JOURNAL OF APPLIED MEASUREMENT, *25*(1/2), 12-25
Copyright© 2025

Assessing Differential Item Functioning in Computerized Adaptive Testing

Ching-Lin Shih
National Sun Yat-sen University

Kuan-Yu Jin
Hong Kong Examinations and Assessment Authority

Chia-Ling Hsu
Hong Kong Examinations and Assessment Authority

To implement computerized adaptive testing (CAT), monitoring the parameter stability of operational items and checking the quality of newly written items are critical. In particular, assessing differential item functioning (DIF) is a vital step in ensuring test fairness and improving test reliability and validity. This study investigated the performance in CAT of several nonparametric DIF assessment methods, the odds ratio (OR; Jin et al., 2018) approach, modified Mantel–Haenszel method (Zwick et al., 1994a, 1994b; Zwick & Thayer, 2002), modified logistic regression method (Lei et al., 2006), and CAT version of the simultaneous item bias test (SIBTEST) method (Nandakumar & Roussos, 2001), via a series of simulation studies. The results showed that the OR outperformed the other three methods in controlling false positive rates and producing high true positive rates when there were many DIF items in a test. Moreover, combining the OR with a scale purification procedure further improved DIF assessment in CAT as the percentage of DIF items exceeded 10%.

Keywords: differential item functioning, computerized adaptive test, scale purification, odds ratio, CATSIB

Requests for reprints should be sent to Chia-Ling Hsu, 68 Gillies Avenue South, Hung Hom, Kowloon, Hong Kong; email: clhsu@hkeaa.edu.hk

Computerized adaptive testing (CAT) has a long history of research and practical application in testing environments. CAT can reach comparable precision in latent trait estimation as its paper-and-pencil counterpart while using only half of the administered items required by the latter (van der Linden & Glas, 2000; Wainer et al., 2000; Weiss, 1982). Furthermore, CAT also exhibits an advantage in that it can yield analogous measurement precision for all examinees (Hsu & Wang, 2015; Hsu et al., 2013; Wang et al., 2019). These advantageous properties of CAT are largely dependent on a well-calibrated item pool and its corresponding common scale. To this end, an essential consideration in the practical implementation of CAT is the routine monitoring of item pool quality. This procedure is critical for preserving the integrity and functionality of the item pool. However, despite its importance, issues related to item pool quality have received comparatively less attention in the literature than other key components of CAT (e.g., item selection algorithms or termination rules). Therefore, this study investigated quality assurance concerns in CAT item pools, contributing to the enhancement of CAT's measurement efficiency and accuracy.

Among the procedures for maintaining a CAT item pool, assessing test fairness is critical for ensuring the appropriate use and application of test scores, as it significantly impacts test reliability and validity. Differential item functioning (DIF) refers to the difference in the probability of a correct response to an item between equally proficient examinees from various demographic groups (Pine, 1977; Zwick, 2000). When DIF items exist, identical test scores may be interpreted differently for examinees from various groups, thereby compromising test validity. To address this, item pools should be regularly reviewed and updated via a thorough examination of DIF. Particularly, implementing methods to assess DIF across crucial demographic variables for items is essential for ensuring item-level fairness in the design and management of a CAT item pool.

In the CAT field, DIF assessment can be performed on items that are adaptively administered for scoring purposes (Zwick, 2000) or on "seeded" items (i.e., pretest items) during the test administration process for calibration purposes (Lim & Choe, 2023). This study focuses on the former approach. Assessing DIF in items during CAT administration is especially important for several key reasons. First, since CAT functions as a continuous and dynamic testing system, keeping the item pool updated and well-calibrated is crucial for sustaining its long-term effectiveness. Second, unlike common DIF assessment practices in non-adaptive (i.e., linear) test scenarios, identifying DIF items in a CAT environment is especially complicated due to the need to find appropriate matching variables. Specifically, to approximate examinees' abilities, number-right scores (e.g., raw total scores) are typically used as matching variables in traditional DIF assessment methods (Holland & Thayer, 1988), which are unsuitable in CAT contexts. Instead, using latent trait estimates is more appropriate as matching variables for implementing DIF assessments in CATs (Zwick, 2000).

Furthermore, because each examinee in a CAT receives fewer items, each item carries greater weight in estimating the examinee's latent traits compared to non-adaptive testing. As a result, any deficiencies in item quality can have a disproportionately large impact on ability estimation (Zwick et al., 1994). Additionally, the computerized nature of CAT administration may introduce other potential sources of DIF compared to traditional testing, such as varying levels of computer anxiety and familiarity, as well as individual preferences for computerized administration (Zwick, 2000). For these reasons, assessing DIF within the context of CAT is more complex than it might appear and remains an essential consideration in maintaining fairness and validity in adaptive testing.

Numerous DIF assessment methods have been proposed, yet most may not be applicable in the CAT context. In general, DIF assessment methods for CAT can be classified into item response theory (IRT)-based and non-IRT-based approaches. Within the IRT-based approach, an IRT model is first fitted to the data, and statistical indices (Lord, 1980; Thissen et al., 1988) are then computed to flag or identify DIF items based on person and item estimates. However, the need to deal with huge data matrices makes these methods inefficient from a practical standpoint. In contrast, the non-IRT-based approaches, such as the Mantel–Haenszel method (MH; Holland & Thayer, 1988), logistic regression method (LR; Swaminathan & Rogers, 1990), and simultaneous item bias test (SIBTEST) method (Shealy & Stout, 1993), are relatively straightforward to implement and demand minimal computation resources.

Zwick et al. (1994a, 1994b) and Zwick and Thayer (2002) introduced the modified MH procedure for CAT DIF assessment by replacing number-right scores with expected IRT true scores. In Zwick et al. (1994a, 1994b), the modified MH procedure yielded DIF assessment results comparable to those obtained through non-adaptive administration approaches, such as administering the entire item pool or number-right scores. Nonetheless, these findings were derived from an impractically small CAT item pool containing only 75 items, raising concerns regarding practical applicability. For instance, as the item pool size increases, the number of sparse bins (i.e., some expected IRT true score intervals without examinees) also increases, resulting in inflated Type I error rate in DIF assessment. To address this issue, Zwick and Thayer (2002) recommended expanding the size of each interval (e.g., intervals of two units in the integer score metric) for large item pools. However, the optimal configuration for determining the units of an interval remains an unresolved issue.

Lei et al. (2006) introduced the modified LR by substituting number-right scores with IRT-based estimates and the modified IRT likelihood ratio test by imputing responses to unadministered items in the CAT item pool. The modified LR method yielded inflated Type I error in DIF assessment when the three-parameter logistic IRT model was applied to the test data and a mean ability difference between groups was present (i.e., *impact*). Conversely, the modified IRT likelihood ratio test demonstrated satisfactory performance in assessing DIF in CAT environments. Although this approach employs imputation to accommodate sparse data matrices, it may magnify some information within the observed data, resulting in a distorted representation (Jin et al., 2018). Additionally, this approach requires the fitting of multiple models across all studied items, leading to computational burdens that, in turn, render it inefficient for operational administration and simulation studies, particularly when processing large data matrices.

Nandakumar and Roussos (2001) proposed an extended version of SIBTEST for CAT, known as CATSIB, which focuses on seeded items. Similar to SIBTEST, CATSIB applies a regression correction method to IRT-based estimates to mitigate bias caused by the presence of impact. Nevertheless, the performance of the CATSIB is highly dependent on CAT settings, particularly regarding its power in assessing DIF. Specifically, the CATSIB generally exhibits good Type I error control, and demonstrates greater power with large, equally sized group samples and higher levels of impact (Lim & Choe, 2023), while its power tends to decrease when larger impacts are present (Lei et al., 2006; Lim et al., 2022; Nandakumar & Roussos, 2004). This inconsistency may arise from challenges in determining the optimal number of IRT ability intervals for matching variables when employing the CATSIB in practice. For instance, intervals with fewer than three examinees in either the reference or focal group are excluded prior to the CATSIB computation. Moreover, the CATSIB assumes uniform estimation precision among all examinees' ability estimates within each group when applying the regression correction. If this assumption is not adequately addressed, such as in a small group sample, the regression correction may perform suboptimally, leading to biased CATSIB outcomes (Lim & Choe, 2023).

Lim and Choe (2023) extended the IRT-based residual DIF (RDIF; Lim et al., 2022) for CAT DIF assessment, focusing on seeded items. In comparison to the abovementioned assessment methods, the RDIF method does not require matching variables, as is necessary with the modified MH and CATSIB methods, nor does it involve separate item calibration for each group or multiple model fittings, as required by the modified IRT likelihood ratio test. The findings revealed that the RDIF method exhibited well-controlled Type I errors and slightly higher power than the CATSIB, particularly when seeded items were calibrated using iterative updates of prior weights and multiple expected-maximization cycles (Kim, 2006). Like other IRT-based DIF assessment methods, the RDIF method's performance depends on the precision of item calibration. This issue may be particularly significant in CAT operational items, where the actual number of responses may be much smaller. Furthermore, the effectiveness of the RDIF method may diminish when examinees' ability estimates lack precision (Lim & Choe, 2023; Lim et al., 2022). These limitations can occur in CAT applications, particularly when the item pool contains outdated or overexposed items with suboptimal properties.

Recently, Jin et al. (2018) proposed an outlier detection strategy to overcome the shortcomings of existing DIF assessment methods. Its concept is straightforward: by comparing the odds ratios (OR) between the reference and focal groups, items with an extreme OR value will be flagged as DIF. The OR method has several attractive features that make it feasible in the CAT context. First, its simple and intuitive operation minimizes the computational processes involved in DIF assessment. Hence, this method can be performed on all the items in a test relatively quickly, regardless of the data size. Second, it is robust with a sparse data matrix, which makes it suitable for CAT, wherein many item responses are missing at random by design (Chen et al., 2024). The performance of the OR method in assessing DIF in CAT is certainly of interest.

Therefore, in this study, we conducted simulations to compare the performance of the OR method with the DIF assessment approaches, including the modified MH, modified LR, and CATSIB methods. The RDIF method (Lim & Choe, 2023) was excluded from the comparison due to limited research concerning its application in CAT. Specifically, its performance may be influenced by the precision of calibrated items and/or its estimation method, with findings primarily based on CAT-seeded items. Furthermore, the simulation settings used in the RDIF research differ from those of this study, which focus on operational items.

The article is organized as follows. First, the procedures of the OR method are introduced. Next, the utilized DIF assessment methods (i.e., OR, modified MH, modified LR, and CATSIB) in this study are presented, followed by scale purification procedures. A series of simulation studies that were used to compare the performances of these methods are presented, and the results are then summarized. Finally, the findings and limitations are described, and the conclusions of this study are drawn.

DIF Assessment Procedures

Odds Ratio (OR) Method

For dichotomously scored item i, odds—calculated as the ratio of the number of examinees who answer the item correctly to the number of examinees who answer the item incorrectly—can be obtained separately for the reference and focal groups. The odds ratio is then calculated to check whether this value deviates significantly from a reference standard. Let $\hat{\lambda}_i$ be the logarithm of the odds ratio of success over failure on item i for the reference and focal groups as follows:

$$\hat{\lambda}_i = \log\left(\frac{n_{R1i}/n_{R0i}}{n_{F1i}/n_{F0i}}\right), \qquad (1)$$

where n_{R1i} and n_{R0i} refer to the number of examinees in the reference group who answer item *i* correctly and incorrectly, respectively; and n_{F1i} and n_{F0i} are the number of examinees in the focal group who answer item *i* correctly and incorrectly, respectively. In addition, the approximate standard error of $\hat{\lambda}_i$ is calculated as follows:

$$\sigma\left(\hat{\lambda}_i\right) = \left(n_{R1i}^{-1} + n_{R0i}^{-1} + n_{F1i}^{-1} + n_{F0i}^{-1}\right)^{1/2}. \tag{2}$$

When the data follows the Rasch model without DIF items, a λ statistic represents the estimate of the mean ability difference between the reference and focal groups (i.e., impact; Jin et al., 2018). For example, a positive λ implies that the reference group generally outperforms the focal group. As the λ values of all test items may vary widely due to randomness, the difference in the latent abilities between groups should be inferred from a measure of central tendency for all λ values. Considering that the sample mean $\bar{\hat{\lambda}}$ is sensitive to outliers (i.e., many DIF items), the sample median $\tilde{\hat{\lambda}}$ is adopted as a robust estimate of the group difference. Specifically, if $\hat{\lambda}_i \pm Z_{\frac{\alpha}{2}} \times \sigma\left(\hat{\lambda}_i\right)$ does not include $\tilde{\hat{\lambda}}$, item *i* will be deemed as exhibiting DIF. Since the OR method is robust to an incomplete data matrix (Chen et al., 2024), it can be used to assess DIF in CAT directly.

Modified Mantel–Haenszel (MH) Method

In the MH method, the examinees are matched on the basis of the observed raw scores (excluding the item being assessed, which can be called the studied item) of the test, which is taken as the matching variable. For studied item *i*, the examinees who are at the same level of the matching variable are divided into four cells according to (1) the group (i.e., focal/reference) in which he/she belongs, and (2) the correctness of the item response. Taking the *k*th ($k = 1, 2, \ldots, K$) level of the matching variable, for instance, the contingency table can be expressed as shown in Table 1.

Table 1

The Contingency Table for the kth Level of Matching Variables on Item i in Mantel-Haenszel Method

	Correctness on item *i*	
	Yes	**No**
Reference group	$n_{R1i,k}$	$n_{R0i,k}$
Focal group	$n_{F1i,k}$	$n_{F0i,k}$

Among the four numbers in Table 1, $n_{R1i,k}$ and $n_{R0i,k}$ indicate the number of examinees in the reference group at the *k*th level of the matching variable who answered item *i* correctly and incorrectly, respectively. Similarly, $n_{F1i,k}$ and $n_{F0i,k}$ represent the number of examinees in the focal group at the *k*th level of the matching variable who answered item *i* correctly and incorrectly, respectively. All four cell numbers form the total sample size N_k. Hence, the DIF measure in the MH method can be expressed as follows (Zwick et al., 1994):

$$\mathrm{MH}_{DIF} = -2.35 \cdot \log(\hat{\alpha}_{MH}) \tag{3}$$

where $\hat{\alpha}_{MH}$ is the MH conditional odds-ratio estimator and is given by

$$\hat{\alpha}_{MH} = \frac{\sum_k \frac{n_{R1i,k} \cdot n_{F0i,k}}{N_k}}{\sum_k \frac{n_{R0i,k} \cdot n_{F1i,k}}{N_k}}. \tag{4}$$

The corresponding standard error is given as $2.35 \cdot \sqrt{Var\left[\log(\hat{\alpha}_{MH})\right]}$ (Holland & Thayer, 1988), where

$$Var\left[\log(\hat{\alpha}_{MH})\right] = \frac{\sum_k \frac{U_k \cdot V_k}{N_k^2}}{2\left[\sum_k \frac{n_{R1i,k} \cdot n_{F0i,k}}{N_k}\right]^2}, \tag{5}$$

with $U_k = (n_{R1i,k} \cdot n_{F0i,k}) + \hat{\alpha}_{MH} \cdot (n_{R0i,k} \cdot n_{F1i,k})$ and $V_k = (n_{R1i,k} + n_{F0i,k}) + \hat{\alpha}_{MH} \cdot (n_{R0i,k} + n_{F1i,k})$. The statistic MH_{DIF} follows a chi-square distribution with one degree of freedom. Rejecting the null hypothesis indicates that item *i* is deemed as exhibiting DIF.

When implementing the MH method in CAT, raw scores cannot be used in a matching variable because different sets of items are administered to different examinees—that is, the examinees cannot be matched directly. Alternatively, examinees' latent trait estimates can be used as a matching variable when assessing DIF in CAT. Specifically, after examinees finish a CAT assessment, the latent trait estimates $\hat{\theta}$ can be obtained for all test-takers according to their responses (known as modified MH). Each of them is then stratified into one of *K* (e.g., 15 or 20) levels, which is used to match participants of reference and focal groups, and DIF analysis with the modified MH method can then be performed.[1] Thus, a set of 2 × 2 × *K* contingency tables will be defined and used for using the modified MH method. For the sake of simplicity, the terms *modified MH* and *MH* are used interchangeably hereafter.

Logistic Regression (LR) Method

Within the LR method, the observed raw scores of the examinees (*X*) and group membership (*G*, usually coded as 0 or 1, indicating the examinees' belonging to the reference or focal group), and their interaction are predictors (Shih et al., 2014). Thus, the probability of answering an item correctly in LR can be expressed as follows:

$$P(\mathrm{u} = 1|X, G) = \frac{\exp\left(\beta_0 + \beta_1 \cdot X + \beta_2 \cdot G + \beta_3 \cdot G \cdot X\right)}{1 + \exp\left(\beta_0 + \beta_1 \cdot X + \beta_2 \cdot G + \beta_3 \cdot G \cdot X\right)}, \tag{6}$$

where β_1, β_2, and β_3 are the regression coefficients for the influence of ability, group membership, and their interaction, respectively. Because more capable examinees have a higher chance of answering an item correctly, a positive and statistically significant β_1 is expected. When β_2 or β_3 significantly deviates from zero, the studied item is deemed as exhibiting uniform or nonuniform DIF, respectively (Jodoin & Gierl, 2001). Due to the current study's focus on uniform DIF, the β_3 term was not considered and tested here. Again, since the raw scores in CAT are no longer a valid indicator to represent the examinees' abilities, the person ability estimates are used here; therefore, the X term in Equation 6 is replaced with $\hat{\theta}$ (known as modified LR). Unlike the modified MH method, $\hat{\theta}$ does not need to be stratified in modified LR. For the sake of simplicity, the terms modified LR and LR are used interchangeably hereafter.

CATSIB Method

In CATSIB, examinees from different groups are matched using their latent trait estimates $E_G\left(\theta|\hat{\theta}\right)$, obtained from CAT and corrected for impact-induced statistical bias (Nandakumar & Roussos, 2004). Specifically, the CATSIB method employs regression correction procedures separately for both groups to obtain an unbiased DIF estimation because the groups are stochastically ordered in terms of the regression of the true score on the observed score when one group is stochastically larger than the other (Nandakumar & Roussos, 2004). After correcting for impact, the matching variable can be calculated as follows:

$$E_G\left(\theta|\hat{\theta}\right) = E_G\left(\theta\right) + \rho_G^2 \cdot \left[\hat{\theta} - E_G\left(\hat{\theta}\right)\right], \tag{7}$$

where ρ_G represents the correlation between θ and $\hat{\theta}$ in group *G*, which is defined as $\sqrt{1 - \frac{\sigma_{e,G}^2}{\sigma_{\hat{\theta},G}^2}}$, where $\sigma_{e,G}^2$ and $\sigma_{\hat{\theta},G}^2$ represent the variance of error scores and latent trait estimates for group *G* in CAT, respectively.

To assess DIF, the CATSIB uses different item response functions for various groups of examinees. Let the item response functions for the reference and focal groups be $P_R(\theta)$ and $P_F(\theta)$, respectively. For a specific latent trait level θ, the magnitude of DIF for a studied item can be expressed as

$$\mathrm{DIF}\left(\theta\right) = P_R\left(\theta\right) - P_F(\theta). \tag{8}$$

The DIF measure of the CATSIB, β, is defined as the average of $\mathrm{DIF}\left(\theta\right)$ over the latent trait scale and can be computed according to the following equation:

$$\beta = \int DIF(\theta) \cdot f(\theta) d\theta, \tag{9}$$

where $f(\theta)$ is an appropriate density function

[1] Fifteen ability levels were used in this study.

on θ that combines the reference and focal groups. Matching the examinees of two groups on the $E_G\left(\theta|\hat{\theta}\right)$ scale (denoted as $\hat{\theta}^*$), β can be approximated by (Nandakumar & Roussos, 2004)

$$\hat{\beta}=\sum_{\hat{\theta}^*=\hat{\theta}^*_{min}}^{\hat{\theta}^*_{max}}\left[\hat{P}_R\left(\hat{\theta}^*\right)-\hat{P}_F\left(\hat{\theta}^*\right)\right]\hat{p}(\hat{\theta}^*), \qquad (10)$$

where $\hat{P}_R\left(\hat{\theta}^*\right)$ and $\hat{P}_F\left(\hat{\theta}^*\right)$ represent the observed proportion-correct score on the studied item for examinees with the latent trait $\hat{\theta}^*$ of the reference and focal groups, respectively. In addition, $\hat{p}(\hat{\theta}^*)$ represents the observed proportion of examinees at $\hat{\theta}^*$. Because it is unlikely that examinees will be able to be matched at the exact same value of $\hat{\theta}^*$, the latent trait continuum should be divided into *K* equal intervals, and Equation 10 can be rewritten as

$$\hat{\beta}=\sum_{k=1}^{K}\left[\hat{P}_{R,k}-\hat{P}_{F,k}\right]\hat{p}_k, \qquad (11)$$

where $\hat{P}_{R,K}$ and $\hat{P}_{F,K}$ indicate the observed proportion-correct score on the studied item for the reference and focal groups in the *k*th interval, respectively. In addition, $\hat{P}_K$ is the observed proportion of examinees classified into the *k*th interval. Thus, the test statistic for CATSIB is then computed as

$$B=\frac{\hat{\beta}}{\hat{\sigma}(\hat{\beta})}, \qquad (12)$$

where $\hat{\sigma}(\beta)$ is the standard error for $\hat{\beta}$ and is estimated as

$$\hat{\sigma}\left(\hat{\beta}\right)=\sqrt{\sum_{k=1}^{K}\left[\frac{\hat{\sigma}^2_{R,k}(Y)}{n_{R,k}}+\frac{\hat{\sigma}^2_{F,k}(Y)}{n_{F,k}}\right]\hat{p}^2_k}, \qquad (13)$$

where *Y* denotes the response to the studied item. In addition, $\hat{\sigma}^2_{R,k}(Y)$ and $\hat{\sigma}^2_{F,k}(Y)$ represent the observed variances of *Y* in the *k*th interval for the reference and focal groups, respectively, whereas $n_{R,K}$ and $n_{F,K}$ are the number of examinees of group *G* in the *k*th interval for the reference and focal groups, respectively. The studied item is deemed as DIF when the absolute value of *B* exceeds 1.96 at the nominal level α = .05.

Scale Purification Procedure

The performance of DIF assessment methods can be influenced by the number of DIF items in the test. This is because the scale could be contaminated by the presence of DIF items. The more DIF items, the more serious the contamination. The effects of scale contamination on DIF assessment, such as inflated false positive rates and deflated true positive rates, can be diminished by applying scale purification procedures to the DIF assessment methods (Lee & Geisinger, 2016; Wang et al., 2009). However, in a CAT scenario, examinees receive different items, and items are administered to different examinees, which is unusual from the perspective of traditional DIF studies. Thus, this study also investigated whether a scale purification procedure could be appropriately applied to DIF assessment methods.

The main idea of a scale purification procedure is to implement DIF assessment iteratively. First, all items are assessed for DIF. Second, after temporarily removing suspected DIF items, all items are assessed again on the basis of using the remaining items as the matching variable. The second step continues until the results of two consecutive iterations show no change. Notably, applying the scale purification procedure may not be suitable for methods that require matching variables to detect DIF items in CAT. As previously explained, the latent trait estimates $\hat{\theta}$ after an exam can be used to replace raw total scores as the matching variable in the MH, LR, and CATSIB methods. However, re-estimating $\hat{\theta}$ after removing suspected DIF items is neither theoretically sound nor practically efficient, suggesting that refining the matching variable in the MH, LR, and CATSIB methods is not applicable. Since the OR method does not require a matching variable, incorporating the scale purification procedure into the OR method (hereafter denoted as OR-P) is viable in the CAT context.

Method

The performance of different methods (OR, OR-P, LR, MH, and CATSIB) in assessing DIF items was compared in a common CAT scenario through a series of simulation studies. Given that the proportions of missingness in the data matrix differ among variable-length CATs, this study focused solely on fixed-length CATs. In the simulations, items were administered to examinees in a fully adaptive manner without incorporating exposure control and content balance procedures into the item selection strategies. Therefore, a set of DIF-free items as anchors was not available in this study. All items in the item pool were assessed for DIF after the CAT administration, in which every test-taker received their latent trait estimate. Therefore, all methods except for OR and OR-P used $\hat{\theta}$ as their matching variable during the DIF assessment. The detailed design of the simulation study is described in the following section.

Design

Four independent variables were manipulated in the study: (a) the DIF assessment method, (b) magnitude of impact, (c) test length, and (d) percentage of DIF items in the pool. Five DIF assessment methods were investigated: OR, OR-P, MH, LR, and CATSIB. The magnitude of the impact was set at 0 or 0.5 logit. When there was no impact, the examinees' latent traits in both groups were generated from *N*(0, 1). When there was an impact, the latent traits were generated from *N*(0.5, 1) and *N*(0, 1) for the reference and focal groups, respectively. To obtain stable latent trait estimates, this study was guided by Linacre's (1994) suggestion that "…at least 8 correct responses and 8 incorrect responses are needed for reasonable confidence that an item calibration is within 1 logit of a stable value" (p. 328). Therefore, the test length was set at 20 and 30 items, which yielded non-blank responses of 6.67% and 10%, respectively, in a 300-item pool. Additionally, it was suggested that each item should be administered to at least 20 examinees. The percentage of DIF items in the pool was set at four levels: 0%, 10%, 20%, and 30%. When the DIF items were generated, 70% favored the reference group, and 30% favored the focal group.

Altogether, this simulation study included 5 × 2 × 2 × 4 = 80 conditions. Because CAT involves a high percentage of missing-by-design situations (e.g., 90% to 93.3%), a large sample of 5,000 examinees was simulated for every condition, in which half were assigned to the reference group and the other half were assigned to the focal group. Non-blank responses were generated using the dichotomous Rasch (1960) model. Item difficulties were sampled from *U*(−2, 2). For the convenience of inferencing, the DIF size was set to a constant 1 logit (Nandakumar & Roussos, 2004). A total of 200 replications were carried out for each condition.

Taking a 20-item test condition with 20% DIF items, for example, the data generation process for a specific replication can be described as follows:

1. The responses to all 300 items (60 DIF items and 240 non-DIF items) are simulated for all examinees using a standard item response generation process.
2. The random item selection method is utilized to select the initial item for each examinee, who is assumed to possess a latent trait equal to the distribution mean.
3. After the examinee's response to the item is made, the expected a posteriori (EAP) estimator is employed to estimate the interim latent trait for the examinee.
4. The amount of information on the interim latent trait is calculated for each of the remaining items in the pool. The item with the highest information is selected for

administration to the examinee, and the examinee's interim latent trait is updated with the EAP. This step is continued until the examinee reaches the pre-specified test length.

5. In cases where any of the 300 items have been administered to less than 20 examinees, the data are regenerated to ensure that every item has at least 20 valid responses (e.g., producing 6.67% non-blank responses in the 300-item pool).

Analysis

The DIF analyses of the five methods were implemented with R and corresponding packages. Two dependent variables were of interest. The *false positive rate* (FPR), also known as the Type I error rate, was calculated as the proportion of a DIF-free item that was mistakenly flagged as DIF across 200 replications. Theoretically, the FPR of an efficient method should not deviate far from a nominal level of 5%. In this study, the tolerance threshold of the FPR was set to 10%, which was consistent with previous studies (Chen et al., 2014; Shih et al., 2014). A method yielding FPRs greater than 10% was considered inflated, rendering the corresponding true positive rates (TPRs), also know as power, meaningless. The TPR was used to determine the proportion of correctly flagging DIF items. A method that produces higher TPRs than other methods is preferred, given that its FPRs are properly controlled.

Results

The means of the FPRs and TPRs over 200 replications for different methods in the manipulated conditions are summarized in Tables 2 and 3. Table 2 shows that for the 20-item test conditions, the FPRs for all the methods increased as the proportion of DIF items increased, which was consistent with previous research findings (Jin et al., 2018). The FPRs for the OR method were well controlled under all conditions, even when the DIF items comprised 30% of the items in the test. The OR-P method yielded slightly lower FPRs than the OR method when the DIF percentage was greater than 10%. Both the MH and LR methods produced well-controlled FPRs when the percentage of DIF items was less than or equal to 10%. When there were 20% or more DIF items in the test, the MH and LR methods lost control of their FPRs. The CATSIB method controlled the FPRs well only when there were no DIF items in the test and the impact was equal to zero. The effects of the magnitude of impact on the DIF assessment methods in CAT could be ignored, except for CATSIB, where greater impact resulted in a more serious inflation of the FPRs. The TPRs of all the methods generally decreased as the percentage of DIF items increased, with significant TPRs greater than 82.8%. The only exception would be the CATSIB method, in which the corresponding FPRs were greater than 10% and, in turn, caused the TPRs to be meaningless. The results for the 30-item test conditions are listed in Table 3, and the findings were similar to those of the 20-item test conditions. In general, as the proportion of DIF items increased, all the methods' FPRs increased, and the TPRs decreased; with the exception of the CATSIB method, all the methods produced meaningful TPRs of no less than 95.3%.

In sum, these findings of the DIF assessment methods' performance in the CAT context revealed that the OR and OR-P methods outperformed the other methods in terms of FPRs and TPRs. More specifically, even for the tests with DIF items at 30%, both OR and OR-P yielded well-controlled FPRs (< 10%) and satisfactory TPRs (> 85%). In addition, the OR-P method showed slightly lower FPRs and higher TPRs than the OR method in most conditions, confirming the benefits of adopting the scale purification procedure in DIF assessment. The MH and LR methods kept the FPRs under control when DIF items were no more than 10% in the test, whereas CATSIB failed to control the FPRs when DIF items were 10% or more in the test.

Table 2

Averaged False Positive Rates and True Positive Rates (%) in a 20-Item CAT

DIF items	False positive rates (FPR)					False positive rates (FPR)				
	OR	OR-P	LR	MH	CS	OR	OR-P	LR	MH	CS
Impact = 0										
0%	5.4	5.4	4.8	4.9	6.1	–	–	–	–	–
10%	5.5	5.5	6.2	6.1	**12.4**	91.0	91.2	90.0	88.3	**86.1**
20%	6.8	6.2	**11.0**	**10.6**	**21.1**	87.9	88.0	**86.9**	**85.0**	**83.2**
30%	8.0	6.9	**12.4**	**12.2**	**22.1**	88.2	88.6	**86.8**	**84.9**	**83.3**
Impact = 0.5										
0%	5.3	5.4	4.8	4.7	**34.2**	–	–	–	–	–
10%	5.9	5.9	6.3	5.9	**42.3**	89.0	89.2	89.0	87.1	**74.8**
20%	7.1	6.8	**10.9**	9.6	**52.8**	85.6	85.7	**85.0**	82.8	**67.5**
30%	8.0	7.1	**11.5**	**10.5**	**52.9**	86.7	87.2	**85.4**	**83.7**	**65.4**

Note. OR = odds ratio. OR-P = odds ratio with iterative purification procedures. LR = logistic regression. MH = Mantel-Haenszel. CS = CATSIB. Inflated FPRs (> 10%) and corresponding TPRs were highlighted in bold.

Table 3

Averaged False Positive Rates and True Positive Rates (%) in a 30-Item CAT

DIF items	False positive rates					True positive rates				
	OR	OR-P	LR	MH	CS	OR	OR-P	LR	MH	CS
Impact = 0										
0%	5.6	5.6	5.1	5.1	5.7	–	–	–	–	–
10%	5.6	5.6	7.3	7.4	**12.6**	98.4	98.3	97.9	97.6	**95.4**
20%	7.5	7.1	**13.9**	**13.5**	**21.6**	97.6	97.5	**97.5**	**97.0**	**95.7**
30%	9.3	7.9	**16.0**	**15.7**	**22.7**	96.9	97.0	**96.1**	**95.5**	**94.5**
Impact = 0.5										
0%	5.8	5.8	4.9	5.1	**24.0**	–	–	–	–	–
10%	6.4	6.4	7.2	6.8	**36.9**	97.2	97.2	97.0	96.7	**90.6**
20%	7.7	7.3	**13.3**	**12.0**	**49.2**	96.5	96.5	**96.6**	**96.0**	**89.2**
30%	8.8	8.0	**14.4**	**13.1**	**49.4**	95.3	95.6	**94.8**	**94.5**	**86.0**

Note. OR = odds ratio. OR-P = odds ratio with iterative purification procedures. LR = logistic regression. MH = Mantel-Haenszel. CS = CATSIB. Inflated FPRs (> 10%) and corresponding TPRs were highlighted in bold.

Surprisingly, the CATSIB method yielded inflated FPRs even when only 10% of the items in the test exhibited DIF, which contradicts previous research findings (Nandakumar & Roussos, 2004). Even under the condition where there were no DIF items with an impact of 0.5, CATSIB failed to produce acceptable FPRs (34.2% and 24.0% for the 20- and 30-item conditions, respectively). Such poor performance might be due to the use of an improper binning approach to obtain a matching variable in CATSIB, and this should be explored further in the future.

Discussion

DIF assessments are indispensable not only for paper-based exams but also for computer-based exams. This study sheds light on the need for the regular and comprehensive review of item quality in a CAT. This is not intended to overturn the original results of an exam. (In practice, even if it is later discovered that some test items in the pool are found to be unfair to a specific group of test-takers, it is not feasible to withdraw personal evaluations released months or even years previously.) Instead, the study emphasizes that early retirement of poor items (those that have not been overexposed) is crucial to ensuring the quality of future tests.

We examined and compared various DIF assessment methods, specifically the OR, MH, LR, and CATSIB, in a CAT environment. Through a series of simulation studies, the results showed that the OR method outperformed the other three methods in controlling FPRs and producing high TPRs. Combining the OR with a scale purification procedure further improved the DIF assessment in terms of the FPR and TPR. These findings demonstrated that the OR method is a useful and robust approach for DIF assessment in operational CAT items. In addition to its computational simplicity, the OR method demonstrated well-controlled Type I error rates and slightly higher power than the MH, LR, and CATSIB methods under the simulated conditions, particularly when used in conjunction with a scale purification procedure.

Relative to the other DIF assessment methods, the CATSIB performed the worst across the conditions used. In particular, the CATSIB yielded exaggerated FPRs when DIF items were only 10% of the test and/or an impact existed among the different groups. This finding was contrary to those in previous studies (Nandakumar & Roussos, 2004). This might be due to differences in the simulated conditions between the studies. One of the important differences is that, in the literature, a set of operational items was regarded as *good* and *qualified* matching variables when assessing DIF for new items, and the obtained latent trait estimates based on these good items were then used as the matching variables to conduct DIF assessments on new items. We discarded this “too perfect” assumption but focused on a more general situation in which DIF could exist in an established pool and with newly added items. The adopted matching variables might be biased, which, in turn, would influence the performance of the DIF assessment. That is, the latent trait estimates could be biased due to the inclusion of DIF item(s) in the calibration, weakening the performance of the CATSIB method.

We also noticed that the performance of the CATSIB was not as robust as that of MH despite the use of the same information as matching variables. This might be because, in the CATSIB, latent trait estimates were stratified into fixed intervals (Equation 11) throughout the DIF assessment. In contrast, the stratification of the matching variable in MH was performed on a rolling basis (i.e., by the studied item) using valid cases that took that studied item. Future research could modify the stratification rule of the CATSIB while re-examining its performance.

The response data in this study were generated using only the dichotomous Rasch (1960) model, which was a common practice in previous DIF studies (Chen et al., 2014; Jin et al., 2018). Because the performance of DIF assessment methods under different response models has been investigated in the literature, this study focused on other independent variables rather than response models. Furthermore, the performance of DIF assessment methods under other DIF patterns (e.g., balanced and constant conditions) has also been explored in the literature, and their findings are quite consistent (Wang & Su, 2004; Wang et al., 2012). Therefore, only the most referenced DIF condition was manipulated in this study.

CAT is a consecutive testing method in which examinees participate with a flexible schedule; thus, regularly assessing DIF in both operational and new items is necessary. As this study demonstrated, the MH and LR methods tended to lose control of FPRs when the DIF items represented 20% or more of the items in the test. This could largely compromise the applicability of these methods. How to develop new strategies, such as scale purification and revising the latent trait scale, for these methods in CAT is worth further investigation. In addition, the CATSIB adopted a regression correction procedure (Nandakumar & Roussos, 2001), which should be able to impede the negative effects impacting DIF assessment. However, this was not found in the current study. Recommended conditions for the CATSIB method that can be appropriately applied to DIF assessment are of great interest. Future studies should look for ways to address the CATSIB’s limitations in CAT environments. Moreover, more investigations are suggested to gain a better understanding of the performance of the OR and OR-P methods in various CAT contexts, such as in assessing DIF with more than two groups and under polytomous items.

Moreover, previous studies on CAT DIF assessment have primarily focused on either operational items or seeded items. Developing a DIF assessment method that accommodates both would be highly valuable. However, in a continuous CAT system incorporating both seeded and operational items, seeded items may receive a limited number of responses in the early stages due to the administration schedule. This constraint can diminish the power of DIF assessment, as an insufficient calibration sample size may hinder the effectiveness of the IRT-based DIF approaches. One potential approach is to extend the OR method to assess DIF for both item types, as it does not rely on calibration derived from a sufficient number of examinee responses. Thus, this approach aligns seamlessly with the consecutive testing framework of CAT, thereby enhancing its overall utility and effectiveness. In addition, evaluating the performance of the OR method in comparison to the recent DIF assessment for CAT, RDIF (Lim & Choe, 2023), warrants further investigation. Both methods offer computational efficiency and robust DIF assessment capabilities, which can significantly enhance the practical implementation of CAT in real-world scenarios.

References

Chen, C.-C., Tang, C.-W., & Jin, K.-Y. (2024). Influences of internet access on civic knowledge measurement in Taiwan. *Large-Scale Assessments in Education*, *12*, 20. https://doi.org/10.1186/s40536-024-00209-8

Chen, J.-H., Chen, C.-T., & Shih, C.-L. (2014). Improving the control of Type I error rate in assessing differential item functioning for hierarchical generalized linear model when impact is presented. *Applied Psychological Measurement*, *38*(1), 18–36. https://doi.org/10.1177/0146621613488643

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). Lawrence Erlbaum Associates.

Hsu, C.-L., & Wang, W.-C. (2015). Variable-length computerized adaptive testing using the higher order DINA model. *Journal of Educational Measurement*, *52*(2), 125–143. http://www.jstor.org/stable/43940561

Hsu, C.-L., Wang, W.-C., & Chen, S.-Y. (2013). Variable-length computerized adaptive testing based on cognitive diagnosis models. *Applied Psychological Measurement*, *37*(7), 563–582. https://doi.org/10.1177/0146621613488642

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2018). Using odds ratios to detect differential item functioning. *Applied Psychological Measurement*, *42*(8), 613–629. https://doi.org/10.1177/0146621618762738

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, *43*(4), 355–381. https://doi.org/10.1111/j.1745-

3984.2006.00021.x

Lee, H., & Geisinger, K. F. (2016). The matching criterion purification for differential item functioning analyses in a large-scale assessment. *Educational and Psychological Measurement*, *76*(1), 141–163. https://doi.org/10.1177/0013164415585166

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*(3), 245–264. https://doi.org/10.1111/j.1745-3984.2006.00015.x

Lim, H., & Choe, E. M. (2023). Detecting differential item functioning in CAT using IRT residual DIF approach. *Journal of Educational Measurement*, *60*(4), 626–650. https://doi.org/10.1111/jedm.12366

Lim, H., Choe, E. M., & Han, K. (2022). A residual-based differential item functioning detection framework in item response theory. *Journal of Educational Measurement*, *59*(1), 80–104. https://doi.org/10.1111/jedm.12313

Linacre, J. M. (1994). Sample size and item calibration [or person measure] stability. *Rasch Measurement Transactions*, *7*(4), 328. https://www.rasch.org/rmt/rmt74m.htm

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

Nandakumar, R., & Roussos, L. (2001). *CATSIB: A modified SIBTEST procedure to detect differential item functioning in computerized adaptive tests*. *Law School Admission Council computerized testing report*. *LSAC Research Report Series*. 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*(2), 177–199. https://doi.org/10.3102/10769986029002177

Pine, S. M. (1977). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing: Proceedings of a symposium presented at the 18th annual convention of the Military Testing Association* (Research Rep. No. 77-1, pp. 37–43). University of Minnesota.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Institute of Education Research.

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. https://doi.org/10.1007/BF02294572

Shih, C.-L., Liu, T.-H., & Wang, W.-C. (2014). Controlling type I error rates in assessing DIF for logistic regression method combined with SIBTEST regression correction procedure and DIF-free-then-DIF strategy. *Educational and Psychological Measurement*, *74*(6), 1018–1048. https://doi.org/10.1177/0013164413520545

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147–172). Lawrence Erlbaum Associate.

van der Linden, W. J., & Glas, C.A.W. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Kluwer.

Wainer, H., Dorans, N. J., Flaugher, R., Mislevy, R. J., Thissen, D., Eignor, D., Green, B. F., & Steinberg, L. (2000). *Computerized adaptive testing: A primer (2nd ed.)*. Lawrence Erlbaum Associates.

Wang, C., Weiss, D. J., & Shang, Z. (2019). Variable-length stopping rules for multidimensional computerized adaptive testing. *Psychometrika*, *84*(3), 749–771 (2019). https://doi.org/10.1007/s11336-018-9644-7

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, *72*(4), 687–708. https://doi.org/10.1177/0013164411426157

Wang, W.-C., Shih, C.-L., & Yang, C.-C. (2009). The MIMIC method with scale purification for detecting differential item functioning. *Educational and Psychological Measurement*, *69*(5), 713–731. https://doi.org/10.1177/0013164409332228

Wang, W.-C., & Su, Y.-H. (2004). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*(2), 113–144. https://doi.org/10.1207/s15324818ame1702_2

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*(4), 473–492. https://doi.org/10.1177/014662168200600408

Zwick, R. (2000). The assessment of differential item functioning in comput adaptive tests. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 221–244). Springer.

Zwick, R., & Thayer, D. T. (2002). Application of an empirical Bayes enhancement of Mantel-Haenszel differential item functioning analysis to computerized adaptive test. *Applied Psychological Measurement*, *26*(1), 57–76. https://doi.org/10.1177/0146621602026001004

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*(2), 121–140. https://doi.org/10.1177/014662169401800203

Zwick, R., Thayer, D. T., & Wingersky, M. (1994b). *DIF analysis for pretest items in computer adaptive testing*. (Research Report No. 94-33). Educational Testing Service.