A Modified CATSIB Procedure for Detecting Differential Item Function on Computer-Based Tests

Johnson Ching-hong Li[1]

Mark J. Gierl[1]

Hollis Lai[1]

Louis Roussos[2]

[1]Centre for Research in Applied Measurement and Evaluation, University of Alberta

[2]Measured Progress

Correspondence concerning this article can be addressed to Johnson Ching-hong Li:

Department of Educational Psychology,

6 - 110 Education North,

University of Alberta,

Edmonton, Alberta, T6G 2G5.

Email: johnson.li@ualberta.ca

Phone: (1) 780-492-3762

Fax: (1) 780-492-0001

Abstract

Nandakumar and Roussos (2004) developed a procedure for evaluating DIF items at the pre-test stage in computerized adaptive testing (called CATSIB-G). This procedure, however, assumes that examinees share the same group-level reliability estimate, despite the fact that some of these students should have stronger or weaker estimates in CAT. Recently, Raju, Price, Oshima, and Nering (2007) have proposed an examinee-level reliability estimate tailored for each examinee. The purpose of this study is to combine CATSIB-G procedure and the examinee-level reliability (called CATSIB-E), and evaluate the Type I error and power rates in detecting non-DIF and DIF items, respectively. Simulation results reveal that CATSIB-E yielded higher power than conventional CATSIB-G, thereby providing a useful method for detecting DIF items in a pre-test setting.

A Modified CATSIB Procedure for Detecting Differential Item Functioning on Computer-Based Tests

Computerized adaptive testing (CAT) is a computer-based approach to measurement of examinees' latent proficiency scores. In CAT, examinees are administered test items that are matched on their performance to previously administered items. Examinees with better performances on previous items will receive a more difficult item whereas examinees with poorer performances will receive a less difficult item. After completing all the test items, examinees' responses are converted to latent proficiency scores (θ) based on item response theory (IRT) models, so that every examinee can be compared and evaluated on a common ability scale (Chang, Plake, Kramer, & Lien, 2011).

An important advantage of CAT is its increased measurement efficiency (Frey & Seitz, 2011). Because the item-selection procedure is focused on tailoring the items that match the ability level of each examinee, test length is typically shorter in CAT than in conventional paper-and-pencil test (PPT). For example, Segall (2005) found that the number of items in CAT can generally be reduced by half when compared with a comparable PPT without a lost in measurement precision. Other benefits of CAT include the capability of accommodating innovative item types, flexible test schedules, fast data analysis, and immediate score reporting (Cui & Li, 2010). Because of these important benefits, many large-scale testing programs use CAT, including the Graduate Record Exam (GRE), the Graduate Management Admission Test (GMAT), the Armed Services Vocational Aptitude Battery (ASVAB), and the Certified Public Accountants Licensure Exam.

Despite the potential benefits of CAT, a constant concern in traditional PPT—differential item functioning (DIF)—is also expected to exist in CAT. An item is considered to display DIF when test takers of equal ability, but from different subgroups (e.g., gender, ethnicity), differ in their probabilities for producing a correct response to that item. This

phenomenon is called DIF. It poses potential threats to test fairness because it means that examinees with the same ability have a different probability of answering an item correctly. Because of this threat, the three professional organizations—the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME)—develop the fourth iteration of *the Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999):

> Fairness in testing refers to perspectives on the ways that scores from tests or items are interpreted in the process of evaluating test takers for a selection or classification decision. Fairness in testing is closely related to test validity, and the evaluation of fairness requires a broad range of evidence that includes empirical data, but may also involve legal, ethical, political, philosophical, and economic reasoning. (AERA et al, 1999, p. 225)

Considerable progress has been made in developing the analytical procedures for detecting DIF in CAT. For example, Zwick, Thayer, and Wingersky (1994, 1995) modified the PPT Mantel-Haenszel DIF procedure for every administered item in a real CAT (known as operational item). As an extension, Nandakumar and Roussos (2001, 2004) argued that the DIF procedure in CAT should be developed for pilot items at the "pre-test stage" (known as pre-test items) rather than for the operational items at the "post-test stage" as in Zwick et al.'s. One benefit of the pre-test approach is that the problematic DIF items can be edited or deleted before they are administered to examinees for scoring purposes. In contrast, if DIF items are detected in operational test forms or at a "post-test stage" on a CAT, they could contaminate the latent proficiency estimates for examinees because the items are not functioning is a consistent manner across all examinee subgroups.

Although Nandakumar and Roussos (2001, 2004) provided a useful procedure for detecting DIF items in a pre-test setting, this procedure is based on the regression correction

formula suggested in Shealey and Stout's (1993) SIBTEST, in which every examinee in the same group (i.e., either reference or focal group) shares the same group-level reliability estimate. This approach ignores the fact that some examinees should have stronger or weaker reliability estimates yielded by a CAT algorithm because each IRT ability estimate has a unique standard error. To solve this problem, Raju, Price, Oshima, and Nering (2007) proposed a procedure for estimating the examinee-level reliability specific to each examinee. This procedure is expected to improve the accuracy of the reliability estimates as well as of the regression correction in Nandakumar and Roussos's procedure for detecting pretest DIF items. We call this combined procedure the CATSIB-E procedure (i.e., CATSIB based on Examinee-level reliability), and the original Nandakumar and Roussos's procedure the CATSIB-G procedure (i.e., CATSIB based on Group-level reliability).

To-date, however, no study has evaluated the potential benefits of CATSIB-E relative to the conventional CATSIB-G. Thus, the purpose of this study is to compare the Type I error rates in detecting non-DIF items and the power rates in detecting DIF items, yielded by the CATSIB-E and CATSIB-G procedures. A Monte Carlo simulation study was conducted to evaluate systematically the performance of the two methods. Our study is divided into four sections. In the first section, we discuss the importance of detecting DIF items at the pre-test stage. In the second section, we present the algorithms for CATSIB-E and CATSIB-G. In the third section, we describe the design, methods, and results of the simulation study. In the fourth section, we offer our conclusions and discussion of the findings in this study.

Detecting DIF Items at the Pretest Stage

In the existing literature, there are two common designs used for identifying DIF items in CAT (Nandakumar & Roussos, 2004). The first design is to detect potential DIF items at the pre-test stage before the items are administered to examinees for real scoring purposes. These items are known as pre-test items, which serve as pilot or experimental items

in a real test, but they are not scored for the examinees. For example, the LSAT administers experimental items to examinees:

> The test consists of five 35-minute sections of multiple-choice questions. Four of the five sections contribute to the test taker's score. The unscored section, commonly referred to as the variable section, typically is used to pretest new test questions or to preequate new test forms (LSAT, 2011).

If an item is found to display statistical bias that favor one specific group of examinees theoretically (i.e., by content experts' reviews) and analytically (i.e., by DIF analysis), it will be edited or discarded from future examination purposes.

The second design evaluates the amount of DIF directly using operationally administered items (i.e., the post-test approach). Using this design, the proficiency estimates for each examinee in a group may need to be adjusted based on the DIF level of each item. This process is often complicated because the scoring and equating procedures of the test may need to be adjusted by both content experts and measurement specialists (Nandakumar & Roussos, 2004).

In the existing literature of DIF in CAT, researchers to-date have focused on the post-test approach, i.e., evaluating the DIF amounts of operational items in a CAT environment (e.g., Zwick et al., 1994, 1995; Roussos, Schnipke, & Pashley, 1999). On one hand, it is reasonable for measurement specialists to ensure that all operational items be free from DIF in a real test. But on the other hand, measurement specialists could choose to be more proactive by preventing these items from appearing on operation forms by evaluating their DIF magnitude at the pre-test stage. Or, said differently, the operational items can be purified from DIF using Nandakumar and Roussos' (2004) CATSIB procedure at the pre-test stage.

The pre-test approach has two main advantages (Nandakumar & Roussos, 2004). The first advantage is the reduced sample size for evaluating DIF items. With CAT, items are

selected adaptively based on examinee's proficiency estimates, meaning that each item is only exposed to a small number of test takers. For an appropriate DIF evaluation, however, each item is often expected to have a reasonable number of test takers in each group of examinees (e.g., 250 in Nandakumar & Roussos, 2004). Given the pre-test approach, each item behaves as an experimental item and can be administered to the entire group of test takers non-adaptively.

The second advantage is the improved scoring. When the test has been purified in the pre-test setting, the proficiency scores estimated for examinees are expected to be more accurate and to be free from DIF. Nandakumar and Roussos (2004, p. 178) called this the "external-matching criterion" approach, meaning that the scores can be matched or purified externally by another group of examinees at the pre-test stage. Conversely, the idea of estimating examinees' proficiency scores and items' DIF amounts simultaneously tend to contaminate the scoring system, as implied in the post-test approach. Nandakumar and Roussos called this the "internal-matching criterion" approach, and that is "why it is important to detect DIF at the pretest stage" (p. 178).

A Modified CATSIB Procedure Based on Examinee-Level Reliability (CATSIB-E)

This section outlines the conventional CATSIB-G procedure proposed by Nandakumar and Roussos (2004) and our modified CATSIB-E procedure.

CATSIB-G

The conventional CATSIB-G procedure can be considered a modified SIBTEST applied to a CAT setting. Generally speaking, SIBTEST is a nonparametric approach to detect DIF (Shealy & Stout, 1993). Assume $DIF(\theta)$ is the level of *DIF* conditional on θ in a studied item,

$$DIF(\theta) = P_R(\theta) - P_F(\theta), \quad (1)$$

where $P_R(\theta)$ [or $P_F(\theta)$] is the probability of a correct response conditional on θ in the reference (or focal) group. Because $DIF(\theta)$ indicates only the probability difference between the two groups at θ, $DIF(\cdot)$ is further defined as an average of $DIF(\theta)$ across all θs, such that it can represent the level of overall group difference. β is used to denote this average, which can be expressed as

$$\beta = \int_{-\infty}^{\infty} DIF(\theta) f(\theta) d\theta, \quad (2)$$

where $f(\theta)$ is the normal density function on θ for the combined reference and focal groups. If $\beta > 0$, then the reference group test takers performs better than the focal group test takers on that item, even when they are matched on θ.

The purpose of SIBTEST is to select reference and focal group examinees matched on their estimated ability ($\hat{\theta}$), thereby ensuring that the difference is due to the DIF but not to their true ability differences. The matching on estimated ability score is accomplished through the regression correction procedure to obtain an unbiased DIF estimation. The corrected ability estimate ($\hat{\theta}^*$) is given by

$$\hat{\theta}^* = E_G(\hat{\theta}) + \hat{\rho}_G^2[\hat{\theta} - E_G(\hat{\theta})], \quad (3)$$

where *G* denotes either reference or focal group, $E_G(\hat{\theta})$ is the expected score (or mean) of the ability scores estimated for group *G*, $\hat{\theta}$ is the ability score estimated for a particular examinee, and $\hat{\rho}_G^2$ is the correlation between the estimated ability scores $\hat{\theta}$ and the true ability scores θ in group *G* (i.e., reliability estimate based on classical test theory), which is given by

$$\hat{\rho}_G^2 = \sqrt{1 - \frac{\sigma_{e\,G}^2}{\sigma_{\hat{\theta}\,G}^2}}, \quad (4)$$

where $\sigma_{\hat{\theta}\,G}^2$ is the variance of the estimated ability scores for group *G*, and $\sigma_{e\,G}^2$ is the variance of the error scores for group *G*.

It is, however, impractical to obtain a sample of examinees which includes all the θ scores ranging from $-\infty$ to ∞ in (2). Hence, combining (1) and (2), an estimated average of *DIF* (denoted as $\hat{\beta}$) can be expressed as

$$\hat{\beta} = \sum_{\hat{\theta}^*=\hat{\theta}^*_{min}}^{\hat{\theta}^*_{max}} [\hat{p}_R(\hat{\theta}^*) - \hat{p}_F(\hat{\theta}^*)]\hat{p}(\hat{\theta}^*), \quad (5)$$

where $\hat{p}_R(\hat{\theta}^*)$ [or $\hat{p}_F(\hat{\theta}^*)$] is the observed proportion of reference group (or focal group) examinees with ability estimate $\hat{\theta}^*$ who answer the studied item correctly, and $\hat{p}(\hat{\theta}^*)$ is the observed proportion of reference and focal group examinees at $\hat{\theta}^*$. Because the observed $\hat{\theta}^*$ are continuous in nature, they can be divided into *K* intervals so that group-performance comparisons can be made in each of the intervals, so that

$$\hat{\beta} = \sum_{k=1}^{K} [\hat{p}_{R,k} - \hat{p}_{F,k}]\hat{p}_k, \quad (6)$$

where $\hat{p}_{R,k}$ (or $\hat{p}_{F,k}$) is the observed proportion of the examinees in the reference group (or focal group) in ability interval *k* who answer the item correctly, $\hat{p}_k$ is the observed proportion of reference and focal group examinees who are classified into interval *k*, and *K* is the total number of intervals, which is generally arbitrary and will be discussed in the Simulation Procedure Section below. The variance of $\hat{\beta}$ can then be estimated based on the observed variance of the responses in each ability interval,

$$Var(\hat{\beta}) = \sum_{k=1}^{K} \left[\frac{\hat{\sigma}^2_{R,k}(Y)}{n_{R,k}} + \frac{\hat{\sigma}^2_{F,k}(Y)}{n_{F,k}}\right]\hat{p}_k^2, \quad (7)$$

where *Y* is the response to the item, $\hat{\sigma}^2_{R,k}(Y)$ [or $\hat{\sigma}^2_{F,k}(Y)$] is the observed variance of *Y* in ability interval *k* for the reference (or focal) group, and $n_{R,k}$ (or $n_{F,k}$) is the number of the examinees in the reference (or focal) group in interval *k*. Given $\hat{\beta}$ in (6) and $Var(\hat{\beta})$ in (7), the 95% confidence interval (CI) for β is given by

$$\hat{\beta} \pm 1.96 \times \sqrt{Var(\hat{\beta})}. \quad (8)$$

For CATSIB-G, two parameters—the ability estimates $\hat{\theta}$ and the reliability estimate $\hat{\rho}_G^2$—are required. The ability estimates of each examinee ($\hat{\theta}$) are typically provided by a CAT algorithm software, as is the case in the present study. The latter estimate $\hat{\rho}_G^2$ is the reliability estimate for group *G*, which depends on the variance of the ability estimates $\sigma^2_{\hat{\theta}\,G}$ and the standard error estimate $\sigma^2_{e\,G}$. $\sigma^2_{\hat{\theta}\,G}$ can be computed easily by the variance of the $\hat{\theta}$ estimates in group *G*, while $\sigma^2_{e\,G}$ is equal to the inverse of the mean test information estimates for group *G*, i.e.,

$$\sigma^2_{e\,G} = \frac{1}{Mean(Info_G)}. \quad (9)$$

CATSIB-E

CATSIB-E follows the same procedure as in CATSIB-G, except for the reliability estimate in (4). Rather than the group-level reliability $\hat{\rho}_G^2$, a reliability estimate based on particular examinee *j* (i.e., $\hat{\rho}_j^2$) is used. This approach is expected to improve the accuracy of the regression correction procedure because a more precise estimate of reliability is tailored for every test taker than is the case in $\hat{\rho}_G^2$ (Raju et al., 2007). Consequently, Equation (4) becomes

$$\hat{\rho}_j^2 = \sqrt{1 - \frac{\sigma^2_{e\,j}}{\sigma^2_{\hat{\theta}\,G}}}. \quad (10)$$

The standard error for examinee *j* (i.e., $\sigma^2_{e\,j}$) is often provided by a CAT algorithm software. Alternatively, it can be estimated by the inverse of each examinee's information estimate, i.e.,

$$\sigma^2_{e\,j} = \frac{1}{Mean(Info_j)}. \quad (11)$$

Simulation Study

Design

Four factors that may affect the accuracy of the CATSIB-G and CATSIB-E procedures were manipulated, including the reference and focal group sizes (N_r and N_f), impact level (d_t), amount of DIF (β), and *a*- and *b*- item parameters ($\emptyset$). For the most part, these factors serve as the same variables manipulated in Nandakumar and Roussos (2004).

Factor 1. Sample sizes of reference and focal groups ($N_r + N_f = N$; three levels). As first described in Nandakumar and Roussos (2004), three sample size levels—(500, 500), (500, 250), and (250, 250)—were manipulated, where the first value indicates the reference group size N_r, and the second value represents the focal group size N_f. The first two levels represent typical sample sizes which may occur in a pre-test CAT setting (Nandakumar & Roussos, 2004). The last level (250, 250) was selected to explore the performance of the CATSIB procedure under a small sample size situation.

Factor 2. Impact level (d_t; three levels). As discussed in Nandakumar and Roussos (2004), the same three impact levels—0, .50, and 1.0—were manipulated in this study. These levels indicate that the means of the reference and focal group ability distribution are 0, .50, and 1.0 standard deviation apart. These ability differences represent a typical situation, where there is no difference between the reference and focal groups (i.e., 0), a small difference (i.e., .50), or a medium to large difference (i.e., 1.0) in ability levels between the two groups.

Factor 3. Amount of DIF (β; four levels). Four amounts of DIF—0, .05, .10, and .15—were manipulated to evaluate the effect of DIF on the CATSIB-G and CATSIB-E procedures. The first three levels, 0, .05, and .10, represent a zero, small-to-medium, and large DIF which are typically observed in a real CAT data set. These same three levels were also manipulated in Nandakumar and Roussos' (2004) simulation study. In addition, the level of .15 was added to the third factor in the current study in order to evaluate the accuracy of the two procedures under an exceptionally large level of DIF.

Factor 4. *a*- and *b*- parameters (∅; six levels). Following Nandakumar and Roussos (2004), six levels of *a*- and *b*- parameters (*a*, *b*) were manipulated in this study, including (0.4, -1.5), (0.4, 1.5), (0.8, 0), (1,-1.5), (1.4, -1.5), and (1.4, 1.5). These items are considered to represent the typical items found in an operational testing situation. For the *a*-parameters, the value of 0.4 indicates a poor item discriminating power, 0.8 and 1 represent medium levels of item discrimination, and 1.4 implies a good level of item discrimination. For the *b*-parameters, the value of -1.5 indicates an easy item, the value of 0 refers to a medium difficulty item, and the value of 1.5 implies a difficult item.

The six levels of the *a*- and *b*-parameters were then factorially combined with the four levels of DIF values to produce 24 items. Because there is no DIF for items 1 to 6, the *b*-parameters (i.e., the required ability level in order to have a 50% for a correct response) for the reference and focal groups are identical. These *b*-parameter values correspond to -1.5, 1.5, 0, -1.5, -1.5, and 1.5. Items 7 to 24 display a specific level of DIF, where the *b*-parameters differ for the two groups (see Table 1). Generally speaking, the reference group examinees contain smaller *b*-parameters than the focal group examinees because the former group is expected to possess a lower ability level than the latter group in order to have the same 50% probability for a correct response. When these *b*-parameters are weighted by the sample sizes and impact levels of the reference and focal groups (denoted as b_w), they can be converted back to the values of (-1.5, 1.5, 0, -1.5, -1.5, 1.5) as those in Items 1 – 6, for each manipulated level of β, respectively.

To summarize, the four factors in our simulation study were factorially combined to produce a design with $3 \times 3 \times 4 \times 6 = 216$ conditions. Each condition was replicated 100 times to evaluate the Type I error and power rates of the CATSIB-G and CATSIB-E procedures.

Simulation Procedure

The simulation study for CATSIB-E and CATSIB-G is presented in two stages. The first stage outlines the requirements for CAT in generating examinees' responses, thereby providing the examinees' theta ability estimates as well as their information scores. This stage mimics an operational testing scenario where examinees are administered a fixed-length CAT composed of operational items with well-established and DIF-free item parameters as well as a specific number of pre-test items for research purposes. The second stage is used to calculate and evaluate DIF for pre-test items using CATSIB-E and CATSIB-G.

1. The CAT stage. For each factorial combination of the sample sizes (N_r and N_f) and impact levels (d_t), the reference group and focal group test takers were simulated from their respective distributions, thereby producing the true ability scores for the reference and the focal groups, respectively. As in Nandakumar and Roussos (2004), each examinee was adaptively administered a fixed length of 25 items from an item bank with 1000 operational items. Given the true ability score (θ) of an examinee, the probability of a correct response for the first item (i.e., $i = 1$) can be estimated by the 3-parameter IRT model,

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-1.7a_i(\theta - b_i)}}, \qquad (12)$$

where a_i is the discrimination parameter for item i, b_i is the difficulty parameter for item i, c_i is the guessing parameter for item i, and e is the exponent function. The a-, b-, and c-parameters of the 1,000 operational items were generated identically as those in Nandakumar and Roussos' study[1]. With (12), a response (either correct or incorrect) to item i can be generated from a Bernoulli distribution with $p = P_i(\theta)$ and $q = 1 - p$. Given this response, the CAT program can provide a temporary estimate of the theta ability for that examinee ($\hat{\theta}_t$) via

[1]Following Nandakumar and Roussos (2004), b ~ normal (0, 1) with range $-3 \leq b \leq 3$. For $b \leq -1$, $\log(a)$ ~ normal (-.357, .25) with range $.4 \leq a \leq 1.1$. For $b > -1$, $\log(a)$ ~ normal (-.223, .34) with range $.4 \leq a \leq 1.7$. c ~ uniform (.12, .22). These item-parameter levels are typical in an operational testing situation.

the standard Maximum Likelihood Estimation (MLE) procedure (e.g., Hambleton, Swaminathan, & Rogers, 1991, p. 33 – 36).

The stage 1 process began with 1000 items in the bank using an adaptive testing environment. An examinee's information estimate conditional on the temporary theta ($\hat{\theta}_t$) was then computed, which can be expressed as

$$I_i(\hat{\theta}_t) = \frac{(1.7a_i)^2(1-c_i)}{\{c_i + e^{[1.7a_i(\hat{\theta}_t - b_i)]}\}\{1 + e^{[-1.7a_i(\hat{\theta}_t - b_i)]}\}^2}. \quad (13)$$

Statistically speaking, the item that yields the largest information should be selected adaptively as the next item. Practically speaking, this approach may lead to the problem that some items are overexposed. Thus, we used an exposure method that was proposed by Kingsbury and Zara (1989) and was used in Nandakumar and Roussos (2004). Specifically, the first item administered was randomly selected from the 10 items associated with the highest information values at $\hat{\theta}_t = 0$. The value of 0 refers to the best estimate of examinee's ability level because we do not have any information about the examinee at the beginning. The second item was then randomly selected from the 9 best items at the new estimate of $\hat{\theta}_t$. The third item, likewise, was randomly chosen from the 8 best items at the new estimate of $\hat{\theta}_t$, and so on; until starting from the 10th item, the item with the highest information was selected. Eventually, the CAT procedure produces a final theta ability estimate ($\hat{\theta}$) for each examinee.

The above procedure represents a prototypical CAT procedure often found in CAT algorithm packages. As noted above, the purpose of the CAT procedure is to provide two statistics—the theta ability estimate for each examinee ($\hat{\theta}$) and the standard error estimate for each examinee (σ_{e_j})—which are required for the CATSIB procedures. After obtaining these estimates, we can proceed to the SIB stage.

2. The SIB stage. Given the two essential estimates ($\hat{\theta}$ and $\sigma_{\bar{e}}^2{}_G$ or $\sigma_{\bar{e}}^2{}_j$), the remaining SIB steps follow those in (3) to (11), thereby producing the 95% CIs for $\hat{\beta}$ in CATSIB-G and CATSIB-E, respectively. As noted above, one important issue is to determine the number of intervals (*K*) required for Equations (5) and (6). In this study, we followed Nandakumar and Roussos' (2004) study that CATSIB-G and CATSIB-E were programmed to begin with an arbitrary large number of ability intervals (i.e., 80). If an interval contained less than 3 examinees in either the reference of focal group, then that interval would be thrown out because it was a "sparse bin". If more than 7.5% of either the reference or focal group examinees were eliminated, then the program would automatically reduce the number of intervals by 1 until the number of eliminated examinees from each group became less than or equal to 7.5%. To avoid the number of intervals becoming 1 (i.e. no bin left for conducting a reasonable CATSIB analysis), the minimum number of intervals was set to 10. Using this constraint, the number of eliminated examinees might exceed 7.5%, but only in rare situations.

Evaluation Criteria

Two types of statistics—Type I error and power—were used to evaluate the precision of the $\hat{\beta}$ for the CATSIB-G and CATSIB-E procedures.

1. Type I error. Type I error refers to the probability that either CATSIB-G or CATSIB-E will incorrectly identify an item as displaying DIF when, in fact, it does not. We call this concept a "false alarm", meaning that a non-DIF item is falsely signaled as a problematic DIF item. For each manipulated condition, the mean Type I error rates of Items 1 - 6 were calculated across 100 replications for CATSIB-G and CATSIB-E, respectively (denoted as $\overline{\alpha_G}$ and $\overline{\alpha_E}$). Because 95% CI was used in this study, the nominal level of Type I error rate should be 5%, meaning that at most 5 out of 100 replicated CIs is expected not to span the true β value of 0.

2. Power. Power refers to the probability of correctly detecting a DIF item. We call this concept a "correct decision" because the procedure can identify DIF items correctly. In this study, Items 7 – 24 represent different levels of DIF items. Hence, for each manipulated condition, the mean power rates of items 7 – 24 yielded by CATSIB-G and CATSIB-E were calculated across 100 replications, respectively (denoted as $\overline{PW_G}$ and $\overline{PW_E}$). Generally speaking, we expect the power rate of an item to be as large as possible. However, due to different magnitudes of manipulated DIF levels, the latter items (e.g., Items 19 – 24 β = .15) are expected to possess larger power than the former items (e.g., Items 7 – 14 with β = .05).

It is generally not feasible to specify the true power rate for each item, and hence we did not use a nominal level of power to compare CATSIB-G and CATSIB-E as in the case of Type I error. Instead, it is desirable if an item possess a high power rate. To determine this rate, we used two comparative statistics: bias and relative bias. Bias is defined as the difference between $\overline{PW_E}$ and $\overline{PW_G}$ (i.e., $\text{bias} = \overline{PW_E} - \overline{PW_G}$). A positive bias means that CATSIB-E produces a higher power rate than CATSIB-G, meaning that CATSIB-E improves the performance of the conventional CATSIB-G. Relative bias is defined as the bias relative to PW_G (i.e., $\text{rbias} = \text{bias}/\overline{PW_G}$). This statistic demonstrates the power rate improved by CATSIB-E relative to the original power rate yielded by CATSIB-G.

Results

1. Type I error. Table 2 shows the Type I error rates for CATSIB-G and CATSIB-E as well as the Type I error rates reported in Nandakumar and Roussos (2004, Table 4) for comparative purposes (called CATSIB-NR). By and large, the Type I error rates for Items 1 - 6 were similar across the three procedures. The rates ranged from 3% to 8% with a mean of 5.2% for CATSIB-NR, ranged from 2% to 9% with a mean of 5.8% for CATSIB-G, and ranged from 1% to 9% with a mean of 5.9% for CATSIB-E. These rates were slightly larger than the nominal level of 5%. Considering the effects of the manipulated factors, enlarging

impact level (d_t) and sample size (N) tended to increase the Type I error rate to a small degree, as shown in Table 3. To summarize, the Type I error rates were close to the nominal level for both CATSIB-G and CATSIB-E, and they were also comparable with the results presented in Nandakumar and Roussos' study.

2. Power. While the Type I error rates were comparable between the CATSIB-G and CATSIB-E, the power rates were not. The mean power rates for CATSIB-G and CATSIB-E as well as their biases and relative biases are shown in Table 4. Note that the results reported in Nandakumar and Roussos (2004) were excluded because the current study had more manipulated conditions, thereby making a direct comparison less feasible. Also, our CATSIB-G procedure replicated Nandakumar and Roussos' CATSIB method and, therefore, our CATSIB-G results can serve as a point of comparison.

In general, increasing the DIF levels tended to increase the power rates for both CATSIB-G and CATSIB-E. To further evaluate the consequence of the manipulated factors on power, we separate Table 4 into two summary tables. First, Table 5 presents the mean biases (i.e., B_m) and mean relative biases (i.e., RB_m) for three levels of sample size (N) by three amounts of impact level (d_t) by three amounts of DIF level (β), while holding the a- and b-parameters constant. This reporting structure is used to evaluate how the 27 factorial conditions influence the CATSIB-G and CATSIB-E procedures in detecting DIF items at the pre-test stage. Second, Table 6 presents the same statistics for the manipulated a- and b-parameter conditions, while holding those 27 factorial conditions constant. The purpose of this reporting structure is to evaluate which level of item discrimination or item difficulty produces better power rates by either CATSIB-E or CATSIB-G.

As presented in Table 5, three outcomes that affect the B_m and RB_m results were observed. First, when $d_t = 0$, there was no obvious difference in power rates between CATSIB-E and CATSIB-G. But when d_t was increased to .50 and to .10, CATSIB-E tended

to produce higher power rates than CATSIB-G, especially when the DIF level was small (i.e., $\beta = .05$). Second, decreasing β (DIF) levels were more likely to produce larger B_m and RB_m, when $d_t \geq .50$. These two outcomes suggest that CATSIB-G is less accurate or robust to some adverse conditions, including a small DIF level and a large amount of impact level. These conditions are generally expected to contaminate the performance of the SIBTEST algorithm. By contrast, the CATSIB-E procedure always yields higher power rates, meaning that this procedure is a more accurate at identifying DIF items. Third, no obvious effect of sample size on B_m and RB_m was observed.

As noted above, the results in Table 6 were used to evaluate the *a*- and *b*-parameter effects on B_m and RB_m. Three outcomes should be noted. First, we found that the power rates for CATSIB-E (i.e., ranging from 59.8% to 91.7% with a mean of 73.8%) were consistently higher than for CATSIB-G (i.e., ranging from 55.9% to 87.4% with a mean of 68.6%), thereby suggesting that CATSIB-E is more powerful than CATSIB-G across different manipulated levels of the *a*- and *b*-parameters. Second, because B_m only measured the power difference (i.e., $\overline{PW_E} - \overline{PW_G}$) without considering the improvement relative to the original $\overline{PW_G}$, we also computed RB_m. Specifically, we observed that a higher level of *b*- (or item difficulty) parameter tended to produce a larger RB_m. In our findings, when *b* = -1.5, RB_m were 7.3%, 8.3%, and 7.5%, for *a* = 0.4, 1, and 1.4, respectively. In comparison, when *b* increased to 0 or 1.5, RB_m became 16.3%, 19.0% and 45.8%, respectively. This outcome suggests that CATSIB-E yielded higher power rates than CATSIB-G across all *b* levels, and the improvement was more substantial as *b* became larger. Third, no obvious pattern was found across the levels of *a*- (or item discrimination) parameter on B_m and RB_m, when *b* was held constant.

Conclusions and Discussion

Computerized adaptive testing (CAT) is becoming more prevalent in operational testing programs and, thus, more important to students and other test takers. It has many advantages over the traditional paper-and-pencil testing (PPT). Although considerable progress has been made in the development of CAT, a constant concern in traditional PPT, DIF, is also an important concern in CAT. Hence, Nandakumar and Roussos (2004) developed a CATSIB procedure for detecting DIF items at the pre-test stage, thereby improving the quality of experimental items before they are used in a real CAT situation.

And while their CATSIB procedure provided an important way to detect DIF items, it is not free from limitations. One important drawback is that their regression correction formula is based on the group-level reliability, meaning that the reliability is assumed to be a constant for all examinees within the same group. In fact, Raju et al. (2007) addressed this limitation explicitly, and proposed a formula for the examinee-level reliability tailored for each examinee in a CAT setting.

The purpose of the present study is, therefore, to evaluate the performance of the combined procedure (i.e., CATSIB-E) for detecting DIF items, and the improvement of the CATSIB-E procedure relative to the conventional CATSIB-G procedure. Our simulation results show that, while maintaining a comparable level of Type I error, CATSIB-E provides consistently better power performances than CATSIB-G, especially when the testing situations are adverse for DIF detection, including a small DIF level (i.e., β), a large impact level (i.e., d_t), and a large b-parameter (i.e., a more difficult item). Thus, CATSIB-E provides an effective and useful method for detecting DIF items in a pre-test setting.

Directions to Future Research

A first area for future research is related to the application of the examinee-level reliability formula to other CAT situations. The present study demonstrated that the CATSIB procedure can be improved by using the examinee-level reliability for the regression

correction formula in Shealey and Stout's (1993) SIBTEST algorithm in a pre-test setting. As an extension, the same improvement is expected to occur if this modified SIBTEST algorithm is used for detecting DIF levels of the operational items that are administered to test takers in a real CAT situation or in a post-test setting. Therefore, more research is required to investigate the improvement of the SIBTEST algorithm based on the examinee-level reliability in other CAT scenarios.

A second area for future research resides directly with the choice of the item parameters used at the pre-test stage. The present study sought to follow the *a*- and *b*-parameters of the operational items in LSAT that were reported in Nandakumar and Roussos (2004). The sample size levels were also identical to those in Nandakumar and Roussos. Hence, future studies can explore the effect of different *a*- and *b*- parameters which are found in standardized tests other than the LSAT, and to investigate whether a larger sample size can further improve the performance of the CATSIB-E procedure.

Acknowledgement

This project was completed with funds provided to the second author by the Medical Council of Canada. We would like to thank the Medical Council of Canada for their support. However, the authors are solely responsible for the methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of the Medical Council of Canada.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Chang, S.-R., Plake, B. S., Kramer, G. A., & Lien, S.-M. (2011). Development and application of detection indices for measuring guessing behaviors and test-taking effort in computerized adaptive testing. *Educational and Psychological Measurement, 71*, 437 – 459.

Cui, Y., & Li, J. C.-H. (2010l). *Building diagnostic computerized adaptive testing: issues and challenges*. Paper presented at the annual meeting of the Canadian Society for the Study of Education, Montreal, Canada.

Frey A., & Seitz, N.-N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the programme for international student assessment. *Educational and Psychological Measurement, 71*, 503 – 522.

Hambleton, R. K., Swaminathan, H, & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Kingsbury, G. G., & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2,* 359-375.

LSAT. (2011). *About the LSAT*. Retrieved from http://www.lsac.org/jd/LSAT/about-the-LSAT.asp

Nandakumar, R., & Roussos, L. A. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests* (Computerized Testing Report No. 97-11). Newtown, PA: Law School Admission Council.

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

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional *SEM*: A case for conditional reliability. *Applied Psychological Measurement, 31,* 169 – 180.

Roussos, L. A., Schnipke, D. L., & Pashley, P. J. (1999). A generalized formula for the Mantel-Haenszel differential item functioning parameter. *Journal of Educational and Behavioral Statistics, 24,* 293-322.

Segall, D. O. (2005). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), Encyclopedia of social measurement. New York, NY: Academic Press.

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

Zwick, R., Thayer, D. T., & Wingersky, M. (1994). 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. (1995). Effect of Rasch calibration on ability and DIF estimation in computer adaptive tests. *Journal of Educational Measurement, 32,* 341-363.

Table 1

The Manipulated Values of a- and b-Parameters for Items 7 to 24

				$d_t = 0$		$d_t = 0.5$		$d_t = 1$	
	Item	β	a	b_r	b_f	b_r	b_f	b_r	b_f
$N_r=N_f$	7	0.05	0.4	-1.738	-1.262	-1.739	-1.261	-1.741	-1.259
	8	0.05	0.4	1.262	1.738	1.261	1.739	1.259	1.741
	9	0.05	0.8	-0.12	0.12	-0.122	0.122	-0.129	0.127
	10	0.05	1	-1.691	-1.309	-1.691	-1.309	-1.689	-1.311
	11	0.05	1.4	-1.697	-1.303	-1.695	-1.305	-1.690	-1.310
	12	0.05	1.4	1.303	1.697	1.305	1.695	1.310	1.690
	13	0.1	0.4	-1.977	-1.023	-1.978	-1.022	-1.983	-1.017
	14	0.1	0.4	1.023	1.977	1.022	1.978	1.017	1.983
	15	0.1	0.8	-0.241	0.241	-0.244	0.244	-0.254	0.254
	16	0.1	1	-1.882	-1.118	-1.881	-1.119	-1.878	-1.122
	17	0.1	1.4	-1.891	-1.109	-1.887	-1.113	-1.878	-1.122
	18	0.1	1.4	1.109	1.891	1.113	1.887	1.122	1.878
	19	0.15	0.4	-2.218	-0.782	-2.219	-0.781	-2.227	-0.773
	20	0.15	0.4	0.782	2.218	0.781	2.219	0.773	2.227
	21	0.15	0.8	-0.363	0.363	-0.367	0.367	-0.38	0.38
	22	0.15	1	-2.071	-0.929	-2.069	-0.931	-2.077	-0.923
	23	0.15	1.4	-2.081	-0.919	-2.075	-0.925	-2.064	-0.936
	24	0.15	1.4	0.919	2.081	0.925	2.075	0.936	2.064
$N_r = 2N_f$	7	0.05	0.4	-1.656	-1.188	-1.656	-1.118	-1.657	-1.186
	8	0.05	0.4	1.338	1.824	1.338	1.824	1.337	1.826
	9	0.05	0.8	-0.08	0.16	-0.081	0.162	-0.084	0.168
	10	0.05	1	-1.622	-1.256	-1.622	-1.256	-1.622	-1.256
	11	0.05	1.4	-1.782	-1.359	-1.778	-1.361	-1.766	-1.367
	12	0.05	1.4	1.359	1.782	1.361	1.778	1.367	1.766
	13	0.05	0.4	-1.806	-0.888	-1.807	-0.886	-1.81	-0.88
	14	0.05	0.4	1.168	2.164	1.167	2.166	1.166	2.168
	15	0.05	0.8	-0.161	0.322	-0.163	0.326	-0.168	0.336
	16	0.05	1	-1.734	-1.032	-1.734	-1.032	-1.736	-1.028
	17	0.05	1.4	-2.104	-1.198	-2.092	-1.204	-2.066	-1.217
	18	0.05	1.4	1.198	2.104	1.204	2.092	1.217	2.066
	19	0.15	0.4	-1.979	-0.543	-1.979	-0.541	-1.985	-0.531
	20	0.15	0.4	1.021	2.457	1.021	2.459	1.015	2.469
	21	0.15	0.8	-0.242	0.484	-0.245	0.489	-0.253	0.507
	22	0.15	1	-1.881	-0.739	-1.879	-0.741	-1.885	-0.731
	23	0.15	1.4	-2.275	-1.113	-2.267	-1.117	-2.252	-1.124
	24	0.15	1.4	1.113	2.275	1.117	2.267	1.124	2.252

Note: d_t is the impact level. β is the DIF level, a is the a-parameter, b_r (or b_f) is the b-parameter for the reference (or focal) group. N_r (or N_f) is the reference (or focal) group size.

Table 2

Mean of the Type I Error Rates Yielded by CATSIB-G and CATSIB-E for Items 1 to 6

				N = (250, 250)								
				d_t = 0			d_t = 0.5			d_t = 1		
Item	β	a	b	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$
1	0	0.4	-1.5	4	2	3	5	7	8	7	7	7
2	0	0.4	1.5	5	6	3	6	8	6	5	8	5
3	0	0.8	0	5	6	6	5	2	3	7	2	4
4	0	1	-1.5	4	7	7	5	4	6	6	8	7
5	0	1.4	-1.5	5	2	4	3	7	5	5	5	7
6	0	1.4	1.5	7	5	7	4	6	7	5	8	9
				N = (500, 250)								
				d_t = 0			d_t = 0.5			d_t = 1		
	β	a	b	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$
1	0	0.4	-1.5	3	6	6	5	8	8	6	6	9
2	0	0.4	1.5	5	6	4	4	6	5	5	8	6
3	0	0.8	0	4	3	1	4	7	6	5	4	5
4	0	1	-1.5	7	5	5	6	5	7	5	5	5
5	0	1.4	-1.5	3	3	3	4	7	2	4	7	7
6	0	1.4	1.5	6	7	8	4	5	8	6	6	8
				N = (500, 500)								
				d_t = 0			d_t = 0.5			d_t = 1		
	β	a	b	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$	$\overline{\alpha_{NR}}$	$\overline{\alpha_G}$	$\overline{\alpha_E}$
1	0	0.4	-1.5	3	6	6	5	4	5	7	8	8
2	0	0.4	1.5	4	4	5	4	8	7	4	3	6
3	0	0.8	0	5	6	6	8	9	8	5	8	6
4	0	1	-1.5	7	8	8	8	5	7	8	6	7
5	0	1.4	-1.5	4	4	5	5	7	4	7	6	6
6	0	1.4	1.5	8	7	5	6	6	8	6	4	7

Note: N is the total sample size of all examinees. The first value in the bracket is the reference group size, and the second value is the focal group size. d_t is the impact level. β is the DIF level, a is the a-parameter, b is the b-parameter. $\overline{\alpha_{NR}}$ is the mean of the Type I error rates reported in Nandakumar and Roussos (2004). $\overline{\alpha_G}$ indicates the mean of the Type I error rates yielded by CATSIB-G in the present study. $\overline{\alpha_E}$ represents the mean of the Type I error rates produced by CATSIB-E in the present study.

Table 3

Mean of the Type I Error Rates for the 9 Manipulated Conditions: Sample Size by Impact Level

		$d_t = 0$	$d_t = 0.5$	$d_t = 1$
N = (250, 250)	$\overline{\alpha_{NR}}$	5.0	4.7	5.8
	$\overline{\alpha_G}$	4.7	5.7	6.3
	$\overline{\alpha_E}$	5.0	5.8	6.5
N = (500, 250)	$\overline{\alpha_{NR}}$	4.7	4.5	5.2
	$\overline{\alpha_G}$	5.0	6.3	6.0
	$\overline{\alpha_E}$	4.5	6.0	6.7
N = (500, 500)	$\overline{\alpha_{NR}}$	5.2	6.0	6.2
	$\overline{\alpha_G}$	5.8	6.5	5.8
	$\overline{\alpha_E}$	5.8	6.5	6.7

Note: N represents the total sample size of all examinees. The first value in the bracket is the reference group size, and the second value is the focal group size. d_t is the impact level. $\overline{\alpha_{NR}}$ is the mean of the Type I error rates reported in Nandakumar and Roussos (2004). $\overline{\alpha_G}$ indicates the mean of the Type I error rates yielded by CATSIB-G in the present study. $\overline{\alpha_E}$ represents the mean of the Type I error rates produced by CATSIB-E in the present study.

Table 4

Mean of the Power Rates, Biases and Relative Biases for Items 7 to 24

				N = (250, 250)											
				d_t = 0				d_t = 0.5				d_t = 1			
Item	β	a	b_w	$\overline{PW_G}$	$\overline{PW_E}$	B	RB	$\overline{PW_G}$	$\overline{PW_E}$	B	RB	$\overline{PW_G}$	$\overline{PW_E}$	B	RB
7	.05	0.4	-1.5	29	30	1	3.4	27	24	-3	-11.1	22	27	5	22.7
8	.05	0.4	1.5	15	15	0	0.0	10	11	1	10.0	13	19	6	46.2
9	.05	0.8	0	23	20	-3	-13.0	10	18	8	80.0	15	25	10	66.7
10	.05	1	-1.5	53	53	0	0.0	40	47	7	17.5	32	52	20	62.5
11	.05	1.4	-1.5	56	56	0	0.0	51	66	15	29.4	40	67	27	67.5
12	.05	1.4	1.5	18	16	-2	-11.1	12	26	14	116.7	9	30	21	233.3
13	.10	0.4	-1.5	76	79	3	3.9	64	69	5	7.8	55	64	9	16.4
14	.10	0.4	1.5	51	56	5	9.8	53	59	6	11.3	27	42	15	55.6
15	.10	0.8	0	60	58	-2	-3.3	46	51	5	10.9	40	49	9	22.5
16	.10	1	-1.5	92	95	3	3.3	88	93	5	5.7	82	91	9	11.0
17	.10	1.4	-1.5	100	99	-1	-1.0	93	97	4	4.3	88	95	7	8.0
18	.10	1.4	1.5	65	65	0	0.0	40	53	13	32.5	41	69	28	68.3
19	.15	0.4	-1.5	99	99	0	0.0	91	94	3	3.3	89	91	2	2.2
20	.15	0.4	1.5	89	88	-1	-1.1	79	80	1	1.3	72	77	5	6.9
21	.15	0.8	0	97	93	-4	-4.1	88	93	5	5.7	86	86	0	0.0
22	.15	1	-1.5	100	100	0	0.0	100	99	-1	-1.0	97	100	3	3.1
23	.15	1.4	-1.5	100	100	0	0.0	99	99	0	0.0	97	97	0	0.0
24	.15	1.4	1.5	85	88	3	3.5	81	87	6	7.4	77	88	11	14.3

Table 4 (Continued)

				N = (500, 250)											
				d_t = 0				d_t = 0.5				d_t = 1			
Item	β	a	b_w	$\overline{PW_G}$	$\overline{PW_E}$	B	RB	$\overline{PW_G}$	$\overline{PW_E}$	B	RB	$\overline{PW_G}$	$\overline{PW_E}$	B	RB
7	.05	0.4	-1.5	32	29	-3	-9.4	41	45	4	9.8	35	48	13	37.1
8	.05	0.4	1.5	28	27	-1	-3.6	25	26	1	4.0	15	31	16	106.7
9	.05	0.8	0	28	27	-1	-3.6	19	26	7	36.8	39	55	16	41.0
10	.05	1	-1.5	54	55	1	1.9	61	64	3	4.9	69	89	20	29.0
11	.05	1.4	-1.5	76	77	1	1.3	75	88	13	17.3	78	93	15	19.2
12	.05	1.4	1.5	22	23	1	4.5	21	29	8	38.1	11	35	24	218.2
13	.10	0.4	-1.5	88	85	-3	-3.4	83	80	-3	-3.6	78	88	10	12.8
14	.10	0.4	1.5	66	66	0	0.0	67	64	-3	-4.5	59	63	4	6.8
15	.10	0.8	0	76	78	2	2.6	72	81	9	12.5	71	83	12	16.9
16	.10	1	-1.5	97	98	1	1.0	96	99	3	3.1	98	99	1	1.0
17	.10	1.4	-1.5	100	100	0	0.0	100	99	-1	-1.0	100	100	0	0.0
18	.10	1.4	1.5	72	79	7	9.7	54	72	18	33.3	61	81	20	32.8
19	.15	0.4	-1.5	98	99	1	1.0	99	99	0	0.0	99	100	1	1.0
20	.15	0.4	1.5	94	94	0	0.0	93	94	1	1.1	86	90	4	4.7
21	.15	0.8	0	99	99	0	0.0	96	97	1	1.0	97	99	2	2.1
22	.15	1	-1.5	100	100	0	0.0	100	100	0	0.0	100	100	0	0.0
23	.15	1.4	-1.5	100	100	0	0.0	100	100	0	0.0	100	100	0	0.0
24	.15	1.4	1.5	89	92	3	3.4	89	96	7	7.9	78	85	7	9.0

Table 4 (Continued)

				N = (500, 500)											
				d_t = 0				d_t = 0.5				d_t = 1			
Item	β	a	b_w	$\overline{PW_G}$	$\overline{PW_E}$	B	RB	$\overline{PW_G}$	$\overline{PW_E}$	B	RB	$\overline{PW_G}$	$\overline{PW_E}$	B	RB
7	.05	0.4	-1.5	45	45	0	0.0	38	47	9	23.7	26	42	16	61.5
8	.05	0.4	1.5	30	28	-2	-6.7	30	30	0	0.0	11	28	17	154.5
9	.05	0.8	0	43	44	1	2.3	24	34	10	41.7	12	31	19	158.3
10	.05	1	-1.5	73	75	2	2.7	53	64	11	20.8	49	73	24	49.0
11	.05	1.4	-1.5	85	85	0	0.0	66	81	15	22.7	60	81	21	35.0
12	.05	1.4	1.5	25	28	3	12.0	17	38	21	123.5	14	45	31	221.4
13	.10	0.4	-1.5	94	91	-3	-3.2	83	89	6	7.2	74	81	7	9.5
14	.10	0.4	1.5	85	87	2	2.4	69	78	9	13.0	60	69	9	15.0
15	.10	0.8	0	92	88	-4	-4.3	74	75	1	1.4	58	77	19	32.8
16	.10	1	-1.5	100	100	0	0.0	98	100	2	2.0	91	98	7	7.7
17	.10	1.4	-1.5	100	100	0	0.0	98	99	1	1.0	98	98	0	0.0
18	.10	1.4	1.5	90	91	1	1.1	77	91	14	18.2	67	90	23	34.3
19	.15	0.4	-1.5	99	100	1	1.0	99	100	1	1.0	96	99	3	3.1
20	.15	0.4	1.5	100	100	0	0.0	98	99	1	1.0	86	92	6	7.0
21	.15	0.8	0	100	100	0	0.0	100	100	0	0.0	89	95	6	6.7
22	.15	1	-1.5	100	100	0	0.0	100	100	0	0.0	100	100	0	0.0
23	.15	1.4	-1.5	100	100	0	0.0	100	100	0	0.0	100	99	-1	-1.0
24	.15	1.4	1.5	100	100	0	0.0	99	100	1	1.0	96	98	2	2.1

Note: N is the total sample size of all examinees. The first value in the bracket is the reference group size, and the second value is the focal group size. d_t is the impact level. β is the DIF level, a is the a- parameter, b_w is the b-parameter weighted by sample sizes and impact levels. $\overline{PW_G}$ is the mean of the power rates yielded by CATSIB-G. $\overline{PW_E}$ is the mean of the power rates yielded by CATSIB-E. B is the bias. RB is the relative bias.

Table 5

Means of the Power Biases and Relative Biases for the 27 Manipulated Conditions: Sample Size by Impact Level by DIF Level

		$d_t = 0$		$d_t = 0.5$		$d_t = 1$	
N	β	B_m	RB_m	B_m	RB_m	B_m	RB_m
(250, 250)	.05	-1	-3.5	7	40.4	15	83.1
	.10	1	2.1	6	12.1	13	30.3
	.15	0	-0.3	2	2.8	4	4.4
		$d_t = 0$		$d_t = 0.5$		$d_t = 1$	
N	β	B_m	RB_m	B_m	RB_m	B_m	RB_m
(500, 250)	.05	0	-1.5	6	18.5	17	75.2
	.10	1	1.7	4	6.6	8	11.7
	.15	1	0.7	2	1.7	2	2.8
		$d_t = 0$		$d_t = 0.5$		$d_t = 1$	
N	β	B_m	RB_m	B_m	RB_m	B_m	RB_m
(500, 500)	.05	1	1.7	11	38.7	21	113.3
	.10	-1	-0.7	6	7.1	11	16.5
	.15	0	0.2	1	0.5	3	3

Note: N is the total sample size of all examinees. The first value in the bracket is the reference group size, and the second value is the focal group size. d_t is the impact level. β is the DIF level. B_m is the mean of the biases for each of the 27 conditions. RB_m is the mean of the relative biases for each of the 27 conditions.

Table 6

Means of the Power Rates, Biases and Relative Biases for the Manipulated a- and b-Parameter Conditions

a	b_w	$\overline{PW_G}$	$\overline{PW_E}$	B_m	RB_m
0.4	-1.5	68.9	72.0	3.1	7.3
0.4	1.5	56.0	59.8	3.8	16.3
0.8	0	61.3	66.0	4.7	19.0
1	-1.5	82.3	86.8	4.5	8.3
1.4	-1.5	87.4	91.7	4.3	7.5
1.4	1.5	55.9	66.5	10.6	45.8

Note: a is the *a*-parameter, b_w is the *b*-parameter weighted by the sample sizes and impact levels. $\overline{PW_G}$ is the mean of the power rates yielded by CATSIB-G. $\overline{PW_E}$ is the mean of the power rates produced by CATSIB-E. B_m is the mean of the biases for each of the 6 conditions. RB_m is the mean of the relative biases for each of the 6 conditions.