300
	15	.302	-.162	.295	-.164	.312	-.135	.309	-.102	.381	.546	.361	.533	.367	.446	.365	.447
	24	.429	-.262	.431	-.318	.461	-.258	.460	-.224	.460	.790	.462	.780	.487	.683	.488	.683
.5	6	.126	-.133	.132	-.023	.134	-.014	.125	-.049	.276	.352	.266	.297	.265	.278	.264	.290
	15	.294	-.346	.319	-.148	.300	-.079	.329	-.202	.373	.665	.378	.457	.358	.421	.387	.475
	24	.454	-.515	.448	-.248	.457	-.218	.465	-.367	.493	.872	.489	.736	.492	.612	.503	.757
1.5	6	.153	-.222	.141	-.042	.120	-.007	.146	-.083	.290	.399	.272	.317	.262	.321	.284	.317
	15	.276	-.443	.332	-.100	.313	-.166	.312	-.316	.341	.738	.397	.526	.369	.416	.373	.558
	24	.452	-.683	.490	-.384	.486	-.186	.472	-.557	.494	1.004	.527	.779	.520	.676	.506	.869
2.5	6	.091	-.251	.149	-.019	.163	-.053	.128	-.126	.253	.444	.287	.338	.280	.340	.275	.351
	15	.247	-.630	.297	-.122	.318	-.088	.302	-.356	.336	.846	.376	.519	.378	.492	.368	.659
	24	.387	-.944	.433	-.691	.461	-.252	.434	-.738	.437	1.230	.473	.952	.495	.726	.470	1.030
3.5	6	-.088	-.318	-.032	-.170	-.045	-.166	-.052	-.184	.170	.476	.091	.273	.106	.264	.115	.286
	15	-.011	-.859	-.002	-.368	-.003	-.249	-.006	-.512	.048	1.073	.015	.530	.020	.365	.035	.737
	24	.000	-1.193	.000	-1.064	.000	-.534	.000	-1.002	.004	1.436	.000	1.279	.000	.755	.000	1.247

F.13 Magnitudes of nonuniform and uniform DIF were .4 and 1.0 respectively (Figures 16 and 31).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.121	.096	.181	.073	.164	.079	.151	.076	.205	.179	.261	.152	.248	.157	.234	.146
	15	.263	.048	.322	.047	.312	.053	.284	.053	.344	.117	.396	.159	.388	.128	.358	.129
	24	.393	.030	.444	.025	.475	.021	.423	.032	.454	.099	.506	.083	.526	.075	.485	.096
-2.5	6	.051	-.034	.125	-.107	.121	-.074	.096	-.061	.277	.251	.295	.286	.275	.300	.282	.277
	15	.241	-.208	.310	-.292	.283	-.224	.255	-.213	.339	.355	.393	.402	.369	.375	.362	.364
	24	.406	-.357	.436	-.423	.456	-.441	.417	-.376	.468	.469	.492	.506	.507	.537	.485	.486
-1.5	6	.074	-.075	.119	-.096	.091	-.051	.096	-.068	.259	.272	.289	.291	.263	.283	.288	.270
	15	.244	-.220	.302	-.274	.254	-.234	.260	-.191	.343	.346	.387	.407	.348	.379	.351	.339
	24	.421	-.430	.438	-.407	.428	-.406	.410	-.394	.484	.525	.493	.518	.478	.514	.469	.502
-.5	6	.124	-.105	.114	-.090	.110	-.079	.132	-.090	.273	.275	.274	.288	.269	.266	.282	.277
	15	.266	-.250	.279	-.258	.251	-.232	.291	-.235	.365	.384	.359	.386	.342	.363	.374	.360
	24	.415	-.411	.418	-.390	.400	-.364	.439	-.379	.468	.501	.475	.500	.452	.477	.488	.487
.5	6	.114	-.125	.120	-.057	.093	-.091	.116	-.090	.253	.294	.271	.258	.261	.277	.273	.277
	15	.278	-.261	.279	-.223	.287	-.196	.288	-.246	.355	.382	.375	.347	.376	.352	.366	.370
	24	.445	-.406	.416	-.381	.424	-.342	.431	-.403	.499	.506	.479	.486	.479	.455	.483	.501
1.5	6	.135	-.109	.107	-.087	.101	-.072	.145	-.090	.281	.275	.262	.270	.262	.276	.300	.267
	15	.260	-.272	.277	-.247	.272	-.221	.262	-.256	.360	.396	.369	.375	.373	.361	.354	.384
	24	.430	-.408	.465	-.402	.424	-.398	.439	-.425	.489	.501	.524	.496	.486	.484	.502	.511
2.5	6	.074	-.101	.126	-.096	.110	-.079	.104	-.074	.266	.284	.290	.302	.277	.299	.274	.294
	15	.239	-.285	.293	-.238	.277	-.256	.254	-.213	.348	.414	.393	.376	.375	.397	.365	.362
	24	.402	-.421	.442	-.455	.446	-.384	.421	-.403	.470	.527	.502	.543	.503	.491	.481	.514
3.5	6	-.090	-.140	-.042	-.166	-.045	-.155	-.061	-.156	.164	.229	.106	.246	.108	.245	.131	.242
	15	-.022	-.279	-.009	-.335	-.011	-.291	-.017	-.260	.072	.364	.049	.424	.047	.380	.059	.354
	24	-.004	-.425	-.003	-.472	-.002	-.467	-.004	-.433	.025	.516	.018	.561	.017	.558	.023	.525

F.14 Magnitudes of nonuniform and uniform DIF were .4 and 1.6 respectively (Figures 17 and 32).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.131	.102	.248	.050	.230	.058	.188	.067	.216	.185	.332	.119	.311	.123	.276	.140
	15	.375	.031	.516	.018	.465	.024	.433	.024	.451	.094	.571	.073	.525	.078	.489	.076
	24	.640	.007	.684	.009	.757	.005	.676	.011	.685	.037	.726	.047	.790	.030	.717	.048
-2.5	6	.104	-.064	.206	-.166	.176	-.118	.166	-.105	.293	.257	.336	.337	.304	.295	.305	.293
	15	.398	-.357	.456	-.441	.453	-.397	.396	-.328	.480	.447	.524	.531	.512	.491	.476	.437
	24	.647	-.590	.701	-.628	.691	-.671	.677	-.589	.686	.656	.737	.684	.722	.717	.714	.647
-1.5	6	.154	-.111	.165	-.137	.162	-.147	.173	-.113	.315	.287	.288	.319	.315	.313	.308	.291
	15	.402	-.383	.427	-.438	.420	-.356	.399	-.354	.474	.476	.491	.548	.482	.478	.469	.471
	24	.654	-.631	.675	-.690	.685	-.639	.685	-.606	.692	.703	.707	.763	.719	.718	.723	.691
-.5	6	.192	-.134	.153	-.158	.166	-.126	.168	-.138	.322	.296	.296	.310	.292	.298	.293	.290
	15	.431	-.421	.430	-.394	.420	-.370	.426	-.372	.495	.502	.495	.495	.476	.473	.493	.474
	24	.681	-.662	.652	-.672	.636	-.611	.679	-.631	.714	.733	.687	.745	.676	.690	.715	.703
.5	6	.188	-.186	.165	-.135	.157	-.125	.157	-.159	.311	.319	.294	.294	.292	.286	.292	.296
	15	.416	-.437	.434	-.365	.421	-.336	.435	-.401	.482	.521	.498	.456	.484	.437	.501	.504
	24	.681	-.660	.688	-.581	.684	-.581	.668	-.637	.717	.723	.724	.651	.720	.650	.709	.698
1.5	6	.155	-.181	.171	-.148	.173	-.130	.162	-.133	.289	.324	.318	.308	.302	.303	.303	.295
	15	.391	-.456	.450	-.364	.419	-.350	.420	-.399	.470	.528	.514	.456	.485	.449	.484	.487
	24	.662	-.670	.712	-.637	.710	-.587	.704	-.664	.701	.732	.752	.706	.750	.658	.744	.730
2.5	6	.070	-.145	.191	-.168	.191	-.137	.132	-.106	.263	.316	.339	.324	.328	.300	.290	.312
	15	.369	-.426	.496	-.419	.480	-.362	.385	-.367	.449	.517	.561	.519	.542	.485	.468	.478
	24	.616	-.674	.675	-.697	.722	-.670	.665	-.651	.662	.748	.715	.763	.754	.737	.700	.726
3.5	6	-.091	-.164	-.030	-.221	-.038	-.186	-.052	-.192	.170	.254	.088	.313	.098	.275	.122	.289
	15	-.005	-.385	-.004	-.477	-.003	-.391	-.009	-.386	.028	.477	.033	.550	.025	.472	.047	.473
	24	.000	-.641	.000	-.708	.000	-.707	.000	-.660	.000	.711	.000	.787	.000	.765	.004	.734

F.15 Magnitudes of nonuniform and uniform DIF were 1.0 and 1.6 respectively (Figures 18 and 33).

θ	#DIF Items	**BIAS** when DIF occurred in								**RMSE** when DIF occurred in							
		First Stages		**Middle Stages**		**Last Stages**		**Across Stages**		**First Stages**		**Middle Stages**		**Last Stages**		**Across Stages**	
		Ref.	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**	**Ref.**	**Foc.**
-3.5	6	.134	.110	.267	.067	.234	.094	.199	.108	.222	.197	.342	.141	.318	.186	.295	.197
	15	.376	.078	.554	.070	.502	.065	.469	.078	.440	.171	.596	.150	.554	.146	.530	.162
	24	.659	.055	.713	.064	.796	.047	.713	.056	.693	.131	.744	.173	.818	.139	.744	.134
-2.5	6	.107	-.014	.216	-.101	.196	-.113	.168	-.083	.290	.285	.341	.314	.313	.297	.317	.284
	15	.377	-.221	.489	-.371	.485	-.338	.421	-.249	.458	.411	.538	.519	.539	.483	.486	.414
	24	.651	-.421	.702	-.508	.736	-.560	.701	-.434	.683	.553	.731	.625	.760	.662	.727	.571
-1.5	6	.153	-.055	.193	-.147	.188	-.085	.177	-.093	.301	.301	.313	.326	.308	.304	.297	.307
	15	.421	-.315	.460	-.376	.458	-.298	.445	-.271	.479	.467	.510	.546	.506	.497	.502	.431
	24	.659	-.588	.704	-.618	.714	-.596	.702	-.529	.692	.738	.730	.784	.739	.747	.727	.674
-.5	6	.200	-.110	.186	-.105	.174	-.102	.200	-.106	.321	.293	.309	.309	.291	.319	.310	.303
	15	.447	-.386	.441	-.329	.438	-.311	.476	-.290	.501	.532	.492	.507	.488	.460	.524	.451
	24	.671	-.631	.671	-.635	.681	-.533	.695	-.564	.697	.778	.699	.783	.708	.683	.723	.703
.5	6	.182	-.188	.189	-.093	.187	-.085	.182	-.099	.301	.338	.306	.297	.301	.292	.311	.292
	15	.422	-.466	.485	-.301	.452	-.262	.473	-.324	.481	.603	.534	.464	.501	.415	.528	.474
	24	.698	-.709	.700	-.576	.720	-.509	.732	-.597	.724	.833	.727	.716	.746	.641	.757	.720
1.5	6	.183	-.196	.205	-.089	.178	-.092	.188	-.129	.301	.355	.321	.302	.303	.301	.311	.309
	15	.412	-.476	.478	-.270	.476	-.313	.432	-.371	.481	.602	.528	.437	.526	.457	.492	.504
	24	.685	-.741	.694	-.597	.770	-.547	.724	-.611	.716	.850	.723	.721	.793	.667	.753	.744
2.5	6	.105	-.192	.223	-.124	.210	-.095	.155	-.155	.259	.351	.349	.332	.346	.322	.302	.334
	15	.409	-.507	.496	-.334	.520	-.261	.433	-.388	.480	.634	.560	.520	.573	.469	.499	.559
	24	.699	-.799	.703	-.787	.756	-.587	.705	-.674	.736	.921	.733	.888	.779	.736	.736	.804
3.5	6	-.078	-.200	-.026	-.198	-.040	-.173	-.047	-.210	.151	.309	.080	.292	.109	.269	.109	.310
	15	-.005	-.540	.000	-.427	-.001	-.353	-.003	-.380	.030	.670	.004	.538	.006	.455	.019	.507
	24	.000	-.799	.000	-.848	.000	-.681	.000	-.769	.000	.925	.000	.967	.000	.790	.000	.898

Appendix G: Observed SE as presented in Study 1.

G.1 Observed SE for the baseline condition or DIF-free CAT

θ		Items/CAT stages																												
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
-3.5	**R**	.35	.41	.44	.47	.49	.49	.49	.48	.40	.37	.33	.31	.29	.27	.26	.25	.24	.23	.22	.21	.20	.20	.18	.18	.17	.17	.16	.16	.16	.17
	F	.34	.39	.41	.44	.47	.51	.51	.50	.45	.41	.37	.34	.31	.29	.26	.23	.22	.21	.20	.19	.18	.18	.17	.17	.17	.16	.16	.16	.16	.16
3.5	**R**	.30	.30	.31	.32	.35	.39	.40	.39	.33	.30	.27	.26	.25	.23	.22	.21	.20	.20	.19	.19	.18	.18	.18	.18	.17	.17	.16	.16	.16	.16
	F	.29	.29	.30	.32	.34	.38	.39	.39	.35	.32	.29	.26	.24	.24	.22	.21	.21	.20	.20	.20	.19	.19	.18	.17	.17	.16	.16	.16	.16	.15

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups.

G.2 Observed SE when 6 items showed uniform DIF with a magnitude of .4 (Figure 34).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.32	.38	.43	.49	.54	.57	.63	.58	.54	.51	.48	.45	.43	.41	.39	.37	.37	.35	.33	.32	.31	.30	.29	.29	.28	.28	.27	.27	.26	.26
		.35	.40	.43	.49	.52	.58	.65	.61	.56	.52	.48	.47	.44	.42	.40	.39	.38	.36	.35	.35	.33	.33	.32	.31	.30	.30	.29	.28	.28	.28
		.35	.41	.45	.49	.53	.57	.62	.57	.53	.50	.48	.46	.44	.43	.41	.39	.38	.38	.36	.35	.34	.33	.33	.32	.31	.30	.30	.30	.28	.28
		.34	.39	.44	.49	.54	.57	.63	.59	.54	.51	.48	.46	.43	.43	.41	.39	.38	.36	.35	.34	.33	.32	.31	.30	.30	.29	.28	.27	.27	.26
	F	.34	.40	.45	.48	.52	.56	.60	.55	.49	.47	.43	.40	.38	.37	.36	.34	.33	.32	.31	.31	.30	.30	.29	.27	.27	.26	.26	.25	.25	.25
		.36	.41	.44	.48	.51	.55	.60	.55	.51	.47	.45	.43	.41	.39	.37	.36	.34	.33	.32	.31	.31	.30	.29	.28	.28	.27	.27	.26	.26	.25
		.38	.45	.49	.52	.57	.61	.65	.59	.55	.51	.49	.46	.43	.41	.40	.38	.37	.35	.34	.33	.32	.31	.31	.30	.29	.28	.28	.27	.27	.26
		.35	.42	.46	.49	.53	.57	.61	.55	.52	.49	.47	.45	.42	.41	.39	.38	.37	.36	.35	.34	.33	.32	.31	.30	.29	.28	.28	.27	.27	.26
θ = 2.5	R	.28	.29	.31	.34	.41	.47	.51	.52	.49	.46	.44	.42	.39	.37	.35	.34	.33	.33	.32	.32	.31	.30	.30	.29	.29	.28	.27	.27	.26	.26
		.30	.32	.35	.38	.45	.51	.54	.53	.50	.47	.45	.43	.41	.39	.38	.37	.35	.34	.34	.33	.33	.32	.31	.31	.30	.30	.30	.29	.28	.27
		.30	.31	.34	.38	.45	.50	.54	.53	.50	.48	.46	.43	.41	.40	.38	.37	.36	.34	.33	.32	.32	.31	.30	.30	.29	.28	.28	.28	.27	.27
		.31	.32	.34	.37	.44	.50	.53	.53	.50	.47	.43	.40	.39	.36	.36	.36	.34	.33	.32	.32	.31	.31	.30	.29	.28	.28	.28	.27	.27	.27
	F	.29	.31	.34	.40	.48	.54	.55	.53	.50	.48	.46	.42	.41	.39	.38	.36	.35	.35	.34	.34	.33	.32	.32	.32	.31	.30	.29	.29	.29	.28
		.28	.29	.32	.36	.44	.50	.55	.54	.50	.48	.46	.44	.41	.40	.37	.36	.35	.34	.33	.33	.32	.32	.31	.30	.30	.29	.28	.28	.27	.27
		.29	.30	.33	.37	.44	.51	.55	.55	.52	.48	.44	.41	.38	.36	.34	.33	.32	.31	.30	.29	.29	.28	.28	.27	.27	.26	.26	.25	.24	.24
		.29	.31	.33	.39	.45	.53	.57	.56	.52	.48	.46	.43	.41	.39	.37	.36	.34	.33	.33	.32	.31	.30	.29	.29	.29	.28	.27	.27	.27	.26

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.3 Observed SE when 24 items showed uniform DIF with a magnitude of .4 (Figure 35).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.33	.38	.42	.46	.51	.56	.61	.54	.50	.46	.43	.42	.41	.40	.38	.37	.35	.33	.33	.33	.32	.31	.30	.29	.29	.27	.27	.26	.25	.25
		.36	.42	.46	.49	.54	.59	.60	.55	.50	.47	.44	.42	.40	.38	.36	.35	.34	.34	.33	.32	.31	.31	.30	.29	.29	.28	.27	.27	.26	.26
		.35	.42	.46	.49	.54	.59	.60	.56	.53	.49	.46	.43	.42	.39	.37	.36	.34	.33	.33	.31	.30	.29	.29	.28	.28	.27	.27	.26	.26	.26
		.33	.38	.41	.45	.51	.57	.61	.55	.49	.48	.44	.43	.41	.40	.39	.38	.37	.36	.35	.34	.33	.33	.32	.31	.31	.30	.30	.29	.28	.28
	F	.35	.40	.44	.48	.51	.55	.59	.55	.49	.46	.44	.41	.40	.39	.38	.36	.35	.34	.33	.32	.32	.31	.30	.30	.29	.29	.28	.27	.27	.27
		.34	.39	.42	.44	.48	.54	.60	.56	.52	.48	.45	.44	.42	.40	.38	.36	.35	.34	.33	.32	.32	.31	.30	.29	.29	.28	.27	.27	.27	.26
		.34	.38	.42	.46	.51	.56	.61	.58	.55	.51	.48	.45	.43	.42	.41	.39	.38	.37	.36	.35	.34	.33	.32	.31	.30	.30	.29	.28	.28	.28
		.36	.40	.43	.48	.52	.55	.62	.59	.54	.50	.48	.46	.44	.43	.41	.40	.37	.36	.35	.34	.33	.33	.31	.30	.30	.29	.28	.28	.27	.27
θ =2.5	R	.30	.31	.33	.37	.43	.48	.54	.53	.51	.48	.45	.43	.42	.40	.38	.37	.36	.35	.34	.34	.33	.32	.32	.31	.31	.30	.30	.29	.28	.28
		.30	.33	.36	.41	.47	.53	.57	.56	.53	.49	.47	.44	.41	.39	.38	.36	.34	.34	.33	.31	.31	.30	.29	.28	.28	.27	.26	.26	.26	.25
		.30	.32	.34	.37	.44	.51	.55	.55	.51	.48	.45	.43	.40	.38	.36	.35	.34	.32	.31	.31	.30	.29	.29	.29	.29	.28	.28	.28	.27	.27
		.31	.31	.34	.37	.43	.50	.56	.56	.53	.48	.47	.45	.42	.40	.39	.38	.37	.36	.35	.34	.33	.31	.30	.29	.29	.28	.28	.28	.27	.27
	F	.30	.32	.37	.42	.48	.54	.56	.53	.49	.47	.44	.42	.40	.39	.37	.35	.34	.33	.32	.32	.31	.30	.29	.28	.27	.27	.26	.26	.26	.25
		.29	.30	.33	.37	.45	.54	.57	.56	.52	.49	.46	.42	.40	.39	.37	.36	.34	.33	.32	.31	.30	.29	.29	.28	.28	.27	.26	.26	.26	.25
		.30	.32	.35	.40	.46	.53	.57	.54	.51	.49	.46	.44	.41	.38	.37	.36	.35	.34	.33	.32	.31	.30	.30	.30	.29	.28	.28	.27	.26	.26
		.29	.31	.34	.39	.47	.55	.59	.57	.53	.50	.47	.44	.43	.41	.39	.38	.36	.35	.34	.33	.32	.32	.31	.30	.30	.29	.29	.28	.28	.27

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.4 Observed SE when 6 items showed uniform DIF with a magnitude of 1.6 (Figure 36).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.37	.45	.50	.55	.60	.65	.69	.63	.57	.53	.53	.49	.47	.44	.42	.41	.39	.37	.35	.34	.33	.32	.31	.31	.30	.30	.29	.28	.28	.27
		.36	.41	.44	.48	.55	.58	.63	.58	.54	.49	.46	.45	.43	.40	.38	.37	.36	.35	.35	.34	.33	.32	.32	.31	.30	.29	.29	.28	.28	.28
		.36	.42	.46	.50	.55	.58	.62	.57	.53	.50	.48	.46	.42	.39	.38	.36	.35	.34	.33	.32	.32	.31	.30	.29	.29	.28	.28	.28	.27	.27
		.36	.41	.45	.50	.57	.60	.67	.61	.55	.51	.47	.45	.42	.41	.39	.37	.36	.34	.33	.32	.31	.31	.30	.30	.29	.28	.28	.27	.27	.27
	F	.30	.30	.30	.32	.35	.38	.44	.46	.46	.44	.42	.41	.38	.36	.35	.34	.34	.33	.32	.32	.31	.31	.30	.30	.29	.29	.28	.27	.26	.26
		.30	.31	.35	.38	.45	.51	.54	.56	.53	.50	.48	.47	.45	.43	.42	.40	.39	.37	.36	.35	.34	.33	.32	.31	.31	.30	.29	.28	.28	.28
		.29	.30	.34	.37	.42	.50	.57	.56	.52	.49	.46	.42	.40	.38	.37	.36	.34	.34	.33	.32	.32	.31	.31	.30	.30	.30	.30	.30	.29	.29
		.29	.30	.32	.35	.40	.46	.51	.51	.47	.44	.42	.41	.40	.39	.37	.35	.35	.33	.33	.32	.31	.30	.30	.30	.29	.29	.28	.27	.27	.27
θ = 2.5	R	.35	.40	.44	.48	.50	.51	.57	.54	.49	.47	.44	.41	.40	.38	.37	.36	.35	.34	.34	.33	.33	.31	.31	.30	.29	.29	.28	.28	.27	.26
		.32	.37	.42	.47	.51	.57	.62	.57	.52	.49	.46	.46	.44	.43	.41	.39	.38	.37	.36	.35	.34	.33	.32	.30	.30	.29	.29	.28	.28	.27
		.36	.41	.45	.49	.54	.59	.65	.58	.53	.49	.45	.43	.41	.40	.38	.38	.37	.36	.35	.34	.33	.33	.32	.31	.31	.30	.30	.29	.28	.28
		.36	.40	.43	.45	.48	.50	.55	.51	.47	.44	.40	.39	.38	.38	.36	.35	.34	.34	.32	.32	.31	.30	.29	.28	.28	.27	.27	.26	.26	.26
	F	.34	.40	.46	.53	.57	.58	.60	.56	.54	.52	.48	.46	.44	.41	.40	.38	.37	.36	.35	.33	.32	.32	.31	.30	.29	.29	.29	.29	.28	.27
		.29	.29	.31	.36	.44	.52	.53	.53	.50	.48	.44	.41	.40	.38	.36	.35	.35	.33	.32	.32	.31	.31	.30	.29	.29	.28	.28	.28	.28	.27
		.29	.31	.35	.38	.45	.53	.55	.54	.52	.48	.45	.43	.41	.40	.38	.37	.36	.35	.34	.33	.33	.32	.30	.30	.29	.29	.28	.27	.27	.26
		.30	.34	.39	.45	.53	.59	.62	.59	.55	.52	.50	.48	.46	.44	.42	.40	.39	.37	.36	.35	.34	.34	.33	.32	.31	.30	.30	.29	.29	.28

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.5 Observed SE when 24 items showed uniform DIF with a magnitude of 1.6 (Figure 37).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.38	.45	.50	.55	.61	.62	.65	.56	.51	.48	.44	.42	.39	.38	.36	.36	.35	.34	.33	.32	.31	.31	.30	.30	.29	.29	.29	.29	.28	.28
		.39	.47	.50	.53	.59	.61	.57	.53	.48	.47	.45	.43	.42	.40	.38	.37	.34	.33	.32	.31	.30	.29	.29	.28	.28	.28	.28	.28	.28	.27
		.32	.36	.39	.43	.47	.52	.56	.49	.48	.46	.43	.40	.39	.37	.36	.36	.34	.33	.32	.31	.30	.30	.29	.28	.28	.27	.27	.26	.26	.26
		.38	.47	.53	.56	.63	.66	.64	.54	.49	.47	.45	.44	.42	.42	.38	.37	.36	.35	.33	.32	.31	.31	.30	.29	.29	.28	.28	.27	.27	.26
	F	.34	.39	.42	.43	.44	.47	.50	.51	.45	.42	.39	.37	.35	.33	.32	.31	.30	.29	.28	.27	.26	.26	.25	.24	.25	.25	.25	.25	.25	.25
		.34	.39	.43	.45	.48	.52	.53	.51	.47	.43	.39	.36	.35	.34	.32	.31	.30	.29	.28	.27	.27	.26	.25	.24	.24	.23	.23	.23	.24	.24
		.34	.39	.43	.48	.53	.57	.66	.65	.59	.55	.50	.47	.44	.41	.39	.37	.35	.34	.32	.31	.29	.28	.27	.27	.26	.26	.25	.25	.24	.24
		.33	.38	.42	.46	.49	.53	.55	.52	.48	.47	.44	.42	.40	.37	.36	.34	.34	.32	.31	.31	.30	.29	.29	.28	.27	.27	.26	.26	.26	.25
θ = 2.5	R	.28	.28	.29	.30	.33	.39	.44	.45	.42	.39	.38	.35	.34	.33	.32	.30	.30	.28	.28	.27	.26	.25	.25	.24	.25	.25	.26	.26	.26	.26
		.30	.31	.33	.36	.40	.44	.48	.49	.46	.42	.41	.38	.36	.34	.33	.31	.29	.28	.27	.27	.26	.25	.24	.23	.23	.23	.23	.23	.24	.24
		.31	.32	.35	.39	.45	.51	.54	.55	.53	.50	.47	.44	.42	.40	.39	.37	.36	.35	.34	.33	.32	.31	.30	.29	.29	.28	.27	.27	.26	.26
		.29	.30	.31	.34	.39	.43	.47	.47	.44	.43	.41	.40	.38	.37	.35	.34	.34	.33	.32	.31	.30	.30	.29	.28	.27	.27	.26	.26	.26	.25
	F	.33	.38	.43	.51	.56	.57	.56	.52	.48	.44	.41	.38	.36	.36	.34	.33	.32	.32	.31	.30	.29	.29	.28	.27	.28	.27	.27	.27	.27	.26
		.29	.29	.35	.40	.48	.52	.52	.49	.45	.43	.41	.39	.38	.37	.36	.35	.34	.33	.33	.33	.31	.30	.30	.29	.28	.28	.28	.27	.27	.27
		.31	.32	.36	.40	.45	.52	.55	.51	.47	.45	.43	.40	.39	.38	.36	.35	.34	.33	.32	.31	.30	.29	.29	.28	.27	.26	.26	.26	.25	.24
		.33	.38	.42	.48	.53	.59	.58	.54	.50	.49	.47	.45	.43	.41	.40	.38	.37	.36	.35	.34	.33	.32	.32	.31	.30	.30	.29	.29	.28	.28

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.6 Observed SE when 6 items showed nonuniform DIF with a magnitude of .4 (Figure 38).

		Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R		.32	.37	.40	.45	.48	.52	.58	.55	.51	.47	.44	.41	.39	.38	.35	.35	.34	.33	.32	.31	.31	.31	.31	.30	.30	.29	.29	.28	.27	.27
			.37	.43	.47	.51	.56	.60	.63	.60	.56	.52	.48	.46	.43	.41	.39	.37	.35	.34	.32	.32	.31	.31	.30	.29	.28	.28	.28	.27	.27	.26
			.35	.41	.44	.48	.51	.55	.59	.53	.51	.48	.45	.44	.42	.40	.38	.37	.35	.34	.34	.33	.32	.31	.30	.30	.29	.28	.27	.27	.26	.25
			.32	.37	.41	.44	.48	.53	.60	.57	.52	.49	.46	.43	.41	.39	.38	.37	.35	.34	.33	.32	.31	.30	.29	.29	.28	.27	.27	.27	.26	.26
	F		.35	.41	.45	.50	.57	.63	.69	.63	.57	.54	.51	.48	.46	.43	.41	.40	.38	.36	.35	.34	.33	.32	.32	.31	.31	.30	.29	.29	.28	.28
			.34	.41	.45	.49	.53	.58	.62	.56	.51	.47	.45	.43	.41	.40	.39	.38	.37	.35	.35	.33	.33	.32	.32	.31	.30	.29	.29	.28	.28	.27
			.34	.39	.44	.48	.53	.58	.65	.62	.57	.53	.49	.46	.44	.42	.40	.39	.38	.36	.35	.34	.32	.31	.31	.30	.30	.30	.29	.28	.28	.28
			.32	.39	.43	.48	.52	.56	.63	.59	.54	.49	.48	.45	.43	.41	.41	.39	.39	.37	.37	.36	.34	.33	.33	.32	.31	.31	.30	.29	.28	.28
θ = 2.5	R		.30	.31	.32	.35	.41	.47	.51	.52	.48	.46	.44	.41	.39	.36	.34	.34	.34	.33	.32	.31	.30	.29	.29	.28	.28	.27	.27	.27	.27	.26
			.30	.31	.33	.39	.46	.52	.56	.53	.49	.47	.45	.42	.40	.37	.36	.34	.33	.32	.31	.30	.29	.29	.28	.28	.27	.27	.26	.26	.25	.25
			.30	.31	.34	.40	.48	.54	.58	.57	.52	.48	.45	.42	.41	.39	.38	.37	.35	.34	.33	.33	.32	.31	.30	.29	.28	.28	.26	.26	.26	.25
			.30	.31	.32	.36	.41	.46	.52	.51	.47	.45	.42	.41	.40	.39	.38	.36	.36	.34	.34	.33	.32	.31	.30	.28	.27	.27	.26	.26	.25	.25
	F		.32	.35	.40	.46	.53	.60	.63	.61	.57	.53	.50	.46	.44	.42	.40	.39	.38	.37	.36	.35	.34	.33	.33	.32	.31	.31	.30	.29	.28	.27
			.29	.31	.33	.37	.43	.50	.55	.55	.50	.47	.45	.43	.41	.40	.39	.38	.37	.36	.35	.34	.33	.32	.32	.31	.31	.30	.29	.29	.28	.28
			.29	.29	.32	.36	.45	.52	.55	.55	.51	.48	.45	.43	.41	.39	.38	.37	.35	.34	.33	.33	.32	.31	.30	.29	.29	.28	.28	.27	.27	.27
			.29	.30	.34	.39	.47	.55	.57	.54	.51	.47	.44	.43	.41	.39	.38	.37	.36	.34	.33	.32	.31	.31	.30	.30	.29	.28	.28	.28	.27	.27

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.7 Observed SE when 24 items showed nonuniform DIF with a magnitude of .4 (Figure 39).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.36	.43	.47	.50	.51	.51	.58	.52	.47	.43	.41	.38	.37	.35	.33	.32	.30	.29	.29	.28	.27	.26	.25	.25	.24	.24	.24	.24	.23	.23
		.37	.42	.45	.46	.47	.50	.54	.51	.45	.42	.40	.38	.35	.34	.33	.31	.30	.29	.28	.27	.27	.26	.25	.25	.24	.23	.23	.23	.23	.22
		.33	.37	.41	.47	.51	.56	.61	.55	.50	.48	.44	.41	.40	.37	.34	.33	.31	.31	.30	.29	.28	.28	.27	.26	.26	.25	.25	.24	.24	.23
		.34	.38	.41	.43	.45	.49	.55	.48	.44	.42	.39	.37	.35	.35	.33	.32	.31	.30	.29	.29	.28	.27	.26	.26	.25	.24	.24	.23	.23	.23
	F	.36	.43	.48	.54	.59	.65	.71	.66	.61	.57	.54	.51	.48	.46	.44	.42	.41	.40	.39	.38	.37	.36	.35	.34	.33	.32	.32	.31	.31	.30
		.35	.41	.47	.53	.59	.64	.68	.65	.59	.56	.53	.50	.48	.47	.46	.44	.42	.41	.40	.39	.38	.37	.36	.35	.35	.35	.33	.33	.32	.32
		.37	.43	.47	.52	.57	.61	.67	.63	.57	.54	.52	.50	.48	.46	.45	.42	.40	.39	.39	.37	.36	.36	.34	.33	.32	.32	.31	.30	.30	.29
		.36	.42	.45	.49	.57	.62	.67	.62	.58	.54	.52	.49	.46	.45	.43	.42	.40	.39	.37	.36	.34	.34	.33	.32	.32	.31	.31	.30	.30	.29
θ = 2.5	R	.30	.31	.32	.35	.43	.47	.52	.49	.45	.42	.39	.37	.34	.32	.31	.30	.29	.29	.28	.27	.26	.25	.25	.24	.23	.24	.23	.23	.23	.22
		.31	.33	.35	.39	.45	.50	.53	.52	.47	.43	.40	.38	.36	.34	.32	.31	.30	.29	.28	.26	.26	.25	.24	.24	.23	.23	.22	.22	.22	.22
		.29	.30	.32	.36	.44	.51	.53	.52	.47	.44	.42	.38	.36	.35	.32	.31	.30	.29	.28	.27	.27	.26	.26	.25	.24	.23	.23	.23	.23	.23
		.29	.30	.32	.35	.41	.49	.51	.49	.44	.43	.41	.39	.37	.34	.33	.31	.30	.28	.27	.27	.26	.26	.25	.24	.24	.24	.23	.23	.22	.22
	F	.30	.33	.36	.43	.51	.57	.61	.61	.58	.56	.54	.52	.50	.48	.47	.45	.44	.42	.41	.40	.40	.39	.38	.38	.37	.36	.35	.34	.33	.32
		.30	.31	.35	.40	.49	.55	.61	.59	.57	.56	.54	.52	.49	.48	.46	.44	.42	.40	.39	.38	.37	.36	.35	.35	.34	.34	.33	.32	.31	.31
		.28	.29	.31	.35	.43	.52	.58	.58	.55	.54	.50	.49	.47	.46	.45	.44	.43	.41	.40	.39	.37	.37	.37	.36	.35	.34	.33	.33	.32	.32
		.32	.34	.39	.44	.52	.62	.66	.66	.61	.57	.55	.51	.50	.48	.47	.46	.44	.42	.40	.39	.38	.37	.37	.37	.36	.35	.34	.33	.32	.31

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.8 Observed SE when 6 items showed nonuniform DIF with a magnitude of 1.6 (Figure 40).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.35	.39	.41	.43	.42	.41	.50	.46	.43	.41	.41	.39	.38	.37	.36	.35	.35	.35	.34	.33	.33	.31	.30	.30	.29	.28	.28	.27	.27	.26
		.35	.41	.46	.51	.53	.58	.60	.56	.52	.48	.46	.42	.37	.34	.31	.28	.27	.26	.26	.26	.25	.25	.24	.24	.24	.24	.23	.23	.23	.22
		.37	.43	.45	.48	.52	.55	.61	.56	.51	.47	.46	.43	.42	.39	.38	.36	.34	.34	.33	.33	.32	.31	.31	.30	.28	.28	.27	.26	.25	.24
		.36	.42	.46	.50	.53	.57	.62	.58	.53	.50	.47	.44	.41	.39	.37	.36	.33	.32	.31	.29	.29	.28	.28	.27	.26	.25	.25	.24	.24	.24
	F	.45	.57	.70	.80	.90	.98	1.02	.92	.86	.82	.78	.73	.69	.65	.61	.58	.56	.53	.50	.47	.46	.44	.42	.40	.39	.38	.36	.35	.34	.33
		.33	.37	.41	.45	.49	.54	.59	.55	.52	.49	.46	.47	.49	.51	.51	.51	.51	.49	.47	.45	.44	.43	.41	.40	.38	.36	.35	.35	.34	.33
		.36	.42	.46	.50	.53	.57	.63	.56	.51	.48	.46	.43	.41	.39	.38	.36	.35	.35	.34	.33	.32	.31	.30	.29	.29	.30	.30	.30	.31	.31
		.35	.47	.54	.64	.70	.74	.80	.73	.66	.60	.56	.52	.49	.47	.44	.41	.42	.40	.39	.40	.39	.38	.36	.36	.35	.35	.34	.34	.32	.32
θ = 2.5	R	.29	.29	.29	.29	.32	.35	.42	.45	.43	.41	.38	.36	.35	.33	.32	.31	.30	.30	.29	.29	.28	.28	.27	.27	.26	.26	.26	.26	.25	.25
		.31	.32	.34	.38	.46	.52	.56	.55	.52	.48	.45	.41	.37	.34	.32	.29	.28	.27	.27	.26	.26	.26	.26	.25	.25	.25	.24	.24	.23	.23
		.30	.32	.35	.39	.45	.49	.53	.54	.50	.48	.47	.45	.43	.41	.39	.37	.36	.35	.34	.33	.32	.31	.31	.30	.28	.27	.26	.25	.24	.23
		.30	.30	.33	.35	.41	.46	.51	.51	.48	.46	.43	.40	.38	.37	.36	.35	.33	.32	.32	.30	.29	.28	.28	.26	.26	.25	.24	.24	.24	.24
	F	.43	.55	.69	.80	.89	.96	.94	.92	.87	.82	.78	.74	.71	.67	.63	.60	.57	.54	.51	.49	.47	.45	.43	.42	.41	.40	.39	.38	.37	.36
		.29	.30	.33	.37	.43	.51	.55	.55	.52	.49	.47	.49	.50	.51	.52	.53	.53	.52	.50	.48	.47	.45	.44	.43	.41	.40	.39	.38	.37	.36
		.30	.31	.34	.37	.45	.52	.56	.54	.51	.47	.44	.42	.39	.37	.36	.34	.33	.32	.31	.31	.30	.30	.29	.28	.29	.29	.30	.30	.31	.31
		.30	.37	.42	.53	.61	.65	.66	.64	.60	.55	.51	.48	.46	.44	.42	.40	.41	.39	.37	.38	.37	.35	.34	.35	.34	.35	.34	.33	.32	.32

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.9 Observed SE when 24 items showed nonuniform DIF with a magnitude of 1.6 (Figure 41).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.35	.39	.43	.45	.46	.44	.48	.39	.36	.33	.30	.29	.27	.25	.24	.23	.22	.21	.20	.19	.19	.18	.18	.17	.18	.18	.18	.18	.18	.18
		.34	.40	.44	.47	.47	.47	.51	.44	.41	.38	.34	.31	.29	.27	.26	.25	.23	.21	.20	.20	.19	.19	.19	.18	.18	.17	.18	.18	.18	.18
		.37	.43	.46	.50	.53	.54	.56	.46	.39	.35	.33	.30	.28	.28	.26	.24	.23	.22	.21	.21	.19	.19	.18	.18	.17	.17	.16	.16	.16	.15
		.35	.39	.43	.45	.45	.44	.50	.44	.39	.38	.34	.35	.31	.29	.27	.25	.25	.23	.23	.22	.22	.22	.21	.20	.19	.19	.19	.19	.19	.18
	F	.44	.57	.70	.80	.91	.97	1.04	1.03	1.01	1.00	.99	.96	.95	.94	.93	.92	.92	.91	.91	.89	.89	.88	.87	.87	.84	.82	.79	.77	.74	.72
		.36	.43	.51	.62	.74	.85	.92	.91	.90	.89	.88	.87	.86	.85	.85	.85	.84	.83	.82	.81	.81	.80	.80	.79	.79	.78	.76	.74	.72	.70
		.35	.42	.47	.52	.56	.58	.68	.70	.69	.69	.68	.68	.67	.66	.65	.65	.65	.64	.64	.63	.63	.63	.62	.62	.62	.61	.60	.60	.60	.59
		.44	.58	.68	.78	.89	.98	1.02	1.00	.97	.91	.90	.85	.85	.84	.83	.83	.79	.79	.78	.78	.76	.74	.73	.72	.71	.71	.70	.68	.67	.67
θ = 2.5	R	.29	.29	.29	.30	.33	.36	.38	.38	.34	.31	.28	.26	.23	.22	.21	.20	.20	.20	.19	.18	.17	.17	.16	.16	.16	.17	.17	.17	.17	.17
		.29	.29	.30	.32	.36	.39	.40	.40	.36	.32	.30	.28	.27	.25	.24	.22	.22	.21	.21	.20	.20	.19	.18	.18	.17	.16	.16	.17	.17	.17
		.28	.30	.33	.38	.46	.54	.52	.47	.41	.36	.33	.31	.28	.26	.24	.22	.21	.21	.20	.19	.18	.18	.17	.17	.16	.16	.15	.15	.15	.15
		.28	.28	.30	.32	.36	.40	.40	.39	.37	.37	.32	.32	.29	.27	.25	.24	.25	.23	.22	.21	.20	.20	.19	.19	.19	.18	.18	.18	.17	.18
	F	.42	.54	.64	.75	.84	.91	.95	.94	.94	.94	.93	.91	.91	.90	.89	.89	.88	.88	.87	.86	.86	.85	.85	.85	.83	.81	.79	.77	.75	.73
		.29	.30	.37	.49	.62	.72	.78	.79	.79	.78	.78	.78	.78	.77	.77	.77	.77	.77	.76	.76	.76	.76	.76	.76	.75	.75	.73	.71	.69	.68
		.31	.33	.36	.40	.47	.52	.62	.68	.68	.68	.68	.67	.68	.68	.68	.68	.67	.67	.67	.67	.66	.67	.67	.67	.67	.67	.67	.67	.67	.67
		.45	.59	.67	.76	.84	.89	.93	.93	.93	.88	.86	.82	.81	.80	.80	.80	.76	.76	.75	.75	.74	.72	.72	.71	.71	.71	.70	.68	.67	.67

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.10 Observed SE when 6 items showed nonuniform and uniform DIF both with magnitudes of .4 (Figure 42).

		Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R		.35	.39	.42	.46	.49	.53	.58	.54	.50	.46	.44	.42	.41	.39	.37	.35	.35	.34	.33	.32	.31	.30	.29	.29	.28	.27	.27	.27	.26	.26
			.32	.37	.41	.46	.52	.57	.61	.55	.51	.47	.44	.42	.38	.36	.33	.32	.31	.30	.29	.28	.28	.27	.27	.26	.26	.25	.25	.25	.24	.24
			.33	.37	.40	.44	.50	.56	.61	.57	.53	.48	.45	.43	.41	.39	.37	.35	.34	.34	.32	.31	.31	.30	.29	.28	.27	.27	.26	.25	.25	.24
			.37	.42	.46	.50	.55	.59	.61	.60	.56	.50	.47	.45	.44	.42	.39	.38	.36	.34	.33	.32	.31	.31	.30	.30	.29	.29	.28	.27	.27	.27
	F		.34	.39	.44	.48	.55	.62	.65	.60	.55	.52	.49	.46	.44	.42	.41	.38	.37	.36	.35	.35	.34	.34	.33	.32	.31	.30	.30	.29	.29	.28	.28
			.38	.43	.45	.47	.51	.53	.59	.56	.52	.48	.46	.42	.42	.41	.39	.38	.37	.35	.35	.34	.33	.32	.31	.31	.29	.29	.29	.28	.27	.27
			.34	.39	.42	.47	.52	.57	.63	.59	.52	.48	.47	.44	.43	.40	.38	.37	.35	.34	.32	.32	.32	.31	.31	.30	.30	.29	.29	.28	.28	.27
			.34	.39	.41	.46	.51	.56	.61	.57	.51	.47	.45	.44	.41	.39	.37	.35	.34	.33	.33	.32	.31	.30	.29	.29	.28	.28	.27	.27	.26	.26
θ = 2.5	R		.29	.29	.29	.31	.36	.41	.48	.50	.49	.45	.43	.40	.38	.37	.35	.34	.33	.33	.32	.31	.31	.31	.30	.29	.29	.28	.28	.28	.27	.26
			.31	.33	.36	.42	.48	.54	.56	.55	.50	.47	.45	.42	.39	.38	.36	.35	.33	.32	.31	.30	.30	.29	.29	.28	.28	.27	.27	.26	.26	.26
			.28	.29	.32	.35	.43	.49	.54	.54	.51	.47	.45	.43	.42	.39	.39	.37	.36	.34	.33	.32	.32	.31	.30	.29	.28	.28	.27	.27	.26	.26
			.31	.33	.34	.37	.42	.48	.53	.53	.49	.46	.43	.41	.38	.37	.36	.35	.33	.32	.31	.31	.30	.29	.28	.27	.27	.26	.26	.26	.25	.25
	F		.31	.34	.39	.45	.52	.60	.61	.58	.55	.51	.47	.45	.43	.41	.39	.37	.36	.35	.34	.34	.32	.32	.31	.30	.29	.28	.27	.27	.26	.27
			.30	.31	.33	.37	.44	.50	.55	.54	.51	.50	.48	.47	.45	.43	.42	.41	.40	.39	.37	.36	.35	.34	.33	.33	.32	.32	.31	.30	.29	.29
			.30	.31	.33	.36	.43	.50	.54	.54	.51	.48	.46	.43	.41	.39	.37	.35	.34	.33	.33	.32	.31	.31	.30	.29	.29	.28	.28	.28	.28	.28
			.29	.30	.34	.41	.51	.56	.61	.59	.54	.51	.48	.45	.42	.40	.38	.37	.36	.35	.34	.34	.33	.32	.32	.31	.31	.30	.29	.28	.27	.27

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.11 Observed SE when 24 items showed nonuniform and uniform DIF both with magnitudes of .4 (Figure 42).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.35	.39	.42	.46	.50	.51	.55	.49	.46	.42	.40	.38	.36	.35	.32	.31	.30	.29	.28	.27	.27	.27	.26	.25	.25	.24	.24	.24	.24	.23
		.33	.38	.42	.47	.50	.51	.56	.49	.45	.40	.39	.38	.36	.34	.34	.33	.31	.30	.28	.27	.27	.26	.26	.25	.24	.24	.24	.23	.23	.22
		.33	.38	.41	.44	.48	.52	.57	.49	.46	.41	.40	.37	.36	.34	.33	.30	.29	.28	.28	.27	.26	.26	.25	.24	.24	.23	.23	.22	.22	.22
		.34	.40	.42	.45	.49	.50	.53	.50	.46	.44	.40	.39	.37	.34	.33	.32	.32	.30	.30	.29	.28	.28	.27	.26	.25	.24	.24	.24	.24	.23
	F	.37	.44	.47	.52	.60	.64	.71	.69	.61	.57	.55	.52	.49	.47	.46	.44	.43	.41	.40	.39	.37	.36	.35	.34	.33	.32	.32	.31	.30	.30
		.38	.44	.47	.51	.56	.61	.66	.63	.59	.55	.53	.50	.49	.46	.44	.42	.41	.40	.39	.38	.37	.36	.36	.35	.35	.34	.33	.32	.30	.30
		.34	.39	.42	.45	.48	.54	.60	.59	.56	.54	.51	.48	.47	.45	.44	.44	.43	.41	.40	.39	.38	.37	.36	.35	.35	.34	.33	.32	.32	.31
		.34	.41	.46	.50	.56	.62	.67	.63	.57	.53	.51	.47	.46	.44	.44	.42	.41	.40	.39	.38	.37	.36	.35	.34	.34	.33	.32	.31	.31	.30
θ = 2.5	R	.30	.30	.31	.33	.36	.39	.45	.45	.42	.39	.37	.35	.34	.33	.33	.32	.31	.30	.29	.29	.28	.27	.26	.25	.25	.25	.25	.24	.24	.23
		.30	.31	.32	.35	.40	.45	.49	.51	.48	.44	.41	.38	.36	.34	.33	.32	.31	.30	.29	.28	.28	.27	.27	.26	.26	.26	.25	.25	.25	.24
		.30	.31	.33	.37	.45	.53	.55	.54	.49	.45	.43	.41	.38	.37	.36	.35	.33	.31	.30	.29	.29	.28	.27	.27	.26	.25	.24	.23	.23	.23
		.30	.30	.31	.34	.40	.44	.48	.49	.45	.44	.42	.40	.38	.36	.34	.33	.32	.30	.30	.29	.28	.27	.26	.26	.25	.25	.24	.24	.23	.23
	F	.34	.38	.43	.49	.56	.61	.63	.61	.57	.54	.52	.50	.48	.46	.46	.44	.43	.42	.41	.40	.39	.38	.38	.37	.36	.34	.33	.33	.32	.32
		.30	.29	.32	.38	.48	.58	.61	.60	.58	.54	.52	.49	.48	.46	.45	.43	.42	.41	.40	.39	.38	.37	.36	.36	.35	.35	.34	.33	.32	.32
		.31	.33	.35	.39	.44	.52	.57	.56	.53	.51	.49	.48	.47	.46	.44	.42	.41	.40	.40	.39	.39	.38	.38	.37	.36	.36	.35	.34	.33	.33
		.30	.32	.35	.41	.51	.58	.63	.63	.60	.57	.53	.49	.48	.46	.45	.43	.41	.40	.39	.37	.37	.36	.35	.34	.33	.32	.32	.30	.30	.30

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.12 Observed SE when 6 items showed nonuniform and uniform DIF both with magnitudes of 1.6 (Figure 44).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.33	.37	.41	.47	.53	.54	.60	.56	.53	.49	.48	.45	.44	.43	.42	.40	.38	.37	.36	.36	.35	.34	.34	.33	.32	.31	.31	.31	.30	.29
		.34	.39	.43	.47	.51	.55	.62	.56	.53	.48	.45	.42	.39	.37	.35	.32	.30	.30	.29	.29	.28	.28	.27	.27	.26	.26	.25	.25	.24	.24
		.35	.40	.42	.47	.52	.55	.60	.57	.52	.49	.45	.43	.41	.39	.36	.36	.34	.33	.32	.31	.30	.30	.29	.29	.29	.28	.27	.27	.26	.25
		.36	.42	.46	.48	.55	.61	.66	.57	.53	.49	.46	.44	.42	.40	.39	.38	.36	.35	.34	.33	.32	.31	.30	.29	.28	.28	.27	.27	.27	.26
	F	.40	.51	.63	.74	.84	.96	.98	.88	.80	.74	.70	.65	.62	.60	.56	.54	.52	.49	.47	.44	.42	.40	.39	.37	.36	.34	.33	.32	.31	.30
		.36	.42	.46	.50	.55	.57	.62	.55	.53	.49	.46	.46	.48	.49	.49	.49	.50	.48	.47	.45	.44	.42	.40	.39	.38	.37	.36	.35	.34	.33
		.34	.40	.45	.49	.53	.57	.62	.58	.52	.48	.44	.42	.40	.38	.37	.35	.34	.33	.32	.31	.30	.29	.28	.28	.28	.29	.29	.29	.29	.30
		.35	.45	.50	.59	.67	.69	.73	.64	.59	.55	.52	.50	.47	.44	.42	.40	.40	.39	.38	.38	.37	.35	.34	.34	.33	.33	.32	.31	.30	.29
θ = 2.5	R	.29	.29	.29	.29	.30	.28	.34	.39	.40	.39	.39	.39	.37	.36	.35	.34	.33	.32	.31	.30	.29	.29	.29	.28	.28	.27	.27	.26	.26	.26
		.30	.32	.34	.40	.46	.53	.56	.57	.53	.49	.46	.45	.42	.39	.37	.35	.33	.33	.33	.32	.31	.31	.30	.30	.29	.28	.28	.27	.27	.27
		.29	.29	.32	.36	.44	.51	.54	.53	.49	.46	.45	.42	.40	.39	.37	.36	.35	.35	.33	.33	.32	.32	.31	.30	.30	.30	.29	.28	.27	.27
		.30	.31	.33	.35	.41	.48	.53	.54	.51	.47	.43	.41	.39	.37	.36	.34	.33	.32	.31	.30	.30	.29	.28	.28	.27	.27	.26	.26	.25	.25
	F	.52	.66	.80	.91	.96	1.00	1.00	.97	.92	.87	.84	.80	.77	.73	.69	.67	.64	.61	.58	.56	.54	.51	.49	.47	.46	.44	.42	.41	.39	.38
		.31	.32	.35	.38	.45	.53	.57	.56	.52	.48	.45	.47	.49	.50	.51	.52	.53	.51	.49	.48	.46	.44	.43	.42	.41	.40	.39	.38	.37	.37
		.30	.31	.34	.38	.46	.54	.57	.56	.51	.47	.43	.41	.39	.38	.36	.34	.34	.33	.32	.31	.30	.29	.28	.28	.29	.30	.31	.31	.32	.32
		.30	.41	.48	.61	.67	.71	.71	.69	.63	.59	.55	.52	.49	.46	.42	.40	.40	.39	.38	.38	.37	.36	.35	.36	.35	.36	.35	.34	.33	.33

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.13 Observed SE when 24 items showed nonuniform and uniform DIF both with magnitudes of 1.6 (Figure 45).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.34	.39	.42	.45	.51	.52	.49	.40	.35	.33	.30	.28	.26	.24	.24	.23	.22	.21	.20	.19	.18	.18	.17	.17	.17	.18	.18	.19	.19	.20
		.34	.39	.41	.44	.52	.50	.48	.38	.34	.32	.29	.28	.27	.25	.24	.23	.22	.21	.21	.20	.20	.19	.18	.18	.17	.17	.17	.17	.18	.18
		.34	.40	.43	.47	.50	.56	.56	.46	.41	.37	.34	.31	.29	.27	.26	.24	.24	.23	.22	.21	.20	.20	.19	.19	.18	.17	.17	.16	.16	.16
		.35	.39	.42	.46	.50	.50	.47	.40	.34	.35	.32	.32	.30	.28	.26	.24	.25	.24	.22	.22	.22	.22	.21	.20	.20	.19	.18	.19	.18	.18
	F	.41	.52	.61	.71	.81	.90	1.00	.97	.95	.94	.93	.92	.89	.88	.87	.86	.84	.83	.81	.80	.79	.78	.77	.76	.74	.71	.68	.66	.63	.61
		.36	.43	.51	.62	.71	.82	.93	.90	.88	.87	.85	.83	.81	.80	.77	.75	.74	.73	.72	.71	.71	.70	.69	.67	.66	.65	.63	.61	.60	.58
		.38	.45	.50	.54	.56	.60	.71	.73	.71	.70	.69	.67	.66	.64	.63	.61	.61	.60	.59	.58	.57	.57	.56	.55	.55	.54	.54	.54	.54	.53
		.42	.55	.62	.72	.84	.95	1.05	1.03	.99	.93	.92	.86	.86	.85	.84	.83	.79	.78	.77	.76	.75	.72	.70	.69	.68	.67	.66	.64	.63	.63
θ = 2.5	R	.29	.29	.29	.29	.30	.29	.30	.30	.29	.27	.26	.24	.23	.22	.22	.21	.21	.20	.20	.20	.19	.19	.18	.18	.19	.19	.20	.20	.20	.21
		.30	.31	.32	.34	.37	.37	.37	.36	.32	.28	.26	.23	.22	.21	.20	.20	.20	.19	.19	.19	.18	.18	.17	.17	.17	.17	.18	.18	.19	.19
		.29	.29	.32	.35	.43	.50	.54	.53	.47	.43	.39	.36	.33	.30	.28	.26	.25	.25	.23	.22	.22	.22	.21	.20	.19	.19	.19	.18	.18	.17
		.30	.30	.32	.34	.37	.39	.38	.35	.31	.33	.30	.30	.29	.27	.26	.25	.27	.26	.25	.23	.22	.22	.22	.21	.20	.20	.19	.19	.19	.19
	F	.50	.66	.79	.89	.96	1.03	1.03	1.04	1.03	1.01	1.00	.99	.97	.96	.96	.95	.95	.94	.94	.93	.93	.92	.92	.90	.89	.87	.85	.84	.82	.80
		.29	.31	.40	.53	.63	.71	.76	.78	.79	.80	.79	.78	.78	.78	.77	.76	.75	.75	.75	.75	.74	.74	.74	.74	.73	.74	.72	.71	.69	.68
		.30	.31	.33	.37	.44	.50	.62	.67	.69	.69	.70	.71	.71	.71	.71	.72	.72	.72	.72	.72	.73	.73	.72	.72	.72	.72	.72	.72	.72	.72
		.55	.70	.79	.88	.93	.96	.97	.97	.97	.93	.92	.88	.88	.88	.88	.87	.84	.83	.82	.82	.82	.80	.79	.79	.78	.78	.77	.75	.75	.74

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.14 Observed SE when 6 items showed nonuniform and uniform DIF with magnitude of 1.6 and .4 respectively (Figure 46).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.34	.39	.42	.44	.45	.43	.52	.52	.48	.45	.42	.41	.39	.38	.37	.37	.35	.34	.33	.32	.31	.30	.30	.29	.28	.27	.27	.27	.26	.26
		.32	.37	.42	.47	.51	.55	.61	.58	.52	.49	.46	.41	.38	.35	.32	.30	.29	.29	.28	.27	.28	.27	.26	.25	.25	.25	.24	.23	.23	.23
		.34	.39	.42	.45	.51	.56	.61	.58	.54	.51	.48	.46	.44	.42	.40	.39	.37	.36	.34	.33	.33	.32	.30	.30	.28	.27	.26	.26	.25	.24
		.36	.43	.48	.52	.58	.62	.67	.61	.58	.54	.50	.46	.44	.40	.39	.37	.35	.34	.34	.32	.31	.30	.30	.28	.27	.26	.26	.26	.25	.25
	F	.45	.58	.72	.81	.91	1.02	1.04	.94	.87	.81	.75	.71	.65	.62	.58	.55	.53	.50	.47	.45	.43	.42	.39	.38	.36	.35	.34	.33	.32	.31
		.34	.39	.42	.47	.53	.58	.62	.58	.53	.50	.47	.47	.49	.48	.49	.50	.51	.48	.46	.45	.43	.42	.40	.39	.38	.37	.37	.35	.34	.34
		.37	.44	.47	.50	.55	.59	.64	.58	.54	.52	.50	.48	.45	.43	.41	.41	.39	.38	.37	.36	.35	.34	.33	.33	.33	.33	.34	.35	.35	.35
		.37	.50	.57	.68	.75	.80	.82	.75	.67	.62	.60	.56	.53	.51	.48	.45	.45	.44	.43	.42	.40	.39	.37	.38	.36	.37	.35	.35	.34	.34
θ = 2.5	R	.29	.29	.29	.30	.32	.30	.39	.42	.42	.42	.41	.40	.39	.37	.36	.35	.34	.33	.32	.32	.31	.30	.30	.29	.28	.27	.27	.26	.26	.25
		.31	.32	.34	.38	.45	.52	.57	.56	.52	.49	.46	.41	.38	.35	.32	.30	.28	.27	.27	.27	.27	.26	.26	.25	.25	.25	.24	.24	.23	.23
		.30	.31	.33	.39	.45	.51	.54	.53	.53	.52	.49	.46	.44	.41	.40	.38	.37	.35	.35	.34	.33	.32	.31	.31	.29	.28	.26	.25	.25	.24
		.30	.30	.32	.34	.40	.47	.52	.50	.48	.46	.43	.40	.40	.38	.36	.35	.33	.32	.31	.29	.29	.28	.27	.26	.26	.25	.25	.24	.23	.23
	F	.48	.63	.78	.87	.93	.99	.98	.95	.91	.87	.83	.79	.75	.72	.68	.65	.62	.60	.57	.55	.53	.51	.49	.47	.46	.45	.43	.41	.39	.38
		.29	.31	.34	.38	.46	.52	.54	.55	.52	.50	.47	.48	.49	.51	.52	.53	.53	.51	.49	.47	.45	.43	.42	.40	.40	.39	.37	.36	.36	.35
		.31	.33	.35	.40	.47	.53	.58	.57	.53	.49	.46	.44	.42	.41	.39	.37	.36	.35	.34	.33	.32	.31	.31	.30	.30	.31	.31	.32	.32	.33
		.29	.37	.43	.55	.60	.63	.65	.62	.59	.55	.53	.50	.48	.45	.43	.40	.41	.40	.40	.41	.39	.38	.37	.37	.37	.37	.36	.35	.35	.34

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.15 Observed SE when 24 items showed nonuniform and uniform DIF with magnitude of 1.6 and .4 respectively (Figure 47).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.34	.40	.43	.46	.47	.44	.48	.39	.34	.31	.29	.29	.26	.24	.24	.23	.22	.21	.20	.19	.19	.19	.18	.18	.18	.18	.18	.19	.18	.18
		.34	.41	.45	.49	.50	.49	.51	.43	.39	.34	.32	.30	.28	.26	.24	.23	.22	.21	.20	.20	.19	.18	.18	.17	.17	.16	.16	.16	.16	.16
		.33	.38	.41	.44	.49	.53	.52	.42	.38	.34	.31	.29	.27	.26	.24	.24	.23	.22	.22	.21	.20	.20	.20	.19	.19	.18	.18	.18	.17	.17
		.34	.39	.42	.45	.44	.42	.47	.38	.34	.33	.31	.31	.29	.27	.25	.25	.25	.24	.23	.22	.22	.22	.21	.21	.20	.20	.20	.20	.19	.18
	F	.45	.58	.71	.79	.90	1.02	1.08	1.04	1.02	1.00	1.01	1.00	.98	.97	.96	.94	.92	.91	.90	.90	.90	.89	.88	.87	.85	.82	.80	.77	.75	.72
		.36	.42	.49	.59	.70	.82	.89	.90	.87	.85	.84	.83	.83	.82	.81	.81	.79	.78	.78	.77	.76	.76	.75	.75	.74	.73	.71	.69	.66	.65
		.36	.42	.45	.50	.55	.57	.69	.72	.71	.70	.70	.70	.69	.68	.68	.67	.66	.65	.65	.64	.64	.63	.62	.62	.61	.61	.60	.60	.59	.59
		.43	.56	.64	.72	.83	.93	1.02	.97	.95	.87	.86	.80	.79	.78	.78	.77	.73	.72	.71	.70	.69	.66	.64	.64	.63	.63	.62	.61	.60	.59
θ = 2.5	R	.29	.29	.29	.29	.31	.30	.35	.39	.36	.32	.29	.28	.26	.24	.23	.22	.20	.20	.20	.19	.18	.17	.17	.17	.17	.18	.18	.18	.18	.19
		.30	.32	.33	.35	.39	.38	.40	.40	.37	.34	.31	.29	.27	.25	.24	.22	.21	.20	.19	.19	.18	.17	.17	.17	.16	.16	.17	.17	.17	.17
		.30	.31	.34	.37	.44	.51	.50	.46	.41	.36	.32	.29	.27	.25	.23	.22	.21	.20	.19	.19	.18	.17	.18	.17	.16	.16	.15	.15	.15	.15
		.29	.29	.30	.33	.37	.38	.39	.40	.36	.35	.31	.31	.28	.26	.24	.22	.23	.23	.22	.21	.20	.20	.19	.18	.18	.17	.17	.17	.17	.17
	F	.46	.59	.73	.85	.90	.95	.96	.97	.96	.95	.94	.93	.92	.90	.90	.89	.89	.87	.87	.87	.86	.85	.85	.85	.83	.81	.79	.77	.75	.74
		.29	.30	.40	.54	.67	.75	.79	.80	.81	.80	.81	.80	.80	.80	.79	.78	.78	.78	.77	.77	.77	.77	.76	.76	.76	.76	.74	.72	.70	.68
		.30	.32	.35	.38	.45	.52	.64	.68	.69	.69	.69	.69	.69	.69	.68	.68	.68	.68	.68	.69	.68	.69	.68	.68	.68	.68	.67	.67	.67	.67
		.47	.60	.68	.77	.85	.90	.94	.94	.94	.88	.87	.84	.83	.82	.81	.80	.77	.77	.76	.76	.76	.73	.73	.73	.72	.72	.71	.70	.69	.68

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.16 Observed SE when 6 items showed nonuniform and uniform DIF with magnitude of .4 and 1.6 respectively (Figure 48).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.37	.42	.45	.52	.59	.60	.64	.55	.53	.50	.49	.46	.43	.42	.40	.38	.37	.36	.35	.34	.33	.32	.31	.31	.30	.29	.28	.28	.28	.27
		.32	.39	.44	.47	.54	.60	.64	.58	.54	.50	.46	.43	.39	.37	.36	.34	.33	.32	.32	.32	.32	.31	.30	.29	.29	.28	.27	.27	.27	.27
		.37	.45	.48	.51	.55	.56	.61	.57	.52	.49	.46	.42	.40	.39	.37	.36	.34	.33	.31	.31	.30	.30	.29	.29	.28	.27	.27	.26	.25	.25
		.36	.44	.48	.53	.57	.61	.66	.59	.54	.49	.46	.43	.40	.39	.37	.36	.35	.33	.33	.32	.31	.30	.29	.29	.28	.27	.27	.27	.26	.26
	F	.35	.40	.43	.47	.51	.55	.60	.57	.53	.49	.46	.43	.41	.39	.37	.36	.34	.33	.32	.32	.31	.30	.30	.29	.28	.28	.27	.26	.26	.25
		.33	.40	.44	.48	.52	.56	.63	.59	.54	.50	.47	.47	.45	.44	.44	.42	.41	.40	.38	.37	.36	.35	.34	.33	.33	.32	.31	.31	.30	.29
		.34	.39	.42	.47	.51	.54	.59	.55	.51	.48	.46	.43	.42	.40	.39	.37	.36	.34	.33	.32	.31	.31	.30	.29	.29	.28	.28	.28	.27	.27
		.38	.46	.50	.53	.56	.57	.63	.56	.51	.49	.46	.44	.41	.39	.37	.37	.36	.35	.34	.33	.33	.32	.30	.30	.29	.29	.29	.29	.28	.27
θ = 2.5	R	.28	.28	.28	.29	.30	.33	.38	.43	.44	.43	.41	.40	.39	.37	.36	.35	.34	.33	.32	.32	.30	.30	.29	.28	.27	.27	.26	.26	.25	.25
		.30	.31	.35	.38	.46	.52	.55	.54	.51	.47	.46	.44	.43	.42	.40	.38	.37	.36	.35	.35	.34	.33	.33	.31	.31	.30	.30	.29	.29	.28
		.31	.33	.34	.40	.47	.52	.56	.54	.51	.47	.44	.42	.39	.38	.37	.36	.35	.34	.33	.32	.31	.30	.30	.29	.29	.28	.28	.28	.27	.27
		.31	.31	.33	.36	.41	.47	.51	.51	.47	.45	.43	.42	.41	.40	.38	.36	.35	.34	.32	.31	.31	.30	.29	.28	.28	.27	.27	.27	.26	.26
	F	.37	.44	.52	.60	.64	.67	.67	.64	.61	.57	.54	.51	.49	.46	.44	.42	.41	.39	.38	.37	.35	.34	.33	.32	.31	.31	.30	.29	.29	.28
		.30	.32	.34	.38	.46	.53	.58	.56	.53	.49	.46	.44	.43	.42	.41	.40	.39	.37	.36	.35	.35	.33	.32	.32	.31	.31	.30	.29	.29	.28
		.28	.29	.31	.36	.42	.50	.55	.54	.49	.46	.43	.41	.39	.36	.35	.34	.32	.31	.30	.30	.29	.29	.28	.28	.28	.28	.27	.27	.27	.27
		.29	.33	.37	.45	.53	.59	.62	.61	.56	.53	.51	.48	.46	.45	.42	.40	.39	.38	.37	.37	.36	.34	.33	.33	.32	.31	.31	.30	.30	.29

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

G.17 Observed SE when 24 items showed nonuniform and uniform DIF with magnitude of .4 and 1.6 respectively (Figure 49).

	Items	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ = -2.5	R	.35	.41	.45	.51	.56	.59	.59	.50	.47	.44	.40	.37	.35	.33	.31	.30	.29	.27	.27	.26	.25	.24	.24	.23	.23	.23	.23	.22	.23	.23
		.33	.37	.40	.44	.51	.55	.54	.48	.44	.41	.39	.36	.35	.33	.32	.31	.30	.29	.28	.27	.26	.26	.25	.25	.24	.23	.23	.23	.23	.23
		.32	.37	.42	.49	.54	.59	.59	.51	.46	.44	.40	.38	.36	.34	.33	.32	.30	.29	.28	.27	.26	.26	.25	.24	.23	.23	.22	.22	.21	.21
		.35	.42	.48	.53	.58	.58	.57	.49	.45	.43	.39	.39	.37	.35	.34	.32	.32	.30	.30	.29	.29	.28	.27	.26	.26	.25	.24	.24	.23	.23
	F	.34	.38	.41	.44	.48	.52	.55	.55	.50	.46	.44	.41	.39	.37	.35	.34	.34	.33	.32	.31	.30	.29	.28	.28	.28	.28	.29	.28	.29	.29
		.36	.41	.44	.47	.50	.55	.58	.56	.52	.48	.45	.43	.40	.39	.37	.36	.34	.33	.32	.31	.31	.30	.29	.28	.28	.27	.27	.27	.27	.27
		.33	.36	.40	.44	.49	.55	.62	.61	.55	.51	.48	.46	.43	.40	.39	.37	.36	.35	.34	.33	.32	.31	.30	.29	.28	.28	.27	.27	.26	.25
		.33	.37	.40	.45	.49	.55	.62	.60	.53	.50	.46	.43	.42	.40	.38	.36	.36	.34	.33	.33	.32	.31	.30	.30	.29	.29	.28	.28	.27	.27
θ = 2.5	R	.29	.29	.29	.30	.32	.34	.37	.39	.36	.34	.32	.31	.29	.28	.28	.27	.26	.25	.25	.24	.24	.23	.23	.22	.23	.23	.23	.24	.24	.24
		.31	.32	.33	.35	.38	.41	.44	.42	.38	.35	.33	.32	.30	.29	.28	.26	.26	.25	.25	.24	.23	.23	.22	.22	.22	.21	.22	.23	.23	.24
		.31	.32	.35	.40	.47	.52	.56	.55	.51	.47	.43	.41	.39	.36	.35	.33	.31	.30	.28	.27	.27	.26	.26	.25	.24	.24	.23	.23	.22	.22
		.28	.28	.31	.33	.38	.43	.45	.43	.38	.38	.36	.36	.33	.33	.30	.29	.29	.28	.28	.26	.26	.26	.25	.24	.24	.23	.23	.23	.23	.22
	F	.37	.43	.52	.60	.65	.68	.67	.62	.59	.57	.54	.51	.49	.47	.46	.44	.43	.41	.40	.39	.38	.37	.36	.36	.35	.34	.34	.33	.33	.32
		.30	.30	.37	.45	.54	.59	.59	.56	.51	.49	.47	.45	.45	.43	.41	.40	.39	.38	.37	.37	.36	.35	.34	.34	.33	.33	.33	.32	.32	.31
		.30	.32	.35	.38	.46	.54	.59	.56	.52	.49	.47	.45	.43	.41	.40	.39	.39	.38	.37	.37	.36	.35	.34	.34	.33	.32	.32	.32	.31	.31
		.34	.40	.46	.54	.60	.64	.65	.59	.56	.54	.52	.50	.48	.45	.44	.43	.42	.40	.39	.38	.37	.37	.36	.35	.35	.34	.34	.33	.33	.32

Note: The observed SEs obtained after each of the 30 operational items was answered are reported for θ = ±2.5 from reference (R) and focal (F) groups. There are four rows for each θ from each group, representing four levels of DIF occurrence or the stages of CAT that DIF items were administered (i.e., first stages, middle stages, last stages, and across stages of CAT). In each row, the observed SEs at DIF items were also highlighted. The rows without highlighted cells represent the case that DIF items occurred across stages of CAT.

Appendix H: Type I error and power of detecting DIF in pretest items.

H.1 Type I Error of detecting DIF in pretest items when the operational test was DIF-free (Figure 50).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00
2	.00	.01	.01	.00	.01	.00	.01	.01	.01	.00	.01	.01	.01	.00	.00	.01	.00	.01
3	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.01	.01	.00	.00	.01	.01	.01	.01
4	.01	.00	.01	.01	.02	.00	.00	.00	.00	.00	.01	.01	.00	.01	.00	.01	.01	.01
5	.00	.01	.01	.00	.00	.00	.01	.00	.01	.02	.00	.00	.02	.02	.02	.00	.00	.00
6	.00	.01	.01	.00	.01	.00	.00	.01	.00	.01	.00	.00	.01	.01	.00	.00	.00	.00
7	.01	.01	.01	.00	.00	.00	.01	.01	.01	.00	.01	.01	.00	.01	.00	.01	.01	.01
8	.00	.00	.01	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.01	.01
9	.00	.01	.00	.00	.00	.00	.00	.00	.00	.01	.01	.00	.01	.01	.01	.00	.00	.00
10	.00	.00	.00	.01	.00	.00	.00	.00	.00	.01	.01	.01	.01	.01	.01	.01	.01	.00
11	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.01	.01	.01	.00	.00	.00
12	.01	.01	.00	.01	.01	.01	.01	.01	.01	.01	.00	.00	.01	.01	.01	.00	.00	.00
13	.03	.01	.01	.01	.03	.01	.00	.00	.00	.01	.00	.00	.01	.02	.01	.00	.00	.00
14	.00	.01	.01	.00	.00	.00	.00	.00	.01	.03	.01	.00	.03	.03	.00	.00	.00	.00
15	.01	.01	.01	.01	.01	.00	.01	.01	.01	.00	.01	.01	.01	.01	.01	.01	.01	.01
16	.01	.00	.00	.01	.02	.02	.00	.00	.00	.01	.00	.00	.01	.01	.01	.00	.00	.00
$\bar{X}$	.00	.00	.00	.00	.01	.00	.00	.00	.00	.01	.00	.00	.01	.01	.00	.00	.00	.00
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.02	.11	.00	.02	.01	.01	.07	.03	.01	.00	.04	.01	.01	.01	.02	.05	.02	.01
2	.00	.25	.01	.00	.01	.00	.26	.11	.01	.01	.05	.01	.00	.00	.00	.08	.03	.01
3	.03	.17	.01	.02	.02	.01	.29	.15	.00	.01	.04	.00	.00	.04	.02	.08	.03	.00
4	.01	.04	.00	.00	.02	.02	.14	.07	.01	.01	.01	.01	.01	.03	.03	.05	.01	.01
5	.01	.02	.00	.03	.04	.06	.14	.07	.01	.01	.00	.01	.02	.03	.03	.04	.04	.00
6	.08	.02	.08	.08	.14	.18	.04	.01	.00	.06	.01	.04	.06	.07	.09	.01	.01	.01
7	.14	.07	.23	.15	.26	.35	.02	.02	.01	.15	.02	.07	.12	.15	.18	.01	.01	.01
8	.01	.21	.01	.01	.00	.00	.10	.05	.01	.00	.09	.01	.00	.01	.02	.06	.02	.00
9	.03	.30	.01	.01	.02	.01	.29	.11	.00	.01	.15	.01	.00	.04	.01	.16	.08	.02
10	.03	.41	.00	.01	.02	.01	.54	.29	.01	.02	.11	.01	.01	.03	.01	.19	.07	.02
11	.02	.15	.00	.01	.04	.04	.45	.28	.03	.02	.04	.00	.03	.03	.04	.14	.07	.01
12	.04	.01	.06	.06	.14	.25	.16	.09	.00	.05	.01	.02	.06	.09	.14	.07	.05	.00
13	.25	.05	.31	.27	.40	.58	.05	.04	.01	.20	.02	.08	.19	.24	.29	.02	.01	.00
14	.45	.24	.55	.45	.54	.67	.00	.00	.01	.27	.07	.20	.24	.26	.37	.01	.01	.00
15	.01	.46	.02	.02	.09	.18	.85	.63	.08	.01	.17	.01	.02	.10	.14	.38	.21	.03
16	.88	.77	.94	.85	.88	.95	.02	.03	.01	.56	.27	.59	.52	.57	.64	.01	.01	.01
$\bar{X}$	.12	.20	.14	.12	.16	.20	.21	.12	.01	.08	.07	.06	.08	.10	.12	.08	.04	.01

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.2 Type I Error of detecting DIF in pretest items when the operational test consisted of 6 items exhibiting uniform DIF with a magnitude of .4 at the beginning of the test (Figure 51).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.00	.01	.01	.00	.00	.01	.00	.00	.02	.00	.01	.01	.00	.00	.00	.01	.01	.02
2	.00	.00	.02	.01	.00	.00	.02	.02	.03	.00	.02	.00	.00	.00	.01	.01	.01	.01
3	.01	.00	.02	.01	.01	.01	.00	.00	.02	.01	.01	.00	.01	.01	.00	.01	.01	.00
4	.00	.01	.03	.00	.01	.00	.01	.01	.03	.00	.00	.02	.00	.01	.01	.00	.01	.01
5	.00	.01	.01	.00	.01	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.01	.00
6	.00	.02	.04	.00	.00	.00	.02	.02	.03	.02	.00	.01	.02	.02	.02	.00	.00	.00
7	.00	.00	.00	.00	.00	.01	.00	.00	.01	.01	.00	.00	.01	.01	.01	.00	.00	.00
8	.01	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
9	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.01	.01	.00
10	.00	.00	.02	.00	.01	.01	.00	.00	.03	.00	.00	.00	.00	.00	.00	.01	.00	.00
11	.01	.00	.04	.01	.02	.02	.02	.02	.04	.01	.02	.00	.01	.03	.01	.00	.00	.00
12	.01	.00	.02	.01	.01	.00	.00	.00	.02	.00	.01	.02	.01	.02	.02	.02	.01	.02
13	.01	.00	.01	.01	.01	.01	.00	.00	.00	.02	.00	.00	.01	.02	.00	.00	.00	.01
14	.01	.00	.01	.01	.02	.01	.00	.00	.00	.00	.01	.01	.00	.00	.00	.01	.01	.01
15	.02	.01	.02	.02	.01	.01	.01	.01	.02	.00	.00	.00	.00	.00	.01	.00	.00	.00
16	.03	.02	.02	.03	.04	.01	.02	.02	.02	.02	.02	.01	.02	.02	.01	.01	.01	.00
$\bar{X}$	.01	.00	.02	.01	.01	.00	.01	.01	.02	.00	.00	.00	.00	.01	.00	.00	.00	.00
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.03	.11	.01	.01	.00	.00	.08	.03	.00	.00	.06	.01	.00	.01	.01	.04	.04	.00
2	.01	.16	.00	.01	.01	.00	.15	.11	.02	.01	.05	.00	.01	.01	.01	.05	.05	.00
3	.02	.16	.01	.01	.02	.01	.22	.14	.00	.01	.07	.00	.01	.01	.00	.08	.05	.00
4	.06	.09	.00	.04	.01	.03	.14	.13	.00	.02	.01	.01	.02	.02	.00	.04	.03	.00
5	.01	.00	.06	.02	.05	.04	.04	.03	.01	.03	.01	.02	.04	.06	.06	.03	.02	.01
6	.07	.01	.20	.08	.11	.22	.02	.02	.01	.06	.01	.09	.07	.07	.08	.01	.01	.01
7	.13	.02	.32	.13	.20	.33	.00	.00	.00	.13	.02	.15	.11	.11	.14	.00	.00	.01
8	.02	.14	.00	.02	.01	.00	.08	.05	.00	.01	.07	.00	.00	.00	.00	.03	.02	.00
9	.03	.39	.00	.01	.01	.00	.29	.16	.00	.00	.16	.00	.00	.02	.01	.11	.07	.00
10	.01	.39	.00	.00	.03	.01	.41	.32	.00	.01	.12	.00	.01	.03	.02	.14	.12	.01
11	.00	.16	.01	.00	.07	.04	.30	.21	.00	.01	.04	.01	.01	.04	.03	.09	.07	.00
12	.09	.01	.21	.10	.17	.31	.10	.08	.02	.06	.01	.05	.06	.12	.15	.05	.04	.00
13	.32	.09	.63	.31	.41	.71	.02	.03	.00	.20	.04	.27	.21	.28	.33	.01	.01	.00
14	.51	.26	.78	.48	.56	.76	.02	.01	.00	.32	.07	.27	.30	.31	.38	.01	.02	.00
15	.02	.48	.02	.03	.11	.25	.67	.62	.00	.02	.16	.00	.05	.07	.16	.24	.18	.01
16	.87	.69	.96	.85	.90	.98	.00	.00	.01	.64	.31	.73	.55	.60	.74	.01	.01	.00
$\bar{X}$	.14	.20	.20	.13	.17	.23	.16	.12	.00	.09	.07	.10	.09	.11	.13	.06	.04	.00

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.3 Type I Error of detecting DIF in pretest items when the operational test consisted of 6 items exhibiting nonuniform DIF with a magnitude of .4 at the beginning of the test (Figure 52).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.00	.00	.00	.00	.00	.02	.01	.01	.00	.00	.01	.00	.00	.00	.00	.00	.00	.02
2	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.02	.00	.00	.00	.00
3	.02	.01	.01	.01	.00	.00	.01	.01	.01	.00	.00	.01	.00	.00	.00	.01	.01	.01
4	.01	.01	.00	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
5	.00	.00	.00	.00	.00	.00	.01	.01	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00
6	.01	.00	.00	.01	.01	.00	.00	.00	.00	.01	.00	.00	.00	.00	.01	.00	.00	.00
7	.02	.01	.01	.01	.01	.01	.00	.00	.00	.01	.01	.01	.00	.01	.00	.01	.01	.01
8	.01	.00	.01	.01	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
9	.01	.00	.00	.01	.01	.00	.01	.01	.01	.00	.01	.00	.00	.00	.00	.00	.00	.00
10	.00	.00	.00	.00	.00	.00	.00	.00	.00	.02	.00	.01	.02	.01	.01	.00	.00	.01
11	.00	.00	.00	.01	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.01	.00	.00	.01
12	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.01	.01	.01	.01	.00	.01	.01	.01
13	.00	.00	.00	.00	.00	.00	.00	.00	.01	.01	.01	.01	.01	.02	.00	.00	.00	.00
14	.00	.00	.00	.00	.01	.02	.00	.00	.00	.01	.01	.00	.02	.02	.02	.01	.00	.00
15	.00	.00	.01	.00	.01	.00	.00	.01	.00	.01	.01	.00	.00	.01	.01	.01	.00	.01
16	.03	.01	.01	.03	.03	.02	.01	.00	.00	.01	.01	.01	.01	.00	.00	.01	.01	.01
$\bar{X}$	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.00	.00	.00	.00
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.02	.14	.01	.01	.00	.00	.15	.07	.00	.00	.03	.01	.00	.01	.00	.06	.02	.00
2	.03	.15	.01	.03	.01	.00	.19	.11	.01	.02	.09	.01	.00	.01	.00	.10	.05	.01
3	.03	.21	.00	.01	.01	.01	.30	.17	.03	.00	.06	.01	.01	.03	.01	.12	.06	.01
4	.01	.04	.00	.00	.05	.00	.19	.07	.01	.01	.03	.00	.01	.04	.03	.08	.05	.01
5	.03	.02	.01	.04	.05	.08	.11	.06	.01	.02	.00	.00	.03	.06	.04	.03	.02	.00
6	.05	.00	.06	.08	.07	.14	.03	.02	.02	.06	.01	.01	.05	.06	.09	.03	.02	.01
7	.12	.06	.16	.11	.17	.22	.02	.03	.02	.12	.02	.07	.12	.17	.18	.01	.00	.01
8	.02	.22	.00	.01	.01	.01	.12	.04	.00	.00	.12	.01	.00	.01	.01	.12	.06	.01
9	.02	.47	.03	.02	.03	.01	.38	.22	.02	.00	.13	.01	.00	.03	.00	.17	.06	.01
10	.03	.40	.02	.01	.01	.02	.50	.32	.01	.01	.10	.01	.01	.04	.01	.17	.09	.01
11	.02	.15	.01	.02	.04	.05	.42	.25	.03	.02	.04	.00	.01	.02	.02	.15	.08	.00
12	.04	.01	.04	.06	.13	.19	.17	.09	.00	.06	.01	.03	.07	.10	.16	.04	.03	.01
13	.28	.08	.40	.30	.38	.56	.04	.02	.00	.15	.03	.12	.14	.22	.28	.03	.02	.00
14	.52	.28	.62	.53	.61	.73	.01	.01	.00	.25	.09	.25	.24	.31	.32	.01	.01	.01
15	.01	.48	.01	.02	.11	.22	.78	.61	.06	.02	.15	.01	.06	.12	.18	.32	.21	.02
16	.84	.70	.94	.85	.89	.98	.01	.02	.00	.58	.32	.63	.53	.60	.73	.01	.01	.01
$\bar{X}$	.13	.21	.14	.13	.16	.20	.21	.13	.01	.08	.07	.07	.08	.11	.13	.09	.05	.01

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.4 Type I Error of detecting DIF in pretest items when the operational test consisted of 6 items exhibiting both nonuniform and uniform DIF with magnitudes of .4 at the beginning of the test (Figure 53).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.00	.00	.01	.00	.01	.01	.00	.00	.00	.00	.01	.01	.00	.00	.00	.01	.01	.01
2	.02	.01	.01	.03	.03	.01	.01	.01	.02	.00	.00	.00	.00	.00	.00	.00	.00	.00
3	.00	.00	.02	.00	.00	.01	.01	.00	.02	.00	.01	.01	.01	.00	.01	.00	.00	.01
4	.01	.00	.02	.00	.01	.00	.00	.00	.02	.01	.00	.00	.01	.01	.01	.00	.00	.00
5	.01	.01	.03	.00	.00	.00	.00	.00	.01	.00	.01	.02	.00	.00	.01	.01	.00	.02
6	.00	.01	.01	.00	.00	.00	.02	.01	.01	.00	.00	.00	.00	.00	.00	.00	.00	.00
7	.01	.00	.01	.01	.01	.00	.00	.00	.00	.01	.00	.00	.00	.01	.00	.00	.00	.00
8	.00	.00	.02	.00	.01	.01	.00	.00	.01	.00	.01	.00	.00	.00	.00	.00	.00	.00
9	.00	.01	.00	.00	.00	.00	.01	.00	.02	.01	.00	.00	.00	.01	.00	.00	.00	.00
10	.00	.00	.03	.01	.01	.00	.01	.01	.02	.02	.01	.01	.01	.01	.01	.02	.01	.01
11	.01	.00	.01	.01	.01	.01	.01	.00	.01	.00	.00	.01	.00	.00	.00	.00	.00	.01
12	.01	.00	.02	.01	.01	.01	.01	.00	.02	.00	.00	.01	.00	.01	.00	.00	.00	.01
13	.03	.00	.01	.02	.01	.01	.00	.00	.01	.01	.00	.00	.01	.01	.00	.00	.00	.01
14	.00	.01	.01	.00	.00	.00	.01	.00	.01	.02	.00	.00	.00	.01	.01	.00	.00	.00
15	.01	.00	.03	.01	.01	.01	.01	.00	.04	.02	.00	.02	.02	.02	.00	.01	.01	.02
16	.01	.00	.01	.01	.01	.01	.00	.00	.00	.01	.00	.00	.01	.01	.01	.00	.00	.00
$\bar{X}$	.01	.00	.01	.00	.01	.00	.01	.00	.01	.01	.00	.00	.00	.01	.00	.00	.00	.01
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.02	.10	.00	.00	.00	.00	.09	.04	.00	.00	.04	.01	.00	.01	.00	.04	.04	.00
2	.02	.18	.00	.02	.02	.00	.20	.13	.02	.01	.09	.00	.01	.00	.01	.08	.04	.00
3	.02	.13	.00	.01	.01	.00	.17	.09	.00	.02	.05	.00	.01	.02	.02	.07	.05	.00
4	.01	.07	.02	.01	.01	.01	.15	.11	.01	.01	.01	.02	.01	.02	.02	.04	.03	.00
5	.02	.01	.05	.02	.05	.08	.05	.05	.00	.02	.01	.02	.02	.04	.05	.03	.02	.00
6	.07	.00	.16	.07	.09	.19	.03	.06	.00	.07	.01	.08	.06	.06	.08	.01	.01	.01
7	.17	.04	.28	.17	.21	.30	.00	.00	.01	.11	.02	.09	.11	.12	.14	.01	.01	.00
8	.01	.15	.00	.01	.01	.00	.10	.05	.00	.00	.09	.00	.00	.01	.00	.04	.02	.00
9	.01	.44	.02	.01	.01	.00	.28	.20	.01	.01	.15	.00	.01	.02	.00	.13	.09	.00
10	.03	.45	.00	.01	.02	.01	.45	.37	.01	.01	.15	.00	.01	.01	.00	.18	.13	.00
11	.01	.23	.00	.00	.03	.06	.41	.33	.00	.03	.04	.00	.04	.05	.03	.09	.07	.00
12	.02	.01	.14	.05	.12	.28	.11	.14	.01	.08	.01	.06	.09	.14	.13	.02	.02	.01
13	.26	.08	.57	.27	.36	.59	.01	.03	.00	.17	.02	.26	.14	.17	.26	.02	.02	.01
14	.54	.26	.76	.50	.60	.75	.00	.00	.00	.31	.07	.32	.26	.26	.37	.01	.01	.00
15	.02	.46	.01	.03	.09	.22	.68	.64	.00	.02	.13	.01	.03	.06	.09	.26	.19	.00
16	.84	.69	.97	.81	.91	.97	.00	.01	.01	.60	.32	.70	.57	.58	.70	.01	.01	.00
$\bar{X}$	.13	.20	.19	.12	.16	.21	.17	.14	.00	.09	.07	.10	.08	.10	.12	.06	.05	.00

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.5 Type I Error of detecting DIF in pretest items when the operational test consisted of 24 items exhibiting uniform DIF with a magnitude of 1.6 at the end of the test (Figure 54).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.04	.00	1.00	.00	.02	.00	.75	.01	1.00	.03	.01	.69	.00	.00	.00	.18	.00	1.00
2	.06	.00	1.00	.01	.02	.01	.90	.00	1.00	.02	.00	.97	.01	.02	.01	.39	.01	1.00
3	.09	.00	1.00	.01	.02	.00	.94	.00	1.00	.03	.00	1.00	.00	.01	.00	.45	.01	1.00
4	.05	.01	1.00	.01	.01	.00	.96	.01	1.00	.02	.01	1.00	.01	.00	.00	.36	.01	1.00
5	.04	.01	1.00	.00	.01	.01	.91	.01	1.00	.01	.01	1.00	.00	.00	.00	.31	.01	.99
6	.03	.02	1.00	.01	.01	.00	.71	.01	1.00	.02	.00	.97	.00	.01	.00	.16	.00	.89
7	.03	.00	1.00	.01	.01	.00	.42	.00	1.00	.01	.01	.68	.01	.02	.01	.06	.00	.43
8	.06	.01	1.00	.00	.02	.01	.78	.01	1.00	.01	.00	.45	.00	.01	.00	.23	.00	1.00
9	.06	.00	1.00	.00	.03	.00	.95	.00	1.00	.03	.00	.99	.01	.02	.00	.56	.00	1.00
10	.09	.01	1.00	.00	.03	.01	.99	.01	1.00	.06	.00	1.00	.01	.03	.00	.62	.00	1.00
11	.10	.01	1.00	.02	.02	.01	.99	.00	1.00	.04	.01	1.00	.01	.02	.00	.66	.01	1.00
12	.06	.00	1.00	.01	.01	.01	.98	.00	1.00	.04	.01	1.00	.02	.02	.00	.46	.01	1.00
13	.03	.00	1.00	.01	.00	.00	.79	.00	1.00	.04	.00	.98	.01	.02	.01	.17	.00	.88
14	.03	.01	.99	.00	.02	.00	.33	.00	1.00	.01	.00	.64	.01	.01	.01	.07	.00	.36
15	.11	.00	1.00	.00	.05	.01	1.00	.01	1.00	.03	.00	1.00	.01	.01	.01	.84	.00	1.00
16	.02	.01	.99	.01	.01	.00	.27	.00	.98	.04	.00	.56	.02	.03	.01	.04	.00	.19
$\bar{X}$	.05	.00	1.00	.00	.02	.00	.79	.00	1.00	.03	.00	.87	.01	.01	.00	.35	.00	.86
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.00	.08	.86	.01	.02	.01	.00	.02	.66	.01	.03	.42	.00	.02	.00	.01	.01	.46
2	.02	.13	.98	.02	.04	.01	.02	.06	.96	.02	.06	.85	.01	.05	.01	.01	.06	.70
3	.01	.13	1.00	.00	.04	.01	.12	.15	1.00	.01	.04	.99	.01	.03	.01	.01	.03	.93
4	.03	.04	1.00	.02	.05	.02	.19	.09	1.00	.02	.03	1.00	.01	.03	.01	.04	.03	.93
5	.10	.02	1.00	.02	.05	.07	.37	.05	1.00	.13	.01	1.00	.07	.07	.04	.03	.03	.92
6	.25	.01	1.00	.11	.18	.17	.45	.02	1.00	.18	.00	1.00	.10	.13	.10	.01	.01	.78
7	.33	.07	1.00	.21	.24	.33	.38	.00	1.00	.23	.03	.99	.14	.15	.13	.03	.01	.55
8	.02	.15	.82	.01	.02	.01	.01	.04	.60	.02	.13	.32	.00	.03	.01	.00	.04	.49
9	.01	.41	.99	.01	.04	.01	.01	.15	.96	.04	.14	.88	.01	.06	.00	.00	.04	.89
10	.01	.37	1.00	.01	.06	.01	.06	.30	1.00	.03	.10	1.00	.01	.08	.01	.02	.11	.97
11	.04	.16	1.00	.03	.07	.02	.25	.25	1.00	.06	.05	1.00	.03	.08	.05	.05	.07	1.00
12	.20	.01	1.00	.08	.22	.28	.61	.09	1.00	.15	.00	1.00	.05	.17	.12	.08	.06	.98
13	.55	.08	1.00	.35	.44	.57	.68	.04	1.00	.45	.04	1.00	.26	.30	.32	.10	.01	.94
14	.75	.29	1.00	.51	.63	.78	.54	.01	1.00	.51	.09	1.00	.32	.33	.41	.05	.01	.57
15	.09	.38	1.00	.05	.23	.21	.24	.59	1.00	.15	.13	1.00	.09	.20	.19	.06	.19	1.00
16	.97	.78	1.00	.88	.95	.98	.72	.00	1.00	.78	.33	1.00	.56	.65	.76	.07	.01	.56
$\bar{X}$	.21	.19	.98	.14	.20	.22	.29	.11	.95	.17	.07	.90	.10	.15	.13	.03	.04	.79

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.6 Type I Error of detecting DIF in pretest items when the operational test consisted of 24 items exhibiting nonuniform DIF with a magnitude of 1.6 at the end of the test (Figure 55).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.02	.02	.01	.04	.10	.10	.14	.01	.02	.01	.01	.01	.01	.05	.04	.05	.01	.00
2	.04	.03	.00	.04	.25	.11	.20	.00	.00	.02	.00	.01	.03	.09	.08	.08	.01	.00
3	.06	.04	.01	.03	.37	.13	.25	.01	.02	.03	.01	.00	.02	.16	.07	.08	.00	.01
4	.10	.01	.00	.08	.54	.21	.26	.00	.00	.02	.00	.00	.01	.20	.06	.10	.01	.01
5	.06	.01	.01	.05	.46	.12	.29	.01	.01	.04	.00	.01	.03	.21	.07	.08	.00	.01
6	.02	.00	.02	.01	.30	.04	.24	.01	.00	.03	.02	.01	.02	.13	.04	.07	.02	.01
7	.01	.00	.01	.01	.13	.05	.19	.02	.01	.00	.00	.00	.00	.04	.03	.03	.01	.00
8	.03	.02	.02	.03	.15	.15	.20	.01	.00	.01	.02	.01	.02	.04	.09	.08	.01	.01
9	.11	.06	.03	.10	.37	.30	.26	.00	.01	.01	.02	.02	.02	.12	.09	.10	.01	.01
10	.14	.04	.00	.10	.55	.38	.29	.00	.03	.02	.03	.01	.02	.21	.07	.14	.01	.01
11	.13	.02	.00	.12	.63	.31	.33	.01	.01	.04	.01	.00	.04	.33	.13	.15	.01	.01
12	.13	.00	.03	.05	.56	.23	.41	.02	.03	.02	.01	.01	.01	.26	.05	.16	.02	.01
13	.04	.00	.04	.03	.36	.11	.34	.03	.01	.02	.00	.00	.02	.14	.05	.17	.00	.01
14	.00	.01	.02	.00	.11	.02	.19	.03	.02	.01	.01	.01	.01	.04	.01	.05	.01	.01
15	.30	.03	.01	.29	.79	.58	.45	.01	.05	.09	.02	.00	.08	.50	.20	.25	.02	.01
16	.03	.00	.01	.02	.18	.06	.22	.01	.01	.01	.01	.01	.01	.04	.01	.06	.03	.02
$\bar{X}$	.07	.02	.01	.06	.36	.18	.26	.01	.01	.02	.01	.01	.02	.16	.07	.10	.01	.01
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.01	.38	.02	.00	.03	.02	.43	.34	.02	.01	.13	.01	.00	.04	.02	.27	.07	.00
2	.02	.47	.01	.00	.05	.03	.50	.53	.02	.01	.20	.01	.01	.12	.05	.34	.12	.02
3	.03	.47	.00	.01	.11	.07	.44	.67	.01	.00	.17	.01	.01	.21	.09	.30	.16	.01
4	.03	.19	.01	.01	.33	.19	.33	.63	.01	.02	.06	.00	.02	.30	.10	.17	.16	.00
5	.03	.04	.10	.01	.46	.35	.26	.53	.01	.02	.01	.04	.01	.38	.19	.11	.10	.01
6	.12	.00	.29	.12	.62	.53	.21	.25	.01	.08	.01	.08	.05	.35	.23	.04	.12	.01
7	.14	.07	.46	.10	.71	.68	.19	.19	.02	.07	.03	.16	.04	.41	.29	.03	.04	.00
8	.01	.60	.04	.02	.08	.07	.61	.42	.01	.00	.30	.04	.02	.07	.05	.50	.10	.02
9	.04	.80	.06	.02	.11	.10	.65	.74	.08	.02	.38	.04	.01	.22	.09	.51	.20	.02
10	.02	.80	.00	.01	.27	.20	.62	.89	.08	.01	.33	.01	.01	.31	.22	.40	.24	.00
11	.01	.38	.02	.04	.57	.52	.49	.88	.05	.02	.15	.00	.02	.49	.31	.23	.21	.00
12	.12	.01	.34	.09	.79	.80	.31	.67	.01	.11	.01	.07	.11	.61	.49	.11	.16	.01
13	.31	.06	.81	.20	.93	.95	.33	.40	.00	.17	.02	.41	.13	.72	.58	.07	.08	.01
14	.41	.20	.85	.32	.97	.94	.28	.19	.01	.22	.07	.44	.16	.73	.54	.02	.04	.00
15	.08	.74	.01	.13	.85	.93	.62	.97	.12	.10	.27	.01	.16	.79	.77	.36	.40	.02
16	.74	.67	1.00	.63	1.00	1.00	.39	.16	.03	.45	.30	.82	.32	.88	.86	.08	.07	.01
$\bar{X}$	.13	.37	.25	.11	.49	.46	.41	.53	.03	.08	.15	.13	.07	.41	.30	.22	.14	.01

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.7 Type I Error of detecting DIF in pretest items when the operational test consisted of 24 items exhibiting both nonuniform and uniform DIF with magnitudes of 1.6 at the end of the test (Figure 56).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.03	.00	1.00	.02	.12	.12	.14	.00	1.00	.00	.02	.40	.00	.05	.04	.04	.01	.97
2	.07	.02	1.00	.03	.22	.18	.27	.00	1.00	.03	.01	.93	.03	.11	.05	.07	.00	1.00
3	.08	.03	1.00	.04	.39	.25	.32	.01	1.00	.03	.03	.98	.02	.13	.06	.13	.01	1.00
4	.09	.02	1.00	.05	.41	.19	.39	.01	1.00	.07	.01	1.00	.06	.26	.09	.15	.02	1.00
5	.06	.00	1.00	.03	.36	.16	.44	.01	1.00	.04	.01	.99	.02	.20	.03	.16	.02	.97
6	.06	.00	1.00	.01	.25	.06	.39	.01	1.00	.01	.01	.93	.01	.11	.03	.12	.01	.80
7	.02	.00	1.00	.00	.09	.04	.26	.01	.97	.02	.00	.63	.01	.07	.03	.05	.01	.39
8	.07	.04	.98	.06	.15	.19	.16	.01	1.00	.01	.02	.17	.01	.05	.05	.05	.01	1.00
9	.13	.03	1.00	.11	.31	.39	.28	.00	1.00	.05	.01	.91	.05	.13	.18	.11	.01	1.00
10	.21	.03	1.00	.14	.55	.44	.35	.00	1.00	.03	.02	1.00	.02	.27	.14	.19	.01	1.00
11	.23	.03	1.00	.12	.60	.39	.46	.02	1.00	.06	.01	1.00	.03	.34	.14	.29	.01	1.00
12	.14	.00	1.00	.05	.53	.25	.56	.01	1.00	.05	.01	1.00	.03	.31	.12	.24	.02	1.00
13	.06	.01	1.00	.02	.30	.08	.48	.04	1.00	.02	.00	.97	.01	.12	.07	.19	.04	.82
14	.02	.00	.99	.01	.12	.04	.30	.01	.98	.02	.01	.56	.02	.06	.02	.04	.02	.30
15	.41	.02	1.00	.28	.77	.72	.59	.01	1.00	.23	.03	1.00	.19	.53	.31	.35	.01	1.00
16	.05	.00	.96	.01	.11	.04	.33	.03	.93	.03	.00	.48	.02	.05	.02	.06	.03	.17
$\bar{X}$	.11	.01	.99	.06	.33	.22	.35	.01	.99	.04	.01	.81	.03	.17	.08	.14	.01	.84
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.03	.33	.61	.02	.02	.03	.29	.29	.46	.00	.20	.28	.00	.04	.02	.16	.08	.28
2	.01	.47	.89	.00	.05	.03	.32	.53	.78	.01	.18	.72	.00	.10	.05	.17	.11	.52
3	.02	.48	.98	.01	.16	.08	.28	.70	.98	.02	.14	.96	.02	.20	.12	.16	.13	.70
4	.02	.20	1.00	.01	.30	.27	.20	.65	1.00	.01	.06	1.00	.01	.25	.17	.08	.16	.84
5	.03	.04	1.00	.03	.51	.50	.16	.59	1.00	.07	.01	1.00	.05	.36	.33	.09	.06	.88
6	.13	.02	1.00	.09	.64	.66	.17	.34	1.00	.08	.00	1.00	.03	.42	.31	.04	.07	.74
7	.22	.06	1.00	.12	.74	.72	.20	.19	.98	.10	.03	1.00	.07	.43	.34	.05	.03	.43
8	.00	.58	.40	.01	.05	.04	.46	.37	.36	.01	.29	.10	.01	.09	.02	.27	.09	.34
9	.01	.85	.90	.01	.10	.12	.47	.74	.76	.02	.40	.67	.03	.16	.09	.31	.18	.65
10	.03	.80	1.00	.02	.29	.28	.44	.86	.99	.01	.34	.99	.02	.29	.20	.28	.22	.93
11	.04	.43	1.00	.04	.59	.67	.30	.88	1.00	.05	.15	1.00	.02	.40	.40	.16	.24	.97
12	.16	.01	1.00	.12	.80	.92	.26	.71	1.00	.12	.01	1.00	.10	.63	.58	.12	.15	.97
13	.44	.05	1.00	.30	.93	.98	.39	.44	1.00	.26	.02	1.00	.12	.71	.68	.12	.09	.90
14	.49	.21	1.00	.36	.95	.96	.43	.18	1.00	.25	.07	1.00	.12	.65	.66	.07	.06	.52
15	.14	.80	1.00	.16	.81	.96	.42	.98	1.00	.11	.35	1.00	.09	.77	.82	.24	.45	1.00
16	.79	.68	1.00	.63	1.00	1.00	.62	.15	1.00	.49	.34	1.00	.33	.89	.90	.16	.07	.50
$\bar{X}$	.16	.37	.92	.12	.49	.51	.34	.53	.89	.10	.16	.86	.06	.40	.35	.15	.14	.70

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.8 Power in detecting uniform DIF in pretest items when the operational test (Figure 57).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.05	.99	.97	.01	.01	.00	1.00	1.00	1.00	.00	.62	.59	.01	.00	.01	.85	.84	.86
2	.01	.39	.30	.01	.00	.00	.68	.68	.70	.01	.12	.13	.00	.00	.01	.17	.18	.21
3	.01	.58	.59	.01	.01	.00	.77	.80	.85	.01	.09	.05	.00	.00	.00	.24	.24	.30
4	.05	1.00	1.00	.00	.01	.01	1.00	1.00	1.00	.02	.79	.84	.01	.01	.00	.93	.92	.94
5	.04	1.00	1.00	.00	.00	.01	1.00	1.00	1.00	.00	.89	.91	.01	.01	.01	.95	.94	.95
6	.01	.76	.70	.00	.00	.01	.82	.83	.83	.00	.23	.22	.00	.00	.00	.27	.31	.29
7	.00	.84	.89	.00	.02	.01	.80	.81	.87	.01	.16	.19	.00	.00	.00	.23	.20	.24
8	.05	1.00	1.00	.01	.02	.02	1.00	1.00	1.00	.03	.96	.98	.00	.01	.00	.97	.96	.99
9	.02	1.00	1.00	.01	.01	.04	1.00	1.00	1.00	.00	.99	.99	.02	.02	.02	.98	.99	.97
10	.01	.85	.85	.00	.01	.01	.84	.85	.83	.01	.29	.28	.01	.02	.01	.24	.25	.20
11	.00	.84	.89	.01	.02	.02	.82	.84	.87	.01	.23	.28	.00	.00	.00	.23	.24	.31
12	.01	1.00	1.00	.06	.06	.11	1.00	1.00	1.00	.02	.88	.93	.02	.01	.03	.91	.92	.95
13	.02	1.00	1.00	.06	.07	.10	1.00	1.00	1.00	.02	.97	.97	.06	.06	.07	.96	.95	.94
14	.01	.72	.74	.02	.02	.04	.76	.77	.75	.00	.26	.25	.02	.03	.01	.23	.24	.21
15	.01	.60	.64	.03	.03	.04	.65	.66	.72	.00	.09	.12	.01	.01	.00	.12	.12	.16
16	.00	1.00	1.00	.17	.20	.25	1.00	1.00	1.00	.01	.52	.55	.02	.03	.03	.66	.67	.73
$\bar{X}$	.02	.85	.85	.02	.03	.04	.88	.89	.90	.01	.50	.51	.01	.01	.01	.56	.56	.58
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.01	1.00	.95	.00	.02	.00	1.00	1.00	.88	.01	.92	.62	.00	.01	.00	.93	.88	.63
2	.01	.98	.54	.01	.00	.00	.94	.86	.45	.01	.63	.23	.00	.01	.01	.67	.51	.21
3	.01	.00	.10	.01	.01	.00	.00	.01	.07	.01	.00	.01	.01	.03	.00	.00	.01	.04
4	.00	.51	.99	.01	.03	.01	.38	.63	.96	.01	.11	.46	.01	.01	.01	.08	.15	.49
5	.04	1.00	1.00	.00	.01	.01	1.00	1.00	1.00	.00	1.00	.93	.02	.05	.02	1.00	.99	.93
6	.05	1.00	.81	.01	.02	.01	1.00	1.00	.82	.01	.80	.43	.00	.00	.01	.85	.75	.42
7	.02	.02	.38	.02	.02	.00	.01	.02	.28	.00	.01	.08	.00	.01	.01	.01	.01	.05
8	.02	.92	1.00	.03	.03	.00	.84	.94	1.00	.03	.52	.91	.01	.02	.01	.26	.39	.73
9	.07	1.00	1.00	.01	.04	.06	1.00	1.00	1.00	.01	1.00	.98	.03	.05	.09	1.00	1.00	.97
10	.04	.99	.77	.01	.04	.05	1.00	1.00	.94	.01	.80	.40	.02	.06	.07	.89	.78	.44
11	.02	.19	.73	.00	.06	.06	.05	.11	.52	.03	.04	.27	.02	.03	.05	.00	.00	.06
12	.06	1.00	1.00	.02	.04	.04	1.00	1.00	1.00	.03	.84	.96	.00	.01	.02	.41	.54	.81
13	.02	1.00	1.00	.06	.15	.37	1.00	1.00	1.00	.02	1.00	.94	.08	.13	.25	1.00	1.00	.98
14	.02	.82	.28	.05	.17	.31	1.00	1.00	.95	.03	.46	.09	.08	.09	.20	.73	.67	.39
15	.12	.59	.93	.06	.12	.15	.16	.27	.69	.13	.20	.57	.08	.11	.12	.02	.02	.10
16	.18	1.00	1.00	.06	.07	.08	1.00	1.00	1.00	.14	.83	.96	.03	.05	.03	.44	.55	.75
$\bar{X}$	.04	.75	.78	.02	.05	.07	.71	.74	.78	.03	.57	.55	.02	.04	.05	.52	.51	.50

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.9 Power in detecting uniform DIF in pretest items when the operational test consisted of 6 items exhibiting uniform DIF with a magnitude of .4 at the beginning of the test (Figure 58).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.04	.97	.85	.00	.00	.01	1.00	1.00	.99	.00	.58	.36	.01	.00	.00	.74	.79	.71
2	.01	.32	.13	.01	.00	.01	.46	.63	.28	.00	.10	.03	.00	.00	.00	.10	.14	.07
3	.02	.55	.86	.00	.01	.00	.85	.81	.98	.01	.03	.08	.00	.00	.01	.28	.23	.42
4	.04	1.00	1.00	.02	.02	.05	1.00	1.00	1.00	.02	.79	.89	.00	.00	.01	.96	.94	1.00
5	.06	1.00	1.00	.02	.02	.01	1.00	1.00	1.00	.01	.86	.68	.00	.00	.01	.88	.95	.87
6	.00	.71	.38	.01	.01	.01	.70	.80	.48	.00	.26	.14	.00	.00	.00	.18	.26	.12
7	.00	.81	.98	.01	.01	.01	.90	.84	.98	.00	.20	.41	.00	.00	.00	.34	.29	.52
8	.01	1.00	1.00	.04	.01	.03	1.00	1.00	1.00	.01	.96	1.00	.01	.02	.01	.97	.96	1.00
9	.02	1.00	1.00	.02	.04	.04	1.00	1.00	1.00	.01	.99	.93	.02	.03	.03	.94	.98	.87
10	.00	.85	.53	.01	.00	.03	.68	.84	.47	.00	.34	.17	.00	.01	.01	.20	.29	.15
11	.01	.82	.98	.01	.01	.01	.89	.79	.98	.01	.18	.45	.00	.01	.01	.25	.18	.50
12	.00	1.00	1.00	.08	.08	.16	1.00	1.00	1.00	.00	.93	.98	.01	.01	.02	.96	.93	.99
13	.00	1.00	1.00	.08	.08	.08	1.00	1.00	1.00	.01	.97	.93	.04	.04	.03	.92	.95	.84
14	.02	.74	.33	.03	.02	.03	.61	.74	.34	.01	.27	.10	.04	.04	.02	.20	.23	.08
15	.01	.55	.88	.02	.02	.03	.72	.65	.96	.01	.16	.29	.01	.01	.00	.22	.21	.36
16	.06	1.00	1.00	.21	.24	.35	1.00	1.00	1.00	.00	.63	.80	.02	.01	.06	.83	.79	.91
$\bar{X}$	.02	.83	.81	.03	.03	.05	.86	.88	.84	.00	.51	.51	.01	.01	.01	.56	.57	.59
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.01	1.00	.85	.00	.01	.01	1.00	1.00	.72	.02	.98	.54	.00	.01	.01	.91	.92	.61
2	.01	.99	.30	.00	.00	.02	.94	.90	.17	.02	.70	.16	.00	.04	.01	.61	.59	.15
3	.02	.00	.26	.02	.04	.00	.00	.01	.22	.01	.00	.01	.00	.03	.01	.00	.00	.06
4	.01	.51	.98	.01	.05	.01	.46	.62	.99	.01	.08	.68	.00	.02	.00	.11	.14	.71
5	.03	1.00	1.00	.01	.02	.01	1.00	1.00	1.00	.02	.99	.84	.02	.04	.02	.99	.98	.81
6	.02	.99	.52	.00	.01	.00	.98	.97	.45	.01	.83	.22	.01	.03	.01	.81	.73	.18
7	.00	.01	.69	.01	.01	.01	.01	.02	.63	.01	.01	.20	.01	.03	.01	.00	.00	.13
8	.02	.90	1.00	.02	.03	.00	.91	.95	1.00	.03	.49	.99	.00	.02	.01	.29	.31	.89
9	.05	1.00	1.00	.02	.05	.07	1.00	1.00	1.00	.03	1.00	.93	.02	.03	.07	1.00	1.00	.94
10	.02	1.00	.43	.02	.05	.07	1.00	1.00	.67	.02	.81	.17	.02	.04	.04	.84	.82	.25
11	.04	.17	.91	.00	.05	.03	.05	.07	.82	.03	.03	.58	.02	.03	.03	.01	.01	.21
12	.10	1.00	1.00	.04	.04	.05	1.00	1.00	1.00	.09	.80	1.00	.02	.04	.01	.45	.48	.94
13	.01	1.00	.98	.09	.17	.40	1.00	1.00	1.00	.03	1.00	.71	.11	.17	.23	1.00	1.00	.90
14	.02	.82	.03	.06	.15	.34	1.00	1.00	.72	.04	.41	.02	.10	.14	.20	.74	.70	.14
15	.11	.56	.99	.04	.09	.17	.23	.23	.94	.08	.18	.73	.07	.06	.08	.02	.02	.21
16	.21	1.00	1.00	.01	.06	.09	1.00	1.00	1.00	.11	.81	.99	.03	.03	.04	.56	.56	.88
$\bar{X}$	.04	.75	.74	.02	.05	.08	.72	.73	.77	.03	.57	.55	.02	.04	.05	.52	.52	.50

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.10 Power in detecting uniform DIF in pretest items when the operational test consisted of 6 items exhibiting nonuniform DIF with a magnitude of .4 at the beginning of the test (Figure 59).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.06	.96	.95	.00	.00	.00	1.00	1.00	1.00	.00	.57	.57	.00	.00	.00	.82	.84	.88
2	.01	.41	.36	.00	.00	.00	.65	.70	.72	.01	.14	.16	.00	.00	.00	.16	.18	.22
3	.02	.57	.54	.00	.00	.00	.69	.70	.72	.03	.04	.02	.00	.01	.01	.18	.16	.17
4	.02	1.00	1.00	.01	.01	.01	1.00	1.00	1.00	.03	.85	.84	.00	.00	.00	.94	.93	.94
5	.05	1.00	1.00	.00	.00	.02	1.00	1.00	1.00	.01	.89	.90	.00	.00	.00	.91	.93	.94
6	.02	.73	.76	.01	.03	.00	.79	.80	.86	.00	.25	.28	.00	.00	.01	.24	.28	.32
7	.01	.82	.84	.01	.01	.00	.82	.82	.83	.01	.19	.20	.01	.01	.02	.29	.26	.27
8	.04	1.00	1.00	.02	.02	.04	1.00	1.00	1.00	.00	.96	.97	.01	.01	.01	.97	.97	.99
9	.01	1.00	1.00	.03	.01	.00	1.00	1.00	1.00	.01	.97	.97	.00	.00	.01	.96	.97	.97
10	.01	.82	.85	.01	.01	.01	.77	.79	.80	.00	.36	.38	.00	.01	.00	.29	.32	.30
11	.00	.80	.82	.01	.01	.00	.78	.77	.80	.01	.21	.24	.01	.01	.01	.24	.25	.30
12	.01	1.00	1.00	.06	.07	.14	1.00	1.00	1.00	.01	.90	.90	.00	.02	.02	.90	.91	.91
13	.01	1.00	1.00	.05	.07	.10	1.00	1.00	1.00	.00	.96	.96	.01	.03	.02	.93	.94	.94
14	.00	.70	.72	.01	.01	.03	.69	.71	.70	.02	.23	.25	.04	.06	.01	.22	.24	.21
15	.02	.49	.48	.07	.07	.09	.57	.57	.63	.02	.12	.14	.01	.01	.01	.14	.15	.17
16	.01	1.00	1.00	.17	.19	.28	1.00	1.00	1.00	.00	.58	.62	.02	.03	.05	.74	.73	.80
$\bar{X}$	.02	.83	.83	.03	.03	.04	.86	.86	.88	.01	.51	.52	.01	.01	.01	.56	.56	.58
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.04	1.00	.96	.00	.00	.00	1.00	1.00	.92	.01	.93	.66	.00	.01	.01	.96	.91	.75
2	.03	.98	.54	.01	.02	.01	.92	.82	.45	.01	.58	.16	.00	.00	.00	.65	.50	.19
3	.01	.01	.09	.02	.02	.01	.01	.01	.09	.02	.01	.02	.00	.02	.01	.01	.01	.05
4	.01	.45	.98	.02	.02	.00	.35	.54	.97	.01	.08	.43	.01	.01	.00	.10	.15	.48
5	.05	1.00	1.00	.01	.03	.02	1.00	1.00	1.00	.03	.99	.96	.03	.03	.03	.99	.98	.96
6	.02	1.00	.84	.01	.03	.02	1.00	1.00	.80	.02	.85	.45	.01	.05	.01	.92	.78	.46
7	.02	.03	.33	.02	.00	.00	.01	.02	.24	.02	.00	.06	.01	.01	.02	.01	.00	.05
8	.02	.91	1.00	.02	.01	.00	.79	.93	1.00	.02	.48	.92	.01	.02	.00	.16	.30	.76
9	.08	1.00	1.00	.00	.03	.03	1.00	1.00	1.00	.03	1.00	.99	.01	.05	.06	1.00	1.00	.98
10	.06	1.00	.82	.02	.06	.05	1.00	1.00	.94	.03	.85	.46	.02	.06	.05	.92	.84	.51
11	.01	.19	.76	.01	.02	.03	.05	.12	.50	.03	.05	.33	.03	.04	.04	.01	.02	.09
12	.07	1.00	1.00	.03	.05	.04	.98	1.00	1.00	.07	.76	.99	.03	.03	.03	.39	.47	.79
13	.03	1.00	1.00	.04	.15	.35	1.00	1.00	1.00	.00	1.00	.94	.08	.12	.18	1.00	1.00	.98
14	.02	.84	.22	.06	.15	.35	1.00	.99	.94	.03	.47	.08	.09	.12	.19	.82	.73	.35
15	.12	.58	.94	.05	.10	.15	.17	.24	.73	.09	.24	.54	.05	.07	.10	.01	.01	.14
16	.15	.99	1.00	.03	.05	.04	1.00	1.00	1.00	.09	.77	.95	.02	.04	.03	.43	.48	.72
$\bar{X}$	.04	.75	.78	.02	.04	.07	.70	.73	.78	.03	.56	.56	.02	.04	.05	.52	.51	.51

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.11 Power in detecting uniform DIF in pretest items when the operational test consisted of 6 items exhibiting both nonuniform and uniform DIF with magnitudes of .4 at the beginning of the test (Figure 60).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.04	.98	.85	.00	.00	.00	1.00	1.00	1.00	.01	.57	.49	.00	.00	.01	.75	.82	.78
2	.01	.35	.18	.01	.00	.03	.51	.61	.41	.00	.11	.06	.00	.00	.00	.14	.17	.13
3	.02	.53	.76	.00	.00	.02	.82	.72	.94	.03	.09	.17	.01	.01	.00	.24	.19	.36
4	.08	1.00	1.00	.01	.01	.01	1.00	1.00	1.00	.03	.87	.94	.00	.00	.01	.97	.96	.99
5	.06	1.00	1.00	.01	.01	.02	1.00	1.00	1.00	.02	.89	.78	.02	.01	.01	.92	.93	.89
6	.00	.71	.34	.01	.00	.01	.65	.77	.56	.00	.28	.19	.00	.00	.01	.24	.31	.18
7	.01	.80	.96	.01	.01	.02	.93	.85	.98	.02	.19	.41	.00	.00	.00	.34	.28	.49
8	.02	1.00	1.00	.01	.01	.05	1.00	1.00	1.00	.01	.97	1.00	.00	.01	.01	.99	.98	.99
9	.01	1.00	1.00	.03	.03	.03	1.00	1.00	1.00	.00	.98	.94	.02	.01	.01	.95	.97	.94
10	.01	.88	.62	.01	.01	.01	.74	.83	.58	.01	.31	.17	.02	.02	.01	.19	.28	.13
11	.02	.82	.99	.03	.03	.02	.86	.81	.99	.02	.22	.41	.01	.01	.00	.30	.24	.45
12	.01	1.00	1.00	.06	.09	.13	1.00	1.00	1.00	.00	.88	.97	.00	.01	.03	.92	.90	.98
13	.01	1.00	1.00	.08	.04	.10	1.00	1.00	1.00	.01	.97	.90	.03	.03	.02	.87	.95	.85
14	.00	.69	.47	.03	.02	.03	.61	.73	.49	.00	.26	.13	.02	.02	.01	.17	.22	.12
15	.00	.51	.79	.03	.04	.03	.69	.60	.87	.01	.07	.27	.01	.01	.00	.22	.12	.33
16	.01	.99	1.00	.18	.18	.25	1.00	1.00	1.00	.00	.61	.78	.00	.02	.04	.82	.78	.89
$\bar{X}$	.02	.83	.81	.03	.03	.05	.86	.87	.86	.01	.51	.54	.01	.01	.01	.56	.57	.59
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.01	1.00	.91	.00	.01	.00	1.00	1.00	.82	.04	.94	.53	.00	.02	.01	.94	.90	.64
2	.06	.98	.31	.02	.01	.01	.90	.89	.26	.00	.68	.17	.00	.02	.00	.60	.54	.15
3	.02	.00	.18	.01	.02	.02	.00	.00	.17	.01	.00	.02	.00	.02	.00	.00	.00	.06
4	.01	.51	1.00	.01	.01	.00	.46	.58	1.00	.01	.07	.64	.00	.05	.01	.10	.14	.68
5	.05	1.00	1.00	.00	.00	.00	1.00	1.00	1.00	.02	1.00	.88	.01	.01	.01	1.00	1.00	.82
6	.06	1.00	.56	.00	.00	.01	1.00	1.00	.58	.01	.85	.25	.00	.02	.01	.80	.78	.23
7	.01	.01	.62	.01	.02	.01	.01	.01	.51	.01	.01	.21	.01	.03	.02	.01	.01	.15
8	.02	.94	1.00	.04	.01	.01	.89	.93	1.00	.01	.47	.99	.00	.01	.01	.27	.32	.85
9	.06	1.00	1.00	.01	.04	.04	1.00	1.00	1.00	.01	1.00	.93	.02	.04	.06	1.00	1.00	.94
10	.03	1.00	.50	.00	.03	.06	1.00	1.00	.82	.00	.80	.23	.02	.05	.05	.84	.82	.33
11	.03	.13	.89	.01	.01	.01	.10	.07	.77	.04	.05	.51	.03	.03	.04	.00	.01	.16
12	.08	.99	1.00	.02	.03	.04	1.00	1.00	1.00	.08	.82	1.00	.03	.04	.04	.50	.55	.93
13	.01	1.00	.98	.03	.11	.26	1.00	1.00	1.00	.00	1.00	.86	.07	.12	.20	1.00	1.00	.96
14	.02	.85	.07	.09	.17	.31	1.00	1.00	.80	.03	.53	.07	.09	.13	.19	.76	.74	.28
15	.09	.48	1.00	.04	.07	.11	.25	.26	.90	.10	.18	.67	.04	.05	.08	.03	.02	.25
16	.18	.99	1.00	.02	.02	.07	1.00	1.00	1.00	.13	.85	1.00	.03	.04	.04	.53	.59	.88
$\bar{X}$	.04	.74	.75	.02	.03	.06	.72	.73	.79	.03	.58	.56	.02	.04	.05	.52	.52	.52

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.12 Power in detecting uniform DIF in pretest items when the operational test consisted of 24 items exhibiting uniform DIF with a magnitude of 1.6 at the end of the test (Figure 61).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.00	.99	.51	.01	.01	.01	.10	1.00	.92	.01	.59	.05	.01	.00	.01	.03	.78	.51
2	.01	.37	1.00	.00	.02	.00	.07	.54	1.00	.01	.08	.44	.00	.00	.00	.01	.12	1.00
3	.30	.51	1.00	.01	.02	.00	1.00	.78	1.00	.10	.05	1.00	.01	.02	.00	1.00	.20	1.00
4	.35	1.00	1.00	.02	.02	.02	1.00	1.00	1.00	.15	.85	1.00	.00	.02	.02	1.00	.95	1.00
5	.01	1.00	.89	.02	.04	.03	.18	1.00	1.00	.00	.86	.24	.01	.01	.00	.05	.90	.65
6	.02	.79	1.00	.02	.02	.01	.09	.83	1.00	.01	.29	.88	.00	.01	.00	.04	.30	1.00
7	.21	.81	1.00	.01	.02	.02	1.00	.83	1.00	.06	.21	1.00	.00	.02	.00	1.00	.26	1.00
8	.37	1.00	1.00	.02	.02	.07	1.00	1.00	1.00	.12	.95	1.00	.01	.01	.01	1.00	.95	1.00
9	.01	1.00	.97	.02	.04	.02	.26	1.00	1.00	.02	.98	.43	.01	.03	.02	.04	.97	.68
10	.05	.83	1.00	.03	.04	.02	.14	.84	1.00	.01	.29	1.00	.01	.03	.02	.06	.26	1.00
11	.10	.85	1.00	.01	.01	.01	1.00	.84	1.00	.07	.24	1.00	.01	.01	.00	1.00	.22	1.00
12	.12	1.00	1.00	.09	.03	.15	1.00	1.00	1.00	.09	.94	1.00	.01	.02	.02	1.00	.93	1.00
13	.05	1.00	.99	.13	.15	.12	.16	1.00	.99	.01	.95	.54	.02	.05	.02	.05	.92	.49
14	.04	.71	1.00	.01	.05	.00	.12	.76	1.00	.03	.22	.99	.03	.04	.02	.01	.24	.96
15	.04	.54	1.00	.04	.03	.02	1.00	.60	1.00	.02	.13	1.00	.01	.01	.01	.92	.15	1.00
16	.02	.99	1.00	.13	.11	.21	1.00	1.00	1.00	.01	.62	1.00	.03	.03	.05	1.00	.75	1.00
$\bar{X}$	.10	.84	.96	.03	.04	.04	.57	.87	.99	.04	.51	.78	.01	.02	.01	.51	.55	.89
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.01	1.00	.13	.00	.02	.00	.32	1.00	.09	.01	.95	.02	.02	.04	.03	.26	.93	.05
2	.00	.98	.62	.01	.03	.01	.04	.81	.42	.01	.71	.22	.01	.05	.00	.05	.51	.32
3	.03	.00	1.00	.02	.04	.01	.55	.01	1.00	.04	.00	1.00	.01	.06	.01	.18	.01	1.00
4	.05	.49	1.00	.04	.05	.00	.99	.59	1.00	.07	.09	1.00	.00	.08	.01	.78	.16	1.00
5	.02	1.00	.16	.01	.03	.01	.69	1.00	.13	.02	1.00	.07	.02	.09	.03	.45	.97	.10
6	.01	1.00	.92	.01	.02	.01	.06	1.00	.89	.01	.87	.62	.01	.05	.03	.07	.78	.53
7	.05	.03	1.00	.01	.08	.00	.89	.03	1.00	.06	.00	1.00	.01	.05	.00	.41	.00	1.00
8	.16	.95	1.00	.04	.05	.00	1.00	.96	1.00	.16	.49	1.00	.01	.07	.01	.93	.36	1.00
9	.01	1.00	.41	.00	.06	.03	.84	1.00	.41	.03	1.00	.19	.03	.13	.04	.59	1.00	.11
10	.01	.99	1.00	.02	.05	.03	.08	1.00	1.00	.01	.80	.93	.01	.10	.05	.07	.83	.72
11	.16	.20	1.00	.04	.10	.03	1.00	.09	1.00	.16	.06	1.00	.03	.07	.04	.54	.01	1.00
12	.34	1.00	1.00	.05	.05	.01	1.00	1.00	1.00	.22	.82	1.00	.02	.05	.01	.98	.55	1.00
13	.02	1.00	.81	.06	.22	.34	.85	1.00	.66	.04	1.00	.61	.10	.21	.24	.50	1.00	.14
14	.08	.77	1.00	.08	.22	.30	.02	1.00	1.00	.12	.45	1.00	.12	.19	.18	.02	.76	.80
15	.48	.55	1.00	.10	.18	.18	1.00	.28	1.00	.27	.24	1.00	.09	.13	.08	.64	.03	1.00
16	.49	.99	1.00	.06	.07	.06	1.00	1.00	1.00	.34	.85	1.00	.08	.08	.04	.98	.51	1.00
$\bar{X}$	.12	.75	.81	.03	.08	.06	.64	.73	.79	.10	.58	.73	.03	.09	.05	.46	.52	.67

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.13 Power in detecting uniform DIF in pretest items when the operational test consisted of 24 items exhibiting nonuniform DIF with a magnitude of 1.6 at the end of the test (Figure 62).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.04	1.00	1.00	.06	.19	.32	.92	1.00	1.00	.01	.81	.77	.05	.12	.15	.82	.75	.66
2	.05	.85	.77	.14	.31	.35	.67	.56	.35	.01	.28	.26	.02	.11	.10	.44	.13	.08
3	.19	.15	.24	.09	.36	.29	.59	.69	.93	.03	.02	.03	.01	.15	.08	.25	.20	.38
4	.26	1.00	1.00	.04	.34	.18	.91	1.00	1.00	.03	.51	.65	.01	.12	.04	.70	.85	.98
5	.05	1.00	1.00	.20	.55	.56	.92	1.00	1.00	.01	.98	.95	.06	.24	.24	.85	.93	.88
6	.05	.96	.89	.17	.59	.55	.63	.78	.55	.02	.53	.38	.05	.25	.20	.35	.22	.14
7	.18	.35	.73	.09	.49	.17	.63	.65	.95	.08	.08	.19	.03	.24	.08	.35	.22	.51
8	.23	1.00	1.00	.04	.39	.08	.96	1.00	1.00	.07	.86	.96	.03	.19	.04	.85	.91	.98
9	.06	1.00	1.00	.25	.83	.77	.87	1.00	1.00	.02	1.00	.98	.06	.38	.28	.80	.96	.92
10	.11	.97	.83	.17	.70	.55	.52	.91	.54	.02	.53	.31	.04	.43	.21	.37	.34	.18
11	.19	.61	.96	.06	.51	.13	.80	.54	.97	.04	.12	.35	.02	.18	.04	.46	.12	.42
12	.16	1.00	1.00	.03	.24	.01	.96	1.00	1.00	.06	.83	.97	.01	.12	.01	.89	.76	.98
13	.08	1.00	1.00	.30	.90	.79	.84	1.00	1.00	.02	.97	.92	.08	.55	.35	.69	.97	.88
14	.16	.81	.34	.22	.79	.61	.38	.90	.36	.02	.30	.13	.04	.45	.17	.19	.39	.10
15	.06	.43	.91	.03	.25	.04	.81	.26	.87	.02	.07	.26	.00	.10	.01	.47	.04	.31
16	.06	1.00	1.00	.03	.08	.00	.99	.98	1.00	.02	.58	.81	.02	.03	.00	.84	.47	.85
$\bar{X}$	.12	.82	.85	.12	.47	.34	.77	.83	.84	.03	.53	.55	.03	.23	.12	.58	.51	.58
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.03	1.00	1.00	.02	.04	.04	.98	1.00	.99	.01	1.00	.87	.02	.13	.10	.97	.93	.73
2	.01	1.00	.78	.02	.09	.06	.91	.99	.70	.00	.88	.41	.03	.14	.12	.81	.66	.20
3	.02	.06	.04	.01	.16	.07	.36	.06	.03	.01	.02	.00	.00	.18	.10	.16	.01	.04
4	.01	.15	.94	.01	.14	.09	.30	.11	.83	.02	.06	.36	.01	.23	.08	.15	.10	.45
5	.03	1.00	1.00	.04	.17	.13	.99	1.00	1.00	.03	1.00	.94	.04	.28	.22	.97	1.00	.86
6	.06	1.00	.84	.02	.19	.14	.95	1.00	.93	.02	.94	.49	.02	.32	.21	.81	.90	.34
7	.02	.01	.37	.01	.31	.24	.32	.06	.08	.03	.01	.09	.02	.30	.20	.14	.02	.05
8	.04	.67	1.00	.02	.32	.18	.57	.35	1.00	.03	.28	.95	.00	.24	.13	.33	.13	.85
9	.06	1.00	1.00	.03	.40	.40	1.00	1.00	1.00	.01	1.00	.99	.05	.50	.36	.98	1.00	.98
10	.07	1.00	.55	.05	.53	.51	.85	1.00	.99	.02	.94	.28	.06	.51	.39	.72	.93	.40
11	.05	.08	.97	.02	.53	.49	.38	.03	.46	.07	.03	.54	.04	.49	.29	.15	.01	.14
12	.20	.99	1.00	.02	.60	.43	.81	.68	1.00	.12	.73	.99	.03	.41	.17	.63	.23	.91
13	.05	1.00	.96	.09	.80	.80	.98	1.00	1.00	.03	.99	.74	.08	.69	.58	.88	1.00	.92
14	.07	.90	.03	.09	.83	.85	.70	1.00	.87	.05	.53	.01	.10	.68	.49	.47	.89	.21
15	.17	.44	1.00	.06	.76	.71	.58	.00	.78	.09	.16	.80	.05	.48	.35	.23	.01	.21
16	.21	1.00	1.00	.04	.66	.47	.93	.79	1.00	.12	.86	.99	.03	.41	.20	.72	.26	.86
$\bar{X}$	.07	.70	.78	.03	.41	.35	.72	.63	.79	.04	.59	.59	.03	.37	.25	.57	.50	.51

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.14 Power in detecting uniform DIF in pretest items when the operational test consisted of 24 items exhibiting both nonuniform and uniform DIF with magnitudes of 1.6 at the end of the test (Figure 63).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.02	1.00	.01	.05	.24	.39	.87	1.00	.46	.01	.84	.00	.06	.14	.16	.72	.71	.27
2	.05	.80	.82	.09	.26	.35	.46	.50	1.00	.03	.33	.15	.04	.13	.15	.27	.15	.95
3	.18	.12	1.00	.07	.34	.30	.71	.73	1.00	.05	.01	1.00	.03	.14	.07	.43	.20	1.00
4	.33	1.00	1.00	.05	.34	.21	.98	1.00	1.00	.12	.59	1.00	.03	.17	.07	.83	.93	1.00
5	.03	1.00	.16	.19	.47	.66	.80	1.00	.70	.03	.98	.00	.07	.22	.30	.70	.90	.41
6	.15	.95	1.00	.21	.48	.55	.40	.73	1.00	.05	.50	.66	.08	.28	.18	.24	.25	.98
7	.23	.44	1.00	.07	.46	.28	.81	.70	1.00	.10	.05	1.00	.04	.25	.10	.60	.20	1.00
8	.39	1.00	1.00	.02	.38	.10	1.00	1.00	1.00	.12	.87	1.00	.02	.16	.03	.93	.94	1.00
9	.09	1.00	.66	.31	.76	.76	.76	1.00	.81	.06	1.00	.15	.10	.40	.31	.61	.95	.42
10	.14	.96	1.00	.19	.72	.66	.38	.87	1.00	.07	.54	.95	.07	.41	.24	.20	.38	.98
11	.27	.65	1.00	.06	.49	.20	.86	.57	1.00	.08	.14	1.00	.03	.30	.04	.73	.15	1.00
12	.21	1.00	1.00	.01	.23	.01	1.00	1.00	1.00	.05	.83	1.00	.02	.11	.01	.94	.72	1.00
13	.12	1.00	.92	.34	.87	.87	.65	1.00	.75	.07	.97	.34	.13	.57	.36	.48	.96	.31
14	.13	.80	1.00	.17	.81	.65	.29	.89	1.00	.06	.30	.96	.07	.46	.18	.13	.43	.88
15	.12	.50	1.00	.02	.26	.05	.91	.26	1.00	.03	.10	1.00	.02	.17	.02	.70	.03	1.00
16	.04	1.00	1.00	.03	.04	.00	1.00	1.00	1.00	.01	.58	1.00	.00	.04	.01	.93	.43	1.00
$\bar{X}$	.15	.82	.85	.11	.45	.38	.74	.83	.92	.06	.54	.70	.05	.24	.14	.59	.52	.82
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.01	1.00	.01	.00	.03	.03	.95	1.00	.01	.01	.99	.00	.05	.14	.12	.89	.97	.00
2	.00	1.00	.18	.01	.07	.06	.83	1.00	.13	.01	.91	.05	.04	.16	.10	.66	.74	.12
3	.02	.05	1.00	.00	.15	.09	.18	.07	1.00	.02	.05	.99	.03	.22	.10	.09	.01	.98
4	.02	.16	1.00	.02	.15	.12	.39	.10	1.00	.04	.01	1.00	.02	.18	.09	.24	.05	1.00
5	.02	1.00	.02	.01	.11	.14	.99	1.00	.01	.00	1.00	.00	.03	.30	.18	.85	1.00	.02
6	.02	1.00	.49	.01	.21	.20	.85	1.00	.33	.02	.96	.46	.05	.34	.21	.61	.88	.31
7	.05	.01	1.00	.03	.29	.33	.22	.06	1.00	.03	.01	1.00	.02	.33	.23	.12	.01	1.00
8	.09	.71	1.00	.01	.30	.34	.73	.38	1.00	.06	.24	1.00	.02	.34	.22	.51	.11	1.00
9	.05	1.00	.15	.03	.37	.46	.98	1.00	.04	.01	1.00	.07	.06	.47	.42	.86	1.00	.02
10	.02	1.00	.92	.04	.47	.49	.75	1.00	.77	.03	.93	.89	.05	.50	.44	.52	.92	.47
11	.11	.06	1.00	.04	.56	.68	.50	.03	1.00	.11	.04	1.00	.05	.39	.34	.26	.00	1.00
12	.18	.96	1.00	.07	.49	.59	.93	.64	1.00	.15	.72	1.00	.03	.38	.30	.75	.19	1.00
13	.04	1.00	.69	.12	.77	.88	.91	1.00	.10	.06	1.00	.57	.08	.68	.60	.74	1.00	.06
14	.07	.94	1.00	.11	.77	.93	.52	1.00	.99	.09	.57	1.00	.09	.71	.66	.31	.87	.57
15	.36	.51	1.00	.09	.80	.91	.76	.02	1.00	.16	.15	1.00	.06	.51	.44	.41	.01	1.00
16	.33	1.00	1.00	.09	.60	.66	.98	.71	1.00	.22	.86	1.00	.04	.43	.32	.84	.25	1.00
$\bar{X}$	.09	.71	.71	.04	.38	.43	.71	.62	.65	.06	.59	.69	.04	.38	.30	.54	.50	.60

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.15 Power in detecting nonuniform DIF in pretest items when the operational test (Figure 64).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.61	1.00	1.00	.91	.75	.99	.73	.74	.83	.05	.19	.27	.06	.05	.13	.11	.12	.13
2	.84	1.00	1.00	1.00	.99	1.00	.99	.99	.99	.09	.79	.91	.22	.18	.42	.45	.42	.45
3	.96	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.14	.98	1.00	.46	.40	.75	.76	.74	.77
4	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.22	1.00	1.00	.66	.57	.90	.93	.94	.95
5	.22	.00	.01	.23	.24	.29	.00	.00	.00	.01	.00	.00	.01	.03	.04	.00	.00	.00
6	.62	.00	.00	.67	.70	.82	.00	.00	.00	.10	.00	.00	.11	.13	.18	.00	.00	.00
7	.93	.01	.01	.95	.96	.99	.01	.01	.01	.23	.00	.01	.25	.32	.41	.00	.00	.00
8	.99	.00	.01	.99	1.00	1.00	.00	.00	.00	.49	.01	.01	.52	.54	.71	.01	.01	.01
9	.88	.01	.02	.91	.91	.99	.01	.00	.00	.19	.00	.00	.17	.23	.31	.00	.00	.00
10	1.00	.01	.05	1.00	1.00	1.00	.00	.00	.00	.50	.00	.01	.49	.52	.76	.00	.00	.00
11	1.00	.02	.07	1.00	1.00	1.00	.00	.00	.00	.68	.00	.01	.69	.75	.92	.00	.00	.00
12	1.00	.01	.04	1.00	1.00	1.00	.00	.00	.00	.84	.00	.01	.85	.88	.97	.00	.00	.00
13	.13	.31	.28	.06	.07	.07	.23	.24	.19	.01	.08	.07	.01	.02	.02	.03	.04	.04
14	.29	.67	.68	.12	.11	.14	.59	.59	.55	.05	.19	.18	.02	.03	.02	.13	.12	.11
15	.50	.96	.97	.23	.23	.27	.92	.93	.92	.09	.42	.39	.03	.03	.03	.32	.32	.30
16	.77	.99	1.00	.38	.40	.45	.99	.98	.98	.11	.74	.76	.06	.06	.09	.61	.61	.60
$\bar{X}$	.73	.44	.44	.71	.71	.75	.40	.40	.40	.23	.27	.29	.29	.29	.41	.21	.21	.21
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.35	.16	.98	.53	.16	.70	.12	.27	.79	.07	.01	.09	.06	.01	.06	.01	.01	.06
2	.69	.94	1.00	.88	.54	.98	.82	.97	1.00	.23	.18	.81	.34	.10	.46	.05	.10	.28
3	.85	1.00	1.00	1.00	.85	1.00	.99	1.00	1.00	.36	.68	.98	.55	.38	.73	.24	.33	.69
4	.87	1.00	1.00	.99	.94	1.00	1.00	1.00	1.00	.52	.93	.99	.81	.67	.92	.59	.70	.87
5	.20	.03	.00	.18	.12	.11	.01	.00	.02	.06	.01	.00	.02	.02	.00	.01	.00	.00
6	.48	.02	.01	.57	.38	.52	.02	.02	.09	.11	.01	.01	.11	.08	.11	.01	.01	.00
7	.77	.01	.01	.84	.77	.85	.02	.05	.24	.35	.00	.00	.35	.21	.33	.00	.00	.00
8	.90	.01	.02	.93	.95	.98	.07	.19	.63	.53	.02	.01	.53	.46	.66	.01	.01	.01
9	.57	.04	.02	.54	.34	.39	.04	.01	.14	.23	.02	.01	.12	.10	.11	.05	.02	.00
10	.88	.01	.04	.89	.83	.95	.01	.03	.44	.44	.00	.01	.36	.33	.41	.02	.00	.00
11	.96	.00	.04	.98	.98	1.00	.09	.19	.75	.69	.00	.00	.68	.59	.71	.01	.00	.00
12	.99	.00	.03	1.00	.99	1.00	.18	.38	.91	.79	.01	.01	.81	.74	.87	.01	.01	.01
13	.02	.12	.03	.00	.00	.00	.30	.21	.06	.00	.03	.01	.01	.01	.01	.11	.09	.03
14	.13	.42	.19	.02	.03	.02	.54	.46	.25	.02	.14	.04	.01	.01	.00	.21	.16	.09
15	.33	.75	.53	.15	.08	.06	.78	.68	.42	.08	.36	.16	.03	.02	.02	.35	.30	.17
16	.48	.95	.78	.26	.17	.20	.94	.89	.73	.13	.48	.32	.06	.05	.05	.50	.41	.25
$\bar{X}$	.59	.34	.35	.61	.51	.61	.37	.40	.53	.29	.18	.21	.30	.23	.34	.13	.13	.15

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.16 Power in detecting nonuniform DIF in pretest items when the operational test consisted of 6 items exhibiting uniform DIF with a magnitude of .4 at the beginning of the test (Figure 65).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.56	1.00	1.00	.87	.74	1.00	.78	.72	.97	.02	.24	.49	.02	.01	.14	.23	.19	.32
2	.91	1.00	1.00	1.00	.98	1.00	.98	.98	1.00	.06	.87	.98	.25	.11	.47	.58	.51	.72
3	.94	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.16	1.00	1.00	.41	.37	.74	.84	.81	.91
4	.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.20	1.00	1.00	.67	.54	.89	.94	.92	.98
5	.26	.00	.02	.29	.28	.38	.00	.00	.01	.03	.02	.03	.02	.01	.03	.02	.01	.02
6	.57	.00	.04	.65	.71	.83	.00	.00	.01	.06	.01	.01	.08	.08	.13	.01	.00	.01
7	.95	.00	.08	.96	.96	1.00	.01	.01	.01	.31	.00	.01	.34	.40	.47	.00	.00	.00
8	1.00	.01	.09	1.00	1.00	1.00	.00	.00	.02	.42	.01	.02	.51	.51	.70	.01	.01	.01
9	.86	.00	.14	.88	.87	.98	.02	.00	.05	.21	.01	.06	.23	.29	.40	.00	.00	.01
10	.99	.03	.24	1.00	1.00	1.00	.02	.01	.05	.42	.00	.04	.49	.52	.72	.00	.00	.00
11	1.00	.03	.29	1.00	1.00	1.00	.00	.00	.04	.61	.01	.02	.64	.69	.94	.01	.01	.00
12	1.00	.00	.30	1.00	1.00	1.00	.00	.00	.02	.84	.00	.02	.86	.86	.99	.00	.00	.00
13	.08	.28	.24	.03	.04	.04	.24	.27	.10	.02	.05	.03	.02	.01	.02	.03	.04	.02
14	.24	.68	.46	.12	.15	.16	.49	.58	.34	.02	.16	.08	.01	.02	.02	.06	.10	.04
15	.49	.91	.84	.19	.22	.31	.81	.87	.71	.04	.38	.28	.02	.01	.03	.29	.31	.18
16	.77	1.00	.99	.41	.42	.50	.98	.98	.94	.13	.65	.60	.08	.07	.12	.49	.55	.42
$\bar{X}$	.72	.43	.48	.71	.71	.76	.39	.40	.39	.22	.27	.29	.29	.28	.42	.22	.21	.23
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.30	.14	1.00	.55	.16	.74	.12	.25	.89	.11	.01	.19	.08	.01	.07	.00	.01	.11
2	.57	.92	1.00	.92	.55	.99	.86	.96	1.00	.22	.13	.87	.30	.12	.41	.04	.06	.42
3	.80	1.00	1.00	.99	.87	1.00	1.00	1.00	1.00	.31	.64	.98	.55	.30	.68	.32	.38	.79
4	.85	1.00	1.00	1.00	.97	1.00	1.00	1.00	1.00	.44	.94	1.00	.76	.62	.88	.66	.70	.94
5	.20	.02	.03	.18	.14	.08	.01	.01	.13	.05	.00	.01	.03	.01	.03	.00	.01	.01
6	.40	.01	.06	.50	.36	.47	.02	.05	.37	.14	.01	.01	.12	.07	.07	.00	.00	.01
7	.76	.00	.04	.83	.75	.88	.04	.06	.61	.35	.01	.03	.32	.27	.34	.01	.01	.03
8	.91	.01	.06	.95	.93	1.00	.17	.22	.84	.49	.00	.00	.53	.48	.58	.00	.00	.00
9	.49	.03	.14	.45	.26	.27	.01	.00	.58	.11	.01	.02	.08	.05	.07	.01	.01	.01
10	.87	.02	.15	.92	.85	.95	.02	.03	.87	.36	.00	.05	.38	.28	.34	.01	.00	.01
11	.92	.00	.22	.96	.92	1.00	.14	.18	.99	.66	.01	.04	.67	.56	.71	.01	.00	.01
12	.98	.00	.20	.99	.99	1.00	.35	.43	.99	.73	.01	.03	.76	.70	.88	.01	.00	.02
13	.02	.12	.01	.01	.00	.00	.28	.24	.02	.01	.04	.00	.01	.01	.01	.09	.08	.01
14	.08	.34	.05	.02	.01	.00	.45	.46	.08	.01	.12	.02	.01	.00	.01	.15	.13	.03
15	.28	.73	.33	.14	.11	.06	.74	.72	.17	.05	.31	.10	.02	.01	.02	.29	.31	.12
16	.52	.90	.62	.28	.24	.20	.88	.93	.42	.12	.51	.25	.05	.04	.05	.47	.47	.23
$\bar{X}$	.56	.33	.37	.60	.50	.60	.38	.41	.62	.26	.17	.22	.29	.22	.32	.13	.13	.17

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.17 Power in detecting nonuniform DIF in pretest items when the operational test consisted of 6 items exhibiting nonuniform DIF with a magnitude of .4 at the beginning of the test (Figure 66).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.46	.99	1.00	.83	.74	1.00	.76	.76	.78	.03	.22	.23	.06	.03	.08	.14	.15	.12
2	.86	1.00	1.00	1.00	.98	1.00	.99	.99	1.00	.13	.87	.96	.28	.20	.47	.52	.50	.47
3	.95	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.20	.99	1.00	.44	.42	.72	.73	.74	.73
4	.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.22	1.00	1.00	.58	.57	.88	.92	.91	.92
5	.21	.01	.01	.25	.28	.31	.01	.01	.01	.02	.00	.00	.02	.02	.02	.00	.00	.00
6	.65	.01	.01	.71	.75	.88	.00	.00	.01	.11	.00	.00	.10	.12	.19	.00	.00	.00
7	.94	.00	.01	.95	.95	.99	.00	.00	.00	.29	.00	.00	.30	.36	.40	.00	.00	.00
8	1.00	.00	.01	1.00	1.00	1.00	.00	.00	.00	.64	.00	.00	.62	.70	.83	.00	.00	.00
9	.84	.01	.01	.88	.89	1.00	.01	.01	.00	.24	.01	.01	.23	.33	.46	.01	.01	.00
10	1.00	.00	.02	1.00	1.00	1.00	.00	.00	.00	.40	.00	.00	.40	.49	.69	.00	.00	.00
11	1.00	.02	.05	1.00	1.00	1.00	.00	.00	.00	.72	.00	.01	.74	.78	.90	.00	.00	.00
12	1.00	.01	.04	1.00	1.00	1.00	.00	.00	.01	.90	.02	.02	.90	.91	.97	.01	.01	.01
13	.10	.28	.30	.07	.09	.08	.18	.18	.18	.02	.07	.07	.01	.02	.01	.04	.04	.03
14	.26	.70	.73	.11	.15	.15	.57	.57	.53	.03	.21	.19	.03	.04	.02	.09	.11	.11
15	.48	.97	.97	.18	.22	.21	.90	.91	.89	.05	.45	.43	.03	.04	.05	.34	.38	.29
16	.75	1.00	1.00	.39	.40	.50	.99	.99	.99	.17	.65	.64	.06	.07	.09	.56	.56	.54
$\bar{X}$	.72	.44	.45	.71	.71	.76	.40	.40	.40	.26	.28	.28	.30	.32	.42	.21	.21	.20
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.35	.33	.96	.49	.19	.71	.11	.22	.76	.08	.00	.07	.05	.01	.11	.00	.00	.03
2	.66	.29	1.00	.94	.66	1.00	.78	.91	1.00	.20	.11	.72	.28	.10	.41	.04	.07	.26
3	.86	.25	1.00	.98	.88	1.00	1.00	1.00	1.00	.32	.64	.97	.56	.38	.73	.24	.34	.65
4	.92	.20	1.00	1.00	.97	1.00	1.00	1.00	1.00	.45	.89	1.00	.76	.59	.90	.41	.58	.86
5	.19	.22	.00	.19	.07	.07	.02	.00	.02	.05	.03	.00	.02	.03	.04	.04	.02	.01
6	.47	.23	.01	.53	.35	.45	.01	.02	.11	.17	.00	.00	.14	.07	.09	.00	.00	.00
7	.76	.24	.02	.86	.75	.89	.05	.06	.36	.35	.00	.00	.33	.26	.32	.00	.00	.00
8	.95	.26	.01	.97	.97	.99	.09	.16	.56	.57	.01	.01	.56	.49	.61	.01	.00	.00
9	.49	.27	.01	.48	.30	.36	.06	.00	.09	.18	.01	.00	.10	.07	.08	.07	.02	.00
10	.89	.29	.01	.91	.83	.95	.04	.04	.49	.44	.00	.00	.34	.25	.37	.02	.00	.00
11	.97	.31	.04	.99	.96	1.00	.09	.18	.72	.60	.00	.00	.58	.46	.62	.00	.00	.00
12	.99	.32	.03	1.00	1.00	1.00	.15	.31	.91	.81	.00	.00	.83	.73	.87	.00	.00	.00
13	.04	.34	.01	.04	.03	.00	.22	.19	.06	.02	.06	.02	.01	.01	.01	.11	.08	.03
14	.10	.34	.21	.04	.03	.03	.50	.43	.20	.03	.14	.04	.01	.00	.00	.20	.15	.07
15	.30	.32	.54	.18	.07	.08	.79	.75	.45	.06	.29	.15	.02	.01	.01	.39	.34	.15
16	.51	.28	.79	.28	.20	.21	.89	.89	.68	.11	.48	.36	.05	.03	.04	.57	.51	.29
$\bar{X}$	.59	.28	.35	.62	.52	.61	.36	.38	.52	.27	.16	.21	.29	.22	.32	.13	.13	.15

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.18 Power in detecting nonuniform DIF in pretest items when the operational test consisted of 6 items exhibiting both nonuniform and uniform DIF with magnitudes of .4 at the beginning of the test (Figure 67).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.46	.99	1.00	.86	.76	1.00	.81	.76	.95	.03	.22	.47	.04	.06	.14	.20	.16	.27
2	.84	1.00	1.00	1.00	.98	1.00	.99	.99	.99	.07	.85	.99	.18	.15	.52	.55	.48	.62
3	.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.16	1.00	1.00	.47	.37	.75	.83	.81	.90
4	.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.22	1.00	1.00	.62	.60	.87	.95	.94	.98
5	.20	.00	.00	.25	.24	.40	.00	.00	.00	.03	.00	.00	.03	.05	.06	.00	.00	.00
6	.66	.00	.03	.69	.72	.84	.00	.01	.00	.07	.00	.00	.08	.11	.17	.00	.00	.00
7	.97	.00	.04	.98	.97	.99	.00	.00	.00	.23	.00	.03	.24	.26	.38	.00	.00	.00
8	.99	.00	.06	1.00	1.00	1.00	.00	.00	.00	.49	.01	.02	.54	.50	.73	.00	.00	.01
9	.79	.01	.19	.84	.87	.97	.01	.01	.03	.16	.00	.00	.19	.27	.42	.00	.00	.00
10	.98	.02	.16	1.00	1.00	1.00	.02	.01	.03	.45	.01	.04	.51	.56	.73	.00	.00	.01
11	1.00	.03	.22	1.00	1.00	1.00	.02	.01	.03	.68	.00	.04	.73	.74	.91	.00	.00	.00
12	1.00	.01	.24	1.00	1.00	1.00	.01	.01	.02	.85	.00	.02	.86	.90	.98	.00	.00	.00
13	.07	.27	.15	.04	.05	.04	.17	.24	.10	.01	.03	.02	.01	.01	.00	.02	.02	.01
14	.27	.64	.48	.11	.14	.13	.41	.56	.34	.04	.17	.12	.02	.02	.03	.08	.11	.06
15	.56	.98	.91	.24	.22	.34	.84	.89	.78	.08	.43	.32	.04	.05	.03	.28	.30	.22
16	.72	1.00	1.00	.32	.38	.49	1.00	1.00	.96	.15	.74	.66	.02	.03	.08	.57	.65	.52
$\bar{X}$	.71	.43	.47	.71	.71	.76	.39	.40	.39	.23	.28	.29	.28	.29	.42	.22	.22	.22
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.25	.15	1.00	.43	.15	.69	.13	.25	.84	.05	.00	.14	.04	.02	.07	.00	.00	.07
2	.68	.94	1.00	.93	.62	1.00	.82	.92	1.00	.16	.15	.87	.26	.09	.39	.08	.08	.39
3	.86	1.00	1.00	1.00	.89	1.00	1.00	1.00	1.00	.30	.65	.98	.58	.35	.71	.23	.30	.71
4	.91	1.00	1.00	1.00	.99	1.00	1.00	1.00	1.00	.43	.90	1.00	.76	.62	.88	.56	.64	.94
5	.20	.02	.02	.19	.12	.12	.00	.00	.08	.05	.01	.01	.02	.02	.03	.01	.01	.01
6	.44	.00	.03	.50	.42	.41	.03	.03	.34	.14	.00	.02	.12	.10	.11	.00	.00	.01
7	.77	.01	.05	.86	.74	.88	.04	.04	.52	.32	.00	.01	.31	.25	.36	.00	.00	.01
8	.93	.00	.06	.96	.95	.98	.14	.13	.80	.54	.00	.02	.56	.47	.62	.01	.00	.01
9	.61	.09	.02	.58	.44	.39	.04	.00	.22	.12	.00	.01	.08	.07	.05	.03	.00	.00
10	.90	.02	.11	.92	.87	.91	.03	.02	.81	.41	.00	.02	.38	.33	.43	.01	.00	.01
11	.96	.01	.12	.97	.97	1.00	.13	.16	.93	.62	.01	.02	.61	.57	.73	.01	.01	.01
12	.99	.00	.15	1.00	1.00	1.00	.26	.35	1.00	.84	.00	.03	.86	.86	.91	.00	.00	.02
13	.04	.07	.01	.02	.01	.00	.14	.14	.02	.00	.06	.01	.00	.00	.01	.10	.09	.02
14	.13	.44	.09	.09	.05	.02	.40	.46	.07	.02	.12	.04	.01	.01	.01	.20	.17	.05
15	.30	.80	.31	.18	.09	.10	.66	.73	.17	.09	.34	.13	.04	.04	.02	.33	.32	.08
16	.52	.94	.71	.28	.17	.20	.86	.90	.53	.15	.53	.29	.05	.04	.06	.51	.47	.22
$\bar{X}$	.59	.34	.35	.62	.53	.61	.35	.38	.58	.26	.17	.22	.29	.24	.33	.13	.13	.16

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.19 Power in detecting nonuniform DIF in pretest items when the operational test consisted of 24 items exhibiting uniform DIF with a magnitude of 1.6 at the end of the test (Figure 68).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.07	.99	1.00	.85	.38	1.00	1.00	.84	1.00	.00	.15	1.00	.05	.00	.16	.99	.20	1.00
2	.34	1.00	1.00	1.00	.89	1.00	1.00	1.00	1.00	.01	.80	1.00	.17	.09	.41	1.00	.52	1.00
3	.60	1.00	1.00	1.00	.99	1.00	1.00	1.00	1.00	.03	.99	1.00	.39	.25	.77	1.00	.87	1.00
4	.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	.08	1.00	1.00	.59	.48	.88	1.00	.94	1.00
5	.08	.01	1.00	.26	.18	.36	.84	.01	1.00	.01	.01	.99	.03	.02	.03	.20	.01	.97
6	.19	.03	1.00	.62	.54	.79	.89	.01	1.00	.04	.00	.98	.08	.07	.16	.16	.00	.95
7	.62	.02	1.00	.95	.90	1.00	.88	.00	1.00	.11	.01	.99	.27	.24	.38	.14	.01	.96
8	.90	.04	1.00	.99	.99	1.00	.92	.01	1.00	.28	.00	1.00	.53	.48	.71	.11	.00	.96
9	.37	.00	1.00	.83	.78	.99	1.00	.00	1.00	.05	.01	1.00	.15	.18	.38	.66	.00	1.00
10	.85	.01	1.00	.99	.99	1.00	1.00	.00	1.00	.17	.00	1.00	.45	.45	.74	.39	.00	1.00
11	.96	.03	1.00	1.00	1.00	1.00	.99	.00	1.00	.45	.01	1.00	.69	.65	.89	.29	.00	1.00
12	1.00	.01	1.00	1.00	1.00	1.00	.98	.00	1.00	.62	.00	1.00	.86	.81	.96	.18	.00	1.00
13	.01	.23	.96	.03	.02	.03	.03	.18	.97	.01	.04	.44	.02	.01	.01	.01	.03	.24
14	.11	.57	.75	.10	.12	.15	.00	.50	.78	.00	.14	.20	.01	.01	.01	.01	.11	.06
15	.29	.89	.31	.22	.20	.30	.04	.84	.40	.02	.32	.10	.02	.03	.06	.01	.22	.03
16	.44	.99	.10	.30	.28	.44	.17	.98	.18	.10	.56	.01	.05	.04	.05	.05	.44	.00
$\bar{X}$	.47	.42	.88	.69	.64	.75	.73	.40	.90	.12	.25	.79	.27	.24	.41	.38	.21	.76
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.08	.14	1.00	.49	.03	.75	.90	.26	1.00	.02	.00	1.00	.06	.00	.10	.30	.00	1.00
2	.31	.89	1.00	.88	.34	.99	1.00	.93	1.00	.06	.11	1.00	.28	.03	.45	.61	.06	1.00
3	.46	1.00	1.00	.99	.70	1.00	1.00	1.00	1.00	.15	.62	1.00	.58	.26	.76	.87	.38	1.00
4	.66	1.00	1.00	1.00	.93	1.00	1.00	1.00	1.00	.22	.94	1.00	.75	.45	.90	.99	.67	1.00
5	.08	.01	1.00	.21	.08	.15	.59	.01	1.00	.02	.01	.98	.04	.02	.00	.06	.01	.92
6	.19	.01	1.00	.52	.28	.56	.83	.02	1.00	.08	.01	.99	.14	.07	.15	.05	.01	.96
7	.48	.01	1.00	.77	.62	.86	.95	.06	1.00	.16	.01	.99	.30	.17	.33	.08	.00	.96
8	.70	.01	1.00	.95	.92	1.00	.99	.19	1.00	.38	.01	.99	.59	.45	.68	.07	.00	.97
9	.17	.02	1.00	.51	.26	.37	.92	.00	1.00	.07	.01	1.00	.14	.06	.13	.12	.02	1.00
10	.49	.01	1.00	.85	.69	.91	1.00	.05	1.00	.21	.01	1.00	.35	.24	.44	.13	.01	1.00
11	.77	.02	1.00	.95	.93	1.00	.99	.17	1.00	.41	.00	1.00	.60	.44	.71	.14	.00	1.00
12	.89	.01	1.00	.98	1.00	1.00	1.00	.41	1.00	.58	.00	1.00	.80	.63	.89	.17	.00	1.00
13	.02	.10	.92	.02	.01	.01	.06	.20	1.00	.02	.05	.69	.02	.01	.02	.00	.06	.29
14	.05	.32	.73	.05	.02	.03	.02	.40	.99	.02	.11	.40	.01	.00	.00	.01	.12	.15
15	.14	.68	.43	.15	.05	.06	.01	.64	.98	.04	.22	.22	.04	.02	.02	.02	.30	.09
16	.25	.89	.18	.25	.14	.21	.01	.86	.95	.07	.50	.04	.05	.05	.04	.03	.43	.02
$\bar{X}$	.36	.32	.89	.60	.44	.62	.70	.39	.99	.16	.16	.83	.30	.18	.35	.23	.13	.77

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.20 Power in detecting nonuniform DIF in pretest items when the operational test consisted of 24 items exhibiting nonuniform DIF with a magnitude of 1.6 at the end of the test (Figure 69).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
	No Test Impact, Balanced Sample									**No Test Impact, Unbalanced Sample**								
1	.07	.66	.91	.13	.14	.08	.45	.80	.95	.01	.05	.07	.00	.02	.00	.17	.25	.34
2	.15	1.00	1.00	.38	.43	.64	.72	1.00	1.00	.02	.42	.66	.03	.03	.02	.37	.54	.69
3	.33	1.00	1.00	.71	.64	.92	.92	1.00	1.00	.03	.82	.97	.04	.07	.12	.64	.78	.89
4	.45	1.00	1.00	.79	.83	1.00	.98	1.00	1.00	.06	.95	1.00	.13	.15	.27	.77	.91	.97
5	.07	.01	.01	.06	.04	.01	.19	.00	.01	.01	.01	.00	.01	.02	.01	.03	.01	.00
6	.20	.03	.01	.19	.06	.10	.18	.01	.01	.01	.01	.01	.01	.00	.01	.02	.01	.01
7	.41	.00	.03	.49	.20	.62	.18	.00	.01	.05	.01	.01	.06	.02	.06	.04	.01	.01
8	.62	.01	.04	.67	.40	.87	.20	.01	.01	.07	.00	.00	.09	.06	.21	.05	.00	.00
9	.10	.01	.05	.10	.07	.04	.36	.00	.04	.01	.00	.02	.00	.04	.01	.15	.00	.01
10	.38	.02	.09	.41	.20	.62	.29	.02	.02	.06	.01	.01	.05	.02	.11	.08	.00	.01
11	.58	.01	.11	.62	.39	.95	.29	.01	.02	.11	.01	.01	.11	.09	.25	.06	.01	.00
12	.73	.01	.12	.80	.60	1.00	.27	.00	.02	.17	.01	.02	.17	.18	.45	.05	.00	.00
13	.05	.35	.11	.03	.08	.00	.12	.42	.10	.01	.06	.03	.01	.04	.01	.03	.08	.03
14	.05	.70	.36	.01	.01	.00	.23	.78	.32	.01	.19	.05	.01	.01	.00	.08	.20	.04
15	.18	.94	.78	.06	.02	.04	.43	.96	.69	.01	.50	.33	.00	.00	.01	.22	.48	.26
16	.30	1.00	.99	.10	.01	.09	.64	1.00	.98	.04	.63	.48	.01	.00	.01	.36	.65	.42
$\bar{X}$	.29	.42	.41	.35	.25	.43	.40	.44	.38	.04	.23	.23	.04	.04	.09	.19	.24	.23
	Test Impact, Balanced Sample									**Test Impact, Unbalanced Sample**								
1	.15	.03	.57	.09	.03	.02	.24	.03	.36	.06	.00	.02	.03	.02	.00	.12	.00	.06
2	.35	.42	.99	.42	.21	.34	.23	.41	.97	.06	.02	.37	.06	.02	.04	.02	.03	.22
3	.48	.89	1.00	.66	.43	.83	.53	.89	1.00	.09	.23	.84	.14	.05	.16	.09	.18	.62
4	.67	.99	1.00	.91	.70	.98	.79	1.00	1.00	.10	.63	.99	.24	.16	.38	.23	.45	.85
5	.17	.05	.04	.12	.01	.01	.19	.10	.05	.02	.02	.01	.02	.00	.00	.07	.02	.01
6	.36	.02	.01	.35	.02	.07	.16	.03	.06	.04	.03	.01	.03	.01	.01	.04	.03	.00
7	.52	.05	.01	.58	.12	.41	.21	.01	.22	.14	.03	.00	.11	.01	.04	.06	.02	.01
8	.72	.04	.05	.81	.39	.74	.18	.01	.43	.18	.02	.01	.18	.07	.21	.03	.01	.01
9	.26	.19	.04	.15	.04	.00	.38	.34	.02	.02	.07	.03	.01	.04	.01	.22	.09	.00
10	.47	.11	.07	.48	.08	.14	.34	.09	.16	.11	.03	.01	.08	.01	.03	.13	.03	.01
11	.71	.05	.06	.77	.30	.66	.30	.02	.47	.16	.01	.02	.11	.04	.10	.07	.00	.01
12	.78	.05	.11	.86	.54	.93	.30	.02	.69	.32	.02	.04	.30	.19	.41	.08	.02	.03
13	.04	.19	.00	.03	.07	.01	.21	.62	.02	.01	.06	.00	.00	.03	.02	.09	.17	.01
14	.13	.42	.06	.06	.03	.00	.31	.86	.11	.01	.17	.02	.01	.01	.01	.14	.33	.05
15	.22	.75	.22	.13	.02	.01	.43	.94	.27	.02	.32	.09	.00	.02	.01	.24	.47	.09
16	.38	.93	.52	.19	.02	.03	.50	1.00	.47	.04	.49	.15	.02	.01	.01	.32	.62	.16
$\bar{X}$	.40	.32	.30	.41	.19	.32	.33	.40	.39	.08	.13	.16	.08	.04	.09	.12	.15	.13

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics

H.21 Power in detecting nonuniform DIF in pretest items when the operational test consisted of 24 items exhibiting both nonuniform and uniform DIF with magnitudes of 1.6 at the end of the test (Figure 70).

Item	LR-UDIF			LR-NDIF			MH			LR-UDIF			LR-NDIF			MH		
	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$	NCS	$\hat{\theta}_{CBT}$	$\hat{\theta}_{CAT}$
No Test Impact, Balanced Sample										**No Test Impact, Unbalanced Sample**								
1	.03	.74	1.00	.09	.16	.03	.65	.80	1.00	.00	.06	1.00	.01	.01	.00	.35	.21	1.00
2	.10	1.00	1.00	.35	.44	.50	.82	.99	1.00	.01	.44	1.00	.02	.05	.02	.54	.54	1.00
3	.21	1.00	1.00	.70	.65	.92	.99	1.00	1.00	.01	.84	1.00	.05	.06	.10	.78	.82	1.00
4	.32	1.00	1.00	.82	.81	.99	.99	1.00	1.00	.01	.98	1.00	.10	.11	.24	.88	.93	1.00
5	.03	.02	1.00	.04	.02	.01	.23	.00	1.00	.00	.01	.95	.00	.01	.01	.08	.01	.95
6	.15	.01	1.00	.22	.06	.08	.32	.01	1.00	.01	.00	.97	.01	.01	.01	.08	.00	.93
7	.30	.01	1.00	.40	.14	.41	.26	.01	1.00	.02	.01	.98	.01	.01	.06	.03	.01	.93
8	.56	.00	1.00	.62	.44	.84	.24	.00	1.00	.09	.01	.97	.12	.11	.20	.05	.00	.88
9	.09	.01	1.00	.11	.07	.08	.51	.01	1.00	.02	.02	1.00	.01	.03	.01	.22	.01	1.00
10	.30	.01	1.00	.37	.18	.45	.43	.01	1.00	.02	.01	1.00	.02	.03	.04	.13	.00	1.00
11	.53	.01	1.00	.65	.46	.90	.35	.01	1.00	.07	.01	1.00	.09	.07	.20	.10	.00	1.00
12	.73	.00	1.00	.82	.61	.99	.28	.00	1.00	.12	.01	.99	.14	.15	.41	.08	.01	.97
13	.01	.27	.82	.03	.04	.00	.06	.35	.80	.01	.04	.31	.01	.03	.01	.01	.08	.22
14	.07	.69	.45	.02	.03	.01	.13	.75	.51	.02	.18	.05	.02	.01	.00	.07	.22	.03
15	.15	.90	.14	.07	.01	.03	.30	.94	.17	.02	.36	.04	.02	.01	.00	.11	.39	.01
16	.29	1.00	.01	.11	.02	.06	.55	1.00	.02	.03	.69	.00	.01	.01	.01	.22	.68	.00
$\bar{X}$	.24	.41	.84	.34	.26	.39	.44	.43	.84	.03	.23	.77	.04	.04	.08	.23	.24	.74
Test Impact, Balanced Sample										**Test Impact, Unbalanced Sample**								
1	.08	.01	1.00	.08	.01	.01	.16	.01	1.00	.02	.01	.99	.01	.01	.00	.06	.00	.97
2	.20	.41	1.00	.35	.20	.26	.30	.38	1.00	.03	.01	1.00	.03	.02	.02	.10	.03	.99
3	.45	.93	1.00	.72	.48	.72	.68	.93	1.00	.08	.20	1.00	.11	.07	.15	.20	.15	1.00
4	.58	1.00	1.00	.90	.75	.96	.88	1.00	1.00	.08	.58	1.00	.19	.13	.28	.40	.40	1.00
5	.11	.06	1.00	.12	.01	.00	.13	.11	1.00	.01	.03	.95	.02	.02	.01	.04	.03	.71
6	.26	.06	.99	.31	.03	.03	.15	.05	1.00	.04	.01	.95	.03	.02	.00	.06	.01	.78
7	.41	.06	1.00	.54	.17	.29	.21	.01	1.00	.12	.01	.96	.13	.01	.07	.02	.01	.87
8	.61	.04	1.00	.75	.36	.59	.23	.00	1.00	.16	.01	.96	.18	.09	.16	.06	.01	.88
9	.12	.25	1.00	.10	.05	.00	.30	.36	1.00	.02	.08	1.00	.01	.05	.00	.14	.11	1.00
10	.36	.06	1.00	.44	.05	.12	.29	.07	1.00	.10	.06	1.00	.07	.01	.02	.10	.08	.98
11	.62	.05	1.00	.76	.32	.50	.33	.02	1.00	.20	.04	1.00	.19	.04	.12	.08	.02	.99
12	.76	.04	1.00	.90	.59	.87	.37	.03	1.00	.26	.02	1.00	.29	.10	.25	.09	.02	.98
13	.03	.15	.90	.01	.06	.04	.11	.67	.93	.01	.07	.66	.00	.05	.01	.03	.21	.24
14	.06	.53	.53	.04	.03	.01	.14	.86	.78	.02	.14	.37	.02	.01	.00	.07	.33	.10
15	.17	.78	.28	.09	.01	.01	.20	.96	.70	.03	.31	.18	.01	.01	.02	.16	.48	.04
16	.26	.92	.12	.18	.03	.01	.31	1.00	.54	.03	.58	.06	.01	.01	.00	.25	.66	.02
$\bar{X}$	.32	.33	.86	.39	.19	.28	.30	.40	.93	.07	.13	.82	.08	.04	.07	.11	.16	.72

Note: Three matching variables: NCS = number-correct score, CBT = $\hat{\theta}_{CBT}$, and CAT = $\hat{\theta}_{CAT}$
Three DIF detection methods: LR-UDIF = the group effect in logistic regression,
LR-NDIF = the interaction effect in logistic regression,
MH = the Mantel-Haenszel statistics