JOURNAL OF APPLIED MEASUREMENT, *25*(1/2), 48-74

Enhancing DIF Detection in Cognitive Diagnostic Models: An Improved LCDM Approach Using TIMSS Data

Su-Pin Hung
National Cheng Kung University

Po-Hsi Chen
National Taiwan Normal University

Hung-Yu Huang
National Cheng Kung University

Cognitive diagnostic models (CDMs) are being increasingly utilized in educational research, especially for analyzing large-scale international assessment datasets comparing skill mastery among various countries and gender groups. However, establishing test invariance before making such comparisons is critical to ensure that differential item functioning (DIF) does not distort the results. Earlier research on DIF detection within CDMs has predominantly dealt with non-compensatory models, leaving several influential factors ambiguous. This study addresses these issues with the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), enhancing its practical applicability. The aim is to improve the LCDM-DIF method and evaluate the efficacy of the purification procedure using total scores as matching criteria with non-parametric methods—logistic regression (LR) and Mantel-Haenszel (MH). Factors examined include test length, percentage of DIF, DIF magnitude, group distributions, and sample size. Using data from the Trends in International Mathematics and Science Study (TIMSS), factoring nationality and gender of the participants, an empirical study gauges the performance of the proposed methods. The results reaffirm that the model-based method surpasses the MH and LR methods in controlling Type I errors and achieving higher power rates. Additionally, the LCDM-based approach offers broader insights into the results. The study discusses its value, potential applications, and future research areas, emphasizing the significance of tackling issues related to contaminated matching criteria in DIF detection within CDMs.

Keywords: cognitive diagnostic models, differential item functioning, log-linear cognitive diagnosis model, purification procedure

Requests for reprints should be sent to Su-Pin Hung, Center of Teacher Education & Institute of Education, National Cheng Kung University, Taiwan; email: sphung@mail.ncku.edu.tw

Acknowledgments: The idea for this work was inspired by Prof. Wang. I would like to express my sincere gratitude for his valuable contributions to the field of Differential Item Functioning (DIF) issues. Additionally, I would like to acknowledge the support received from the Grants provided by the Ministry of Science and Technology, Taiwan, which funded the preparation of this study (No. 103-2410-H-006 -124 -)

Introduction

Cognitive diagnostic models (CDMs) provide a detailed understanding of examinees' mastery levels of a specific set of attributes, not just an estimation of overall ability. These attributes—representing an examinee's distinct skills or knowledge components—are fundamental for understanding their performance on a series of items or tasks, each relevant to the assessment domain. The varying attribute mastery patterns among examinees, namely mastering all, some, or none of the attributes, influence their item responses. Accordingly, CDMs portray the relationship between students' skills mastery and their observed item score. These diagnostic results assist teachers in adjusting curricular materials, while the individual mastery profiles inform students of areas that need to improve.

CDMs allow for compensatory or non-compensatory relationships among attributes and are classified accordingly (see Rupp et al., 2010, for an overview). Non-compensatory models, such as the deterministic inputs, noisy "and" gate (DINA) model (Haertel, 1989; Junker & Sijtsma, 2001), implement a conjunctive rule that requires a respondent to master all required attributes to obtain a score on a tested item. In contrast, compensatory models, like the deterministic inputs, noisy "or" gate (DINO) model (Templin & Henson, 2006), apply a disjunctive rule. This rule only requires a subset of the necessary attributes for an item and allows mastery of other attributes to compensate. More recent models, such as the general diagnostic model (von Davier, 2005), the log-linear cognitive diagnosis model (LCDM; Henson et al., 2009), and the generalized DINA model (de la Torre, 2011), incorporate both types of relationships. This dual relationship makes them more practical in real-world scenarios.

The development of CDMs and their application to large-scale educational surveys for comparing skill mastery levels across countries or gender groups (e.g., Lee et al., 2011; Park et al., 2018) have led to the emergence of several related issues. One such issue is the differential item functioning (DIF) within cognitive diagnostic contexts. DIF implies that individuals from different groups (e.g., gender, ethnicity, age) with identical underlying abilities or trait levels demonstrate varying success rates for specific test items (Holland & Wainer, 1993). In the context of CDMs, DIF occurs when examinees possessing the same latent attribute profile from different groups have different probabilities of correctly answering an item (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Zhang, 2006).

A few studies have investigated the impact of DIF in CDMs (e.g., Eren et al., 2023; Gierl et al., 2008; Hou et al., 2014; Li, 2008; Li & Wang, 2015; Svetina et al., 2018; Qiu et al., 2019; Zhang, 2006). The DIF methods developed under CDMs primarily stem from classical test theory (CTT), such as the application of the MH method (Holland & Thayer, 1988; Mantel & Haenszel, 1959), the Simultaneous Item Bias Test (SIBTEST; Shealy & Stout, 1993), and the logistic regression method (LR; Swaminathan & Rogers, 1990). There are also CDM-based DIF assessment methods, like the modified item response theory (IRT)-based DIF approach under CDMs. For instance, Hou et al. (2014) proposed a two-stage strategy, which involves estimating item parameters for reference and focal groups with separate calibration in the first stage and then applying the Wald test to detect DIF in the second stage. Other studies, like that of Li (2008), adopted single joint analysis. In this case, Li extended the higher-order deterministic inputs, noisy, and gate (HO-DINA) model, incorporating a group variable on the attribute level to detect DIF and group differences at the attribute level, known as DAF (differential attribute functioning). Similarly, Li and Wang (2015) proposed the LCDM-based DIF method, which involves regressing item parameters on grouping variables.

Traditionally, a matching variable is used to detect DIF items, aligning the reference and focal groups on the same metric to assess item performance for DIF. Accordingly, building a common metric before DIF detection is necessary, regardless of the DIF detection method applied. Previous studies focused on comparing differing DIF detection methods under CDMs and exploring various matching variables when implementing non-parametric DIF methods within CDMs. The scale purification problem was considered to identify a non-DIF item set from the evaluated instrument for use as a matching criterion in DIF detection using the MH method in CDMs (e.g., Qiu et al., 2019). However, all studies that were simulated primarily targeted non-compensatory models. The estimation requirement for the compensatory model may not align with the DINA model. Furthermore, the effects of differing test lengths, DIF percentages, magnitudes, and sample sizes in a compensatory dataset warrant further investigation. In another instance, studies used the DINA model in their simulation study to compare the MH, LR, and Wald tests (Svetina et al., 2018). Yet, the efficacy of applying LR with scale purification under CDMs remains to be seen. Lastly, as Paulsen et al. (2019) pointed out, previous studies solely focused on developing DIF estimation methods in CDMs, failing to document how different DIF-related scenarios impact classification accuracy.

The current study aims to broaden the general CDM-based DIF method and incorporate a compensatory model to assess the need for stable item parameter recovery under various circumstances. Specifically, DIF is assessed by incorporating group variables into the LCDM model. To address the research gap in previous studies, the LR method and MH method, coupled with scale purification procedures, were adopted as a competitor to the LCDM-DIF method. This study strives to strengthen the LCDM-DIF method and evaluate the efficacy of the purification procedure when using total scores as matching criteria with non-parametric techniques. The study seeks to pinpoint the most effective conditions for matching DIF detection strategies through simulation studies. The gained insights will enable practitioners to make informed decisions during data analysis.

Background

The LCDM Model

The LCDM is one of the most general CDMs and can accommodate most common CDMs. It functions similarly to common CDMs, including the DINA and DINO models, the compensatory reparameterized unified models (C-RUM), and the Reduced RUM through the use of model constraints. Like every CDM, the LCDM consists of three essential components: the Q-matrix, the alpha-matrix, and the latent response. The Q-matrix specifies which attributes each item assesses, typically including *j* items in the rows and *k* attributes in the columns. Here, entries of 1 or 0 indicate whether an item measures a given attribute (e.g., $q_{jk} = 1$). As it operationalizes a substantive theory informing diagnostic assessments' construction, the Q-matrix plays a central role in CDMs (Rupp et al., 2010). It functions much like a test blueprint and is usually developed by domain experts. In contrast, the alpha-matrix defines examinees' mastery profiles across the specified attributes, traditionally arranging examinees in rows and attributes in columns. The latent response is an indicator depending on both the Q-matrix and the alpha-matrix. In the current study, we aim to aid practitioners in assessing DIF in practical CDM applications. We chose the LCDM (Henson et al., 2009) as its general frame for empirical applications.

Under the LCDM, the log-odds of the probability of success (a score of 1) on item *j* for examinee *i* with attribute profile $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik}), = \alpha_{ik}$ 1 indicates mastery of attribute *k* by examinee *i* and 0 otherwise, is defined as follows:

$$P\left(Y_{ij} = 1 \middle| \alpha_i\right) = \frac{\exp[\lambda_{j,0} + \lambda_j^T h(\alpha_i, q_j)]}{1 + \exp[\lambda_{j,0} + \lambda_j^T h(\alpha_i, q_j)]}, \quad (1)$$

where $\lambda_{j,0}$ is related to the probability of success for non-masters of all measured attributes required by item *j*; the vector λ_j^T represents a $1 \times \left(2^k - 1\right)$ vector of weights with the main effect and intercept parameters for the *j*-th item; q_j is a vector of q_{jk} , $q_{jk} = 1$ indicates a given attribute *k* is measured by an item *j*, and 0 otherwise; the $h(\alpha_i, q_j)$ represents a set of linear combinations of the α_i and q_j ; and the $\lambda_j^T h(\alpha_i, q_j)$ can be written as follows:

$$\lambda_j^T h(\alpha_i, q_j) = \sum_{k=1}^{K} \lambda_{jk}\left(\alpha_{ik} q_{jk}\right) + \sum_{k=1}^{K} \sum_{\nu > k} \lambda_{jkv}\left(\alpha_{ik}\alpha_{iv} q_{jk} q_{jv}\right) + \ldots \quad (2)$$

Therefore, the LCDM defines the log-odds as a linear function that includes all attribute main effects and all possible interactions. Such a model can be expanded to include all possible conditional relations. By constraining a portion of the parameters of the LCDM, the full model can be reduced to a set of commonly used compensatory and non-compensatory CDMs.

Specifically, the DINA model can be represented by constraining all main effects to "0" and keeping only the intercept and the highest-order interaction terms in the LCDM. Besides producing a conjunctive model in the LCDM, it also generates a model with a disjunctive condensation rule aligning with the DINO model. In the DINO model, any attribute can entirely compensate for the absence of all others. The values of the main effect and interaction effect parameters are constrained to be equal in the DINO model (i.e., if two attributes are required for the item *j*, then $\lambda_{j1} = \lambda_{j2} = \lambda_{j12} = \lambda_j$). Furthermore, the interaction is defined as the negative of any of the main effects. Other subsumed models can be specified under the framework of the LCDM, and readers who are interested in the variety of manifestations of the LCDM can refer to the original study by Henson et al. (2009).

DIF Detection Approaches Within the Context of CDMs

DIF detection has long been recognized as a standard part of test development to ensure fairness at the item or test level (American Educational Research Association et al., 1999). When developing new tests, items exhibiting DIF are typically revised or discarded (Tennant & Pallant, 2007). Hence, for valid inferences and to ensure measurement invariance, it is also necessary to detect DIF items in the context of cognitive diagnostic measurements. Consequently, detecting DIF items in cognitive diagnostic measurements is necessary to make valid inferences and maintain measurement invariance. All DIF detection methods can be broadly classified into IRT-based and non-IRT-based approaches (see Magis et al., 2010, 2011 for an overview). The IRT-based approach presumes a specific IRT model applies to each group and examines DIF using model fitting. Tests such as the IRT-based chi-square test (Lord, 1980) or the IRT-based likelihood ratio test method (Thissen et al., 1993) spot differences in item parameters between groups. Conversely, non-IRT methods (i.e., MH, SIBTEST, and LR methods) hinge on statistical techniques assessing the presence of DIF with the total score on ability, or some other variable as the matching criterion, not requiring an IRT solution, which can decrease computation times.

Two types of DIF can be identified: uniform DIF and non-uniform DIF. Uniform DIF indicates a consistently higher likelihood of one group answering an item correctly across the entire trait continuum, while non-uniform DIF suggests that the direction of DIF changes over the trait continuum; that is, an item may favor one group at low trait levels and the other group at high trait levels. The strength of the IRT-based approach is its ability to detect both uniform and non-uniform DIF. In contrast, most non-IRT methods were primarily designed for uniform DIF detection only, such as the MH method and the SIBTEST. The Logistic Regression (LR) method can serve as a bridge between IRT-based and non-IRT-based methods, being able to detect both uniform and non-uniform DIF (Magis et al., 2010). Given the pros and cons of both IRT-based and non-IRT-based methods, previous research has often investigated the differences in their efficacy in identifying DIF items under various conditions.

To ensure test invariance within the CDM framework, numerous studies have adapted and modified the aforementioned non-IRT methods (non-parametric methods) to investigate DIF. These studies primarily focus on comparing various methods based on Type I error control and power rates (Hou et al., 2014; Li, 2008; Li & Wang, 2015; Paulsen et al., 2019; Qiu et al., 2019; Svetina et al., 2018; Zhang, 2006). In this context, the power rate is determined by calculating the percentage of correctly identified known DIF items. Conversely, the Type I error rate refers to the percentage of non-DIF items incorrectly identified as such. Zhang (2006) compared two traditional non-parametric DIF methods (MH and SIBTEST) at the item level, taking into account two different matching variables (total test scores and attribute mastery profile). In Zhang's study, the attribute mastery profile was first extracted based on the DINA model and utilized as a matching variable. Results revealed that matching mastery profile scores led to superior control of Type I error and higher power rates compared to using total test scores. While the MH method showed a slight edge over SIBTEST in identifying uniform DIF, both methods fell short in pinpointing non-uniform DIF.

In contrast to the earlier non-parametric method, several studies have proposed the CDM-based DIF method. Li (2008) modified the higher-order (HO)-DINA model by adding a group indicator as a covariate at the attribute level and by allowing the item parameters to differ across various groups to detect DIF and DAF simultaneously. She discovered that the CDM-based DIF approach provided better control of Type I error rates and power rates than the MH method. Unsurprisingly, the model-based approach outperformed the non-parametric method.

Hou et al. (2014) utilized the DINA model to calibrate item parameters for the reference and focal group separately and detected DIF items through the Wald statistic. In their simulation design, only the test length and number of attributes were held constant, while the impact of DIF percentage and DIF size was investigated. They found that the Wald test resulted in inflated Type I errors and low power rates, specifically with poorly discriminated items and small DIF sizes. The Wald statistic used in the CDM-based approach can detect uniform and non-uniform DIF, with results indicating that its performance was comparable to or better than the MH and SIBTEST procedures in general.

Li and Wang (2015) introduced the LCDM-based DIF method and applied the Bayesian Markov chain Monte Carlo (MCMC) methods (de la Torre & Douglas, 2004) for parameter estimation. The Bayesian MCMC provided a more accurate estimation of parameters than the Wald test used by Hou et al. (2014). This was primarily due to its inaccurate estimation of standard errors (Qiu et al., 2019). The LCDM-based DIF method outperformed the Wald method when item parameters (i.e., poorly discriminated items) were high. This new method straightforwardly applies to the DIF detection of multi-group variables. Furthermore, the grouping variables in DIF detection can be quantitative or latent.

Though previous simulation studies suggested that the CDM-based DIF method performed better than traditional DIF methods (i.e., MH and SIBTEST), especially in detecting non-uniform DIF, it is important to note that CDM-based DIF methods are more computationally demanding. This is why previous DIF studies continue to investigate possible traditional DIF methods within the CDM context. Svetina et al. (2018) delved into this concept by comparing LR to MH and the Wald method under a Q-matrix misspecification, which could lead to differential Q-matrix functioning (i.e., reference and focal groups using different attributes to solve items). Their results indicated that LR yielded equally high or higher power rates when compared to the Wald method and MH method. Additionally, Svetina et al. found that while both MH and LR maintained reasonable Type I error rates, the Wald method, however, yielded inflated Type I error rates across most conditions.

Qiu et al. (2019) addressed the issue of how to implement the MH method with various purification procedures within CDM. They identified three types of DIF that could occur in CDMs. GS-DIF occurs when different groups have unique guessing and slipping item parameters, and the q-matrix of an item is consistent across groups. MM-DIF happens when different groups follow different measurement models for the studied item, while Q-DIF indicates that different groups have unique Q-matrices. While dealing with the GS-DIF, previous studies reported that the equal model and Q-matrix (EMQ) method performed effectively in DIF detection without the need for matching variables and scale purification (Hou et al., 2014; Li & Wang, 2015). In Qiu et al.'s study (2019), the EMQ method was compared to four purification strategies of the MH method. The findings showed that the purification procedure with the MH method, using profile scores as matching variables, yielded superiorly controlled Type I error rates and higher power rates than the EMQ method in detecting Q-DIF and MM-DIF. While the EMQ method exhibited better performance in detecting GS-DIF and maintained promising power rates across all conditions, it yielded inflated Type I error rates when the DIF source was from varying Q-matrices or models, especially under the highest DIF percentage (e.g., 40%). Qiu et al. concluded that the effectiveness of the DIF-free-then-DIF, DFTD strategy (Wang, 2008), the PMH-D (i.e., using MH with profile score as matching and utilizing DFTD strategy) outperformed the PMH-P strategy (i.e., using MH with profile score as matching and utilizing scale purification procedures, comparable to all the other item purification, AOI-P method in IRT-DIF), particularly in cases of high DIF percentages or varying Q-matrices/models. They also determined that the PMH-P method excelled over the PMH-S method (i.e., using MH with profile score as matching), indicating that scale purification methods were beneficial.

Rationale for the Current Study

In previous studies, all items were primarily simulated from the DINA model. However, there is less information about DIF items occurring in other CDMs. While the generalized framework of Li and Wang (2015) is promising, further investigation is necessary to understand the performance of the LCDM-DIF method under various conditions and different sub-models. Moreover, while previous DIF detection studies within the framework of CDMs (e.g., the DINA model; the HO-DINA model) considered factors like sample size, structures of the Q-matrix, the distribution of ability, discrimination of attributes, and types of matching criteria that could influence type I error control and power with different DIF detection methods, the effect of various conditions (e.g., unequal sample size, test length, extreme amount of DIF items, different levels of DIF magnitude) on detecting DIF under compensatory CDMs remains unknown.

For instance, Paulsen et al. (2020) used the DINA model to investigate how DIF-related scenarios affect classification accuracy. They observed that the magnitude of DIF (e.g., DIF of 0.10) and the number of DIF items (e.g., 30%) must be large to significantly diminish classification accuracy at the individual attribute level. Additionally, unequal distributions between groups could reduce classification accuracy. It would be interesting to further investigate how DIF items affect classification accuracy with a different model.

Moreover, most prior studies have frequently assumed that the matching criterion is DIF-free, often neglecting the purification procedure of the matching criterion (that is, the test total score or the mastery profile score). Qiu et al. (2019) employed profile scores as matching variables, delving into the effectiveness of MH with various purification procedures in contrast to the CDM-based DIF method. Considering previous studies have highlighted the non-parametric DIF method in DIF detection, the effectiveness of other non-parametric DIF methods against model-based DIF methods within the CDM context remains an area worthy of further investigation. In multiple earlier studies within the IRT framework (Finch & French, 2007; French & Maller, 2007; Li & Stout, 1996; Shih & Wang, 2009), LR has been demonstrated to be reasonably practical, boasting higher power rates and lower Type I error rates than SIBTEST. Noted for its simplicity and ability to detect non-uniform DIF, LR remains another practical choice, although its application with purification in CDMs is yet to be explored.

As noted in previous studies (Zhang, 2006), using the profile as a matching variable is more computationally intensive than using the total score. This study aims to investigate the use of the total score for matching and to implement the purification procedure with the LR and MH methods, comparing them to the CDM-based DIF method. Paulsen et al. (2020) also noted that understanding the impact of DIF on classification accuracy is vital in CDM-DIF studies. Therefore, this study aims to enhance the robustness of the LCDM-DIF method and assess its effectiveness in attribute classification accuracy, a facet often overlooked in previous CDM-DIF research.

By addressing these gaps, this study aims to fortify the LCDM-DIF method and assess the efficiency of the purification procedure when total scores are used as matching criteria with non-parametric methods. Through well-crafted simulation studies, the research aims to pinpoint the optimal conditions for method-to-match DIF detection strategies. The insights garnered will equip practitioners to make well-informed decisions during the data analysis process.

DIF Detection Methods in the Present Study

Compared to previous studies, the purpose of DIF detection assumed here is the GS-DIF. The LCDM-based DIF and the two non-IRT-based approaches are implemented. Following Qiu et al. (2019), the LCDM-based DIF method used in this context is henceforth referred to as the EMQ method. Brief introductions to the three DIF methods implemented in this study are provided.

EMQ Method (LCDM-Based DIF)

Li and Wang (2015) expanded the LCDM to account for the DIF effect in CDMs. They accomplished this by allowing the reference and focal groups to have different sets of item parameters and introduced the subscript ‘c’ to the LCDM. The model can, therefore, be formulated as follows:

$$P(Y_{ij} = 1 \,|\alpha_{ic}) = \frac{\exp[\lambda_{j,0,c} + \lambda_{j,c}^{T} h(\alpha_{ic}, q_j)]}{1 + \exp[\lambda_{j,0,c} + \lambda_{j,c}^{T} h(\alpha_{ic}, q_j)]}, \tag{3}$$

the subscript c stands for group membership (c = 1, …, C). Thus, while an item is DIF-free, any person with the same attribute profile α, who belongs to any group, would have the same probability of answering this item correctly. The grouping variable could be categorical or continuous. In the present study, only two groups were manipulated, where c = 1 denotes the reference group and c = 2 is the focal group. The $h(\alpha_{ic}, q_j)$ represents a set of a linear combination of the α_{ic} and the Q-matrix entries for the j_{th} item, q_j. The $\lambda_{j,0,c}$ is the intercept term that is related to the probability of correct responses for the non-masters and belong to group c. The $\lambda_{j,c}^{T}$ is a vector of weights for item j with a length of $2^k - 1$ for group c; K is the number of latent binary attributes; q_{jk} (k = 1, …, K) is the entry for item j in the Q-matrix.

As previously noted, the LCDM can be simplified to less intricate models. The current work reduces the LCDM to accommodate both the DINA and DINO models. Each model requires unique constraints. In the DINA model, the primary effects are set to zero, with only the intercept term and the highest-order interaction being evaluated; these must be positive. On the other hand, to accommodate the DINO model, the main effects are assessed, and the interaction matches the magnitude of the main effects but is negative. The focal and reference groups in the present study fall under the same common metric, meaning they share the same model and Q-matrix. As a result, the EMQ method does not require matching variables or scale purification (Qiu et al., 2019).

To model the DIF effect in the LCDM, we take the item parameters of LCDM with DINA constraints as an example. The DINA model contains two item-level parameters: guessing and slip. The guessing parameter g_j specifies the probability of success on item j for those examinees who have not mastered all required attributes; the slip parameter s_j specifies the probability of failure on item j for those who have mastered all required attributes. Corresponding to the parameters in the LCDM and considering the DIF effect, the item parameters are as follows:

$$g_{jc} = \frac{\exp\left(\lambda_{j,0} + g_{djc}\right)}{1 + \exp\left(\lambda_{j,0} + g_{djc}\right)}, \tag{4}$$

$$s_{jc} = 1 - \frac{\exp\left(\lambda_{j,0} + \lambda_{j,1} + s_{djc}\right)}{1 + \exp\left(\lambda_{j,0} + \lambda_{j,1} + s_{djc}\right)}, \tag{5}$$

where g_{jc} and s_{jc} are the guessing and slipping parameters for item j and group c; the $\lambda_{j,0}$ and $\lambda_{j,1}$ parameters are the intercept and the highest-way interaction effect parameters, respectively, and can be thought of as a mean parameter across the two groups. The g_{djc} and s_{djc} parameters are indicative of the deviation from the mean parameter (i.e., the $\lambda_{j,0}$ and $\lambda_{j,1}$ parameters) between the two groups for item j; in the present study only two groups are considered, thus for computation purposes only the DIF parameters for the reference group (i.e., the g_{djc} and s_{djc}; where c = 1 means the reference group and c = 2 belong to the focal group) were freely estimated and the DIF parameters for reference group plus the DIF parameters for focal groups were constrained as zero (i.e., $g_{dj2} = [-1] \times g_{dj1}$). The same constrained way can be found in very popular commercial DIF software (e.g., ConQuest; Wu et al., 2007).

To obtain an estimate for DIF, it is necessary to estimate the s_{dj} and g_{dj} for each group and their corresponding $100\left(1 - \alpha\right)\%$. CI is computed to determine if the two noise parameters of a specific group are significantly different from those of another group (i.e., to check if the range contains 0). The result is the same with the use of a Wald-test estimate.

The results of statistical hypothesis testing are prone to vulnerability due to a large sample size. To quantify the magnitude of DIF contamination, we can use the mean item difficulty difference (MIDD; Wang & Yeh, 2003) between reference and focus groups as an index for identifying DIF items. The MIDD is directly associated with the signed Area measure proposed by Raju (Wang & Su, 2004a). Hence, applying the concept of MIDD within the framework of CDMs results in:

$$\text{MIDD} = \lambda_{j,0,R} - \lambda_{j,0,F} \tag{6}$$

where the $\lambda_{j,0,R}$ and $\lambda_{j,0,F}$ denote the mean item guessing of the reference and focal groups, respectively.

The same concept can be applied to the average item slippage between the reference and focal groups. Wang (2008) suggested that a DIF magnitude of 0.5 logits could be considered as a cut-off point for identifying DIF. Given that the item parameters are on a logit scale, a significant DIF can be determined using the empirical cut-off value of 0.5 logit.

Given the different constraints of the sub-models in the LCDM-based DIF model, it is noted that there are different meanings when interpreting the results of DIF parameters within different models. For example, in the DINA model, a positive value for s_{dj} indicates that the item favors the focal group for examinees mastering all attributes required by item j, and a positive value for g_{dj} indicates that the item favors the reference group for examinees who have not mastered at least one of the attributes required by item j. However, in the LCDM with DINO-like constraints, a positive value for s_{dj} indicates that the item favors the focal group for examinees mastering at least those attributes required by item j, and a positive value for g_{dj} indicates that the item favors the reference group for examinees who have not mastered one of the attributes required by item j. Thus, there are four combinations of s_{dj} and g_{dj} to form uniform and non-uniform DIF types:

1. Both s_{dj} and g_{dj} are positive, $s_{dj} > 0$ or $s_{Fj} - s_{Rj} < 0$; $g_{dj} > 0$ or $g_{Fj} - g_{Rj} > 0$. That is one of the uniform DIF, the item favors the focal group for both masters

and non-masters.

2. Both s_{dj} and g_{dj} are negative, $s_{dj} < 0$ or $s_{Fj} - s_{Rj} > 0$; $g_{dj} < 0$ or $g_{Fj} - g_{Rj} < 0$. That is another one of uniform DIF, where the item favors the reference group for masters and non-masters.
3. s_{dj} is positive, and g_{dj} is negative, $s_{dj} > 0$ or $s_{Fj} - s_{Rj} < 0$; $g_{dj} < 0$ or $g_{Fj} - g_{Rj} < 0$. That is one of the non-uniform DIF, the item favors the masters in the focal group but favors non-masters in the reference group.
4. s_{dj} is negative, and g_{dj} is positive, $s_{dj} < 0$ or $s_{Fj} - s_{Rj} > 0$; $g_{dj} > 0$ or $g_{Fj} - g_{Rj} > 0$. That is another one of the non-uniform DIF, where the item favors masters in the reference group but favors non-masters in the focal group.

The Purification Procedures for the MH and LR Methods

Purification is a procedure designed to identify a set of non-DIF items from the instrument under evaluation to use as a matching criterion in DIF detection. Lord (1980) suggested the use of the scale purification process to mitigate the impact of DIF contamination embedded in an internal matching variable on DIF assessment. Concurrently, a similar concept, DIF-free-then-DIF (DFTD), was proposed by Wang (2008). This implies that when initiating DIF detection, one needs to ensure that at least a set of items in the test are DIF-free. These DIF-free items are then used as matching criteria to detect DIF.

Scale purification has been employed in both IRT-based and non-IRT-based DIF detection methods. These include the iterative linking IRT-based method (Candell & Drasgow, 1988), a two-stage version of MH where the total score is employed as the matching criterion for initial DIF screening. A new score, solely composed of items not exhibiting DIF, is utilized for the final analysis (e.g., Holland & Thayer, 1988). The purification procedure has also been implemented in Mantel method as well as the generalized MH method, (e.g., Wang & Su, 2004a, 2004b). Additionally, it has been employed in the iterative logistic regression (French & Maller, 2007) and the iterative constant item method (ICI; e.g., Wang et al., 2009). Empirical evidence backs the use of purification with both IRT and non-IRT-DIF methods (Candell & Drasgow, 1988; Holland & Thayer, 1988).

In the current study, both the MH and LR methods were implemented. The purification procedure to iteratively elimination of DIF items for these methods is detailed as follows:

1. Test all items one by one and conduct LR / MH analysis for all items (N) with the total summed score as the matching criterion.
2. Identify DIF items (n) based on set criteria.
3. Rerun the analysis for all items with N – n total score as the matching criterion (omitting the items from the set obtained in Step 2) and identify DIF items.
4. Rerun the analysis for all items with N – n total scores (omitting the items based on Step 3) as the matching criterion.

Continue Steps 3 and 4 until the same set of DIF items is identified in two consecutive analyses or no other items are identified, so that the purification process is stopped.

Method

Design of Simulation Study I

To evaluate the parameter recovery of the LCDM-DIF method with compensatory and non-compensatory models, we generated simulated item responses in line with the LCDMs with DINA- and DINO-like constraints. These simulated data sets helped maintain direct control over certain factors, such as the structure of the Q-Matrix, the DIF percentage, and the DIF amount that influences DIF detection. Nevertheless, conditions were simulated to approximate real-world situations to ensure the generalizability of the results.

Controlled Variables

We constructed a single 20 × 5 Q-matrix that balanced complexity and effectiveness. This Q-matrix was similar to those used in prior studies (e.g., de la Torre & Douglas, 2004; Huang, 2018; Li, 2008). For the initial five items, each item was simulated as if estimating a single attribute; for items 6 to 15, each item was simulated as if estimating two attributes; and for items 16 to 20, each item was simulated as if estimating three attributes. To prevent interference with the inference of the test length effect, which was manipulated in this study, the same Q-matrix structure was repeated twice and thrice with increasing test length.

The sample size was set at 1,000 per group. In the present study, we assumed the attribute vector of α is based on a broadly defined latent trait resembling the θ of the item response model. The conditional independence assumption is similar to the higher-order DINA model (de la Torre & Douglas, 2004). Thus, in the present study, the θ is sampled from normal distribution with the mean and variances set 0 and 1. This method can be found in previous CDM studies (de la Torre & Douglas, 2004; Henson et al., 2009). Attribute difficulty in CDMs represents the level of challenge associated with specific attributes. Varying the range of attribute difficulty in simulation design can provide insights into model evaluation, bias detection, person classification, and item quality assessment. The range of attribute difficulty and item parameters were set according to previous CDM studies (de la Torre & Douglas, 2004). In the current study, the range of attribute difficulty was set at [−1.5, −1.0, −1.0, −0.5, −0.5] for both groups; the intercept and slope parameters were randomly generated from *uniform* (−2.2, −0.85). As such, the guessing and slip parameters in the DINA and DINO models fell within the range of (.1, .3) concerning probability.

According to Li's (2008) simulation study, the DIF amount was set at .10, providing sufficient power. Therefore, in simulation study 1, we set the DIF magnitude at .54 logit, close to the .10 difference between subgroups concerning probability at .10. Thus, in the present study, the slip and guessing parameters in the focal group were formed by adding or subtracting .27 logit from the values for the reference group. Finally, the percentage of DIF items was set to 20% to represent a well-developed test in real-world situations, a level also used in previous DIF studies within CDMs (Li, 2008; Li & Wang, 2015; Zhang, 2006).

Manipulated Variables

Three factors were manipulated:

1. Group distributions: We simulated equal and non-equal examinee distributions, where the non-equal distribution was simulated by setting the normal focal group distribution one standard deviation lower than the reference group.
2. Test lengths (20/40/60), which represent short, moderate, and long tests, respectively.
3. DIF patterns: No DIF, provide baseline information for comparing Type I error rates; One-sided DIF, where all DIF items were adjusted to favor the reference group by decreasing the slip parameter and increasing the guessing parameter by equal amounts (DIF magnitude set to 0.27 logit); and Balanced DIF, in which half of the DIF items favored the reference group, and the rest favored the focal group.

These settings were chosen to reflect actual test data, as used in previous DIF detection studies (e.g., Finch & French, 2007; Huang, 2014; Li, 2008; Roussos & Stout, 1996; Shih & Wang, 2009).

The recovery analysis bore three aspects in focus: recovery of the simulated item parameters, attribute difficulty parameters, and attribute mastery classifications. Parameter recovery was measured considering two variables: bias and root mean square error (RMSE) for each condition. Additionally, the accuracy of attribute recovery estimation was gauged using the proportion correctly classified (PC; de la Torre & Douglas, 2004). This was utilized to scrutinize the accuracy in recapturing latent classes using the maximum posterior probability for each attribute (ACA) and the overall profile classification accuracy (PCA).

Design of Simulation Study II

Simulation Study II was designed to compare the efficacy of a model-based method against two non-parametric DIF detection methods. Data was generated using LCDM, along with DINA and DINO constraints. Consequently, factors like test length and the Q-matrix were fixed and simulated in Study II to achieve accurate results.

Controlled Variables

We employed a 20-item test to represent a short-length test. We chose this test length to mirror most diagnostic assessments used in previous research (e.g., de la Torre & Douglas, 2008; Lee et al., 2011). Additionally, the structure of the Q-matrix, item parameters, and attribute difficulty were controlled and simulated as in Study I. Finally, we simulated all DIF items to favor the reference group to demonstrate the effectiveness of different DIF detection methods.

Five variables were manipulated in this study: (a) Group distributions: As created in simulation one; (b) DIF percentage (0%, 10%, 20%, and 30%), signifying varying levels of DIF items on a test; (c) Sample size (F500/R500, F1000/R1000, F1000/R2000), representing both small and large sample sizes, as well as imbalanced subgroup ratios; (d) DIF magnitude values of 0.4, 0.6, and 0.8 logits, correspond to small, moderate, and significant differences; and lastly (e) DIF method (MH/LR/Model-based).

The values for these simulations were based on previous DIF detection studies (Finch & French, 2007; French & Maller, 2007; Huang, 2014; Rogers & Swaminathan, 1993; Shih & Wang, 2009).

Analysis

Numerous CDMs have been developed according to various diagnostic requirements, each estimable through different methods (e.g., expectation maximization, EM, or MCMC) and software (refer to the detailed classification in Rupp et al., 2010). Roussos et al. (2007) noted that the statistical information procured from MCMC estimation (a complete posterior distribution) is richer than that obtained from an EM algorithm (an estimate and its standard error). Furthermore, Li and Wang (2015) explored different estimation algorithms (MCMC vs. MML) for item parameter estimations in CDMs, discovering that underestimated item parameters using MML estimation led to an inflation in Type I error rates when the Wald method was implemented in DIF detection. MCMC demonstrated superior parameter recovery and effectiveness in DIF detection compared to the Wald method in a range of simulation studies. Consequently, MCMC estimation was selected for this study. For each condition, 25 replications were simulated to ensure consistency in results. The number of replications was determined by the amount used in previous simulation work utilizing CDMs estimated through an MCMC algorithm (de la Torre & Douglas, 2004) and to compare methods or criteria for accurate DIF detection rates with CDMs (Li, 2008; Zhang, 2006). The MCMC chain for each replication was run for 14,000 iterations, utilizing the WinBUGS 1.4 program (Spiegelhalter et al., 2003) via a Gibbs sampling algorithm. The iteration number was determined based on the convergence diagnostics proposed by Gelman and Rubin (1992). Hence, a conservative burn of 4,000 and an additional 10,000 post-burn-in iterations were used for all conditions. A significance level of $\alpha = .05$ was employed to analyze the Type I error rate. The MH and LR analyses were executed using the free R software package (R Development Core Team, 2011), leveraging the dif-R package (Magis et al., 2011).

Empirical Study

The Trends in International Mathematics and Science Study (TIMSS), which collects manifest information on participants, such as nationality and gender, was used as an empirical example to apply the LCDM-based DIF method in practical analyses. Moreover, the well-developed framework of test construction employed by TIMSS can be transformed appropriately into a Q-matrix, as required by CDMs. Data were taken from Booklets 4 and 5 of the TIMSS 2007 mathematics assessment for fourth-grade students, consisting of 25 items: 15 multiple-choice and 10 constructed response items (Foy & Olson, 2008).

Three reasons informed the choice of Booklets 4 and 5 for the current study. First, they encompass the most significant number of dichotomously scored items, which the DINA and DINO models require. Second, the selected dataset aligns with the overall domains established by the test developers. Lastly, this same dataset was chosen to fit the DINA model in a previous study (Lee et al., 2011).

Lee et al. (2011), relying on the TIMSS 2007 mathematics framework (Mullis et al., 2005), identified nine topic areas within the content domains as indicated in the 2007 TIMSS framework that measured 15 attributes. They created the Q-matrix, specifically for Booklets 4 and 5, based on these 15 attributes through expert domain reviewer discussions. However, due to the limited size of the item bank (25 items) for analyzing such a vast scope of attributes, certain attributes were only tested by a single item. As a result, this study reduced the attributes from 15 to 9 according to the

Table 1

TIMSS 2007 Fourth Grade Mathematics Q-Matrix

		Number				Geometric shapes & measurement			Data display	
Item		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	M041052	1	0	0	0	0	0	0	0	0
2	M041056	0	1	0	0	0	0	0	0	0
3	M041069	1	1	0	0	0	0	0	0	0
4	M041076	1	1	0	0	0	0	0	0	0
5	M041281	1	0	0	1	0	0	0	0	0
6	M041164	0	0	0	0	0	1	1	0	0
7	M041146	0	0	0	0	1	1	1	0	0
8	M041152	1	0	0	0	0	0	0	0	0
9	M041258A	0	0	0	0	0	1	0	0	0
10	M041258B	0	0	0	0	1	1	0	0	0
11	M041131	1	0	0	0	1	0	0	0	0
12	M041275	1	0	0	0	0	0	0	1	1
13	M041186	1	0	0	0	0	0	0	1	0
14	M041336	1	1	0	0	0	0	0	1	0
15	M031303	1	0	0	0	0	0	0	0	0
16	M031309	1	0	0	0	0	0	0	0	0
17	M031245	1	0	1	0	0	0	0	0	0
18	M031242A	1	0	0	1	0	0	0	0	0
19	M031242B	1	0	0	0	0	0	0	1	0
20	M031242C	1	0	0	1	0	0	0	1	0
21	M031247	1	0	1	0	0	0	0	0	0
22	M031219	0	0	0	0	0	1	1	0	0
23	M031173	1	0	0	0	0	0	0	0	0
24	M031085	1	0	0	0	0	1	0	0	0
25	M031172	1	0	0	0	0	0	0	1	1

nine topic areas within the content domains, as indicated in the 2007 TIMSS framework (Mullis et al., 2005). The entries of the item content Q-matrix are presented in Table 1. In the Q-matrix structure, all items measured at least one attribute, and some items tested two or three attributes, forming a so-called 'complex structure.'

Additionally, merely considering the number of examinees who complete Booklets 4 and 5 simultaneously from a single country may not provide sufficient data for analysis. Thus, the study incorporated data from 858 examinees in fourth grade from two countries (Taiwan and the United States) who took Booklets 4 and 5 of the TIMSS 2007 fourth-grade mathematics assessment. This comprised 452 girls and 406 boys. Historical studies exploring patterns of gender difference in mathematics concluded that factors such as the context of the item and the presentation method can impact gender performance disparities in mathematics (Harris & Carlton, 1993; Ryan & Chiu, 2001). These factors tend to be more related to issues of test equality than innate ability. Bearing in mind that the analysis of extensive data is frequently used as guidance for practitioners, it is essential to identify the item characteristics that favor different genders. Consequently, gender DIF was explored using this dataset. This study compared LCDM-based DIF with different constraints (i.e., DINA and DINO) using Akaike's information criterion (AIC) and a corrected Bayesian information criterion (BIC). The model with a lower AIC or BIC value indicated a better fit to the data (Akaike, 1974; Schwarz, 1978).

Results

Parameter Recovery of the LCDM-DIF Method: DINA and DINO Example

Due to space constraints, Table 2 presents the mean RMSE and mean absolute bias for the item parameter estimates for the applications of the DINA and DINO models. The mean RMSEs ranged from .08 to .14 (M = .10), and the mean absolute bias values ranged from .00 to .04 (M = .01) for the DINA example. For the DINO example, the mean RMSEs ranged from .06 to .39 (M = .15), and the mean absolute bias values ranged from .00 to .20 (M = 0.03). These results for mean RMSE and mean absolute bias appeared satisfactory and are similar to the findings of previous studies using MCMC methods which were implemented in the CDM model under similar conditions (e.g., Li & Wang, 2015).

Table 2

Mean RMSE and Mean Absolute Bias for Item Parameters in the Parameter Recovery Study

T.L.	DIF pattern	Ability	DINA-like								DINO-like							
			RMSE				Bias				RMSE				Bias			
			$\lambda_{j.0}$	$\lambda_{j.1}$	DIF_s	DIF_g	$\lambda_{j.0}$	$\lambda_{j.1}$	DIF_s	DIF_g	$\lambda_{j.0}$	$\lambda_{j.1}$	DIF_s	DIF_g	$\lambda_{j.0}$	$\lambda_{j.1}$	DIF_s	DIF_g
20	NO	equal	0.14	0.08	0.08	0.13	0.04	0.00	0.00	0.01	0.36	0.12	0.09	0.29	0.17	0.03	0.01	0.01
		unequal	0.11	0.12	0.11	0.11	0.03	0.02	0.01	0.00	0.25	0.11	0.09	0.25	0.12	0.02	0.01	0.09
	one-sided	equal	0.13	0.09	0.08	0.12	0.03	0.01	0.00	0.00	0.39	0.10	0.08	0.34	0.20	0.02	0.00	0.00
		unequal	0.11	0.11	0.12	0.10	0.02	0.02	0.02	0.01	0.25	0.11	0.10	0.26	0.10	0.03	0.01	0.07
	balanced	equal	0.13	0.09	0.08	0.12	0.04	0.01	0.01	0.01	0.36	0.12	0.08	0.31	0.17	0.02	0.00	0.00
		unequal	0.12	0.12	0.08	0.13	0.03	0.03	0.02	0.01	0.26	0.12	0.09	0.25	0.12	0.03	0.01	0.07
40	NO	equal	0.11	0.08	0.08	0.10	0.03	0.01	0.00	0.00	0.23	0.07	0.07	0.21	0.08	0.01	0.00	0.00
		unequal	0.09	0.10	0.10	0.09	0.02	0.01	0.01	0.01	0.17	0.07	0.07	0.16	0.04	0.01	0.00	0.03
	one-sided	equal	0.11	0.08	0.08	0.10	0.03	0.01	0.00	0.00	0.22	0.07	0.07	0.21	0.08	0.01	0.00	0.00
		unequal	0.09	0.10	0.10	0.09	0.02	0.02	0.01	0.01	0.18	0.07	0.07	0.16	0.05	0.01	0.00	0.02
	balanced	equal	0.11	0.08	0.08	0.11	0.02	0.01	0.01	0.00	0.23	0.07	0.07	0.21	0.07	0.01	0.00	0.01
		unequal	0.10	0.10	0.10	0.09	0.02	0.02	0.01	0.01	0.19	0.07	0.07	0.18	0.05	0.01	0.00	0.03
60	NO	equal	0.10	0.08	0.08	0.10	0.02	0.01	0.00	0.00	0.20	0.07	0.06	0.20	0.06	0.01	0.00	0.00
		unequal	0.08	0.10	0.10	0.09	0.01	0.01	0.00	0.00	0.16	0.07	0.07	0.16	0.04	0.01	0.00	0.02
	one-sided	equal	0.10	0.08	0.08	0.10	0.02	0.01	0.00	0.00	0.20	0.07	0.06	0.18	0.06	0.01	0.00	0.00
		unequal	0.09	0.10	0.10	0.09	0.01	0.02	0.00	0.00	0.15	0.07	0.07	0.15	0.03	0.01	0.00	0.01
	balanced	equal	0.10	0.08	0.08	0.10	0.02	0.01	0.00	0.00	0.20	0.07	0.07	0.19	0.07	0.01	0.00	0.00
		unequal	0.09	0.10	0.10	0.09	0.01	0.01	0.00	0.00	0.17	0.07	0.07	0.16	0.04	0.01	0.00	0.03

Note. BA denotes balanced DIF pattern; ON denotes one-sided DIF pattern;
NO denotes no DIF pattern

It can be seen that both intercept $\lambda_{j.0}$ and slope parameters $\lambda_{j.1}$ were recovered well. DIF patterns and test length did not affect the recovery of intercept and slope parameters. The recovery of $\lambda_{j.1}$ was slightly better than the recovery of $\lambda_{j.0}$. This may have occurred because the attribute difficulty generated ranges from −1.5 to −0.5, compared to ability distribution for both focal and reference groups generated in this study, these items were set as easy, and one can expect the sample size for masters to be larger than non-masters. The estimation of $\lambda_{j.1}$ parameters is related to the master group; thus, the recovery was slightly better than the recovery. Besides, the ability distribution difference seems to have impacted the recovery of item parameters. The RMSE was slightly lower in the unequal ability distribution than in the equal ability. On the other hand, the RMSE of $\lambda_{j.1}$ was slightly higher in the unequal ability distribution than in the equal ability. This is because the non-masters increased for the unequal ability distribution and increased the accuracy of the estimation of $\lambda_{j.0}$ parameters. The population of masters decreased in the unequal ability distribution and thus decreased the accuracy of the estimation of $\lambda_{j.1}$ parameters.

A similar pattern was detected for the DINO-like model. In short, the parameter recovery of DINO shows poorer recovery when intercept parameters are compared to the DINA-like model. This might be due to a characteristic of the compensatory model: the guessing parameter is related to examinees who have not mastered at least one attribute tested in an item. The parameter estimation may not be accurate when insufficient sample sizes and test lengths offer information.

The recovery results from the two LCDM sub-models were generally favorable. Upon close inspection of the DINA and DINO data's recovery analysis, items with a simple structure (i.e., those that test only one attribute or fewer) demonstrated better recovery of the slipping parameter. However, items testing a single attribute showed a lower estimation of the guessing parameter than items testing more than one attribute (due to space constraints, this result could not be presented in Table 1, but is available upon request). This may stem from the accessibility of items with simpler structures compared to those with complex structures. Consequently, the sample size of masters was more substantial than that of non-masters, which enhanced the estimation of slipping parameters. Hence, in simple structure items, the DIF-s estimation was more accurate than in complex structure items. Conversely, DIF-g estimation was more precise in complex structures. Additionally, the difference in ability distribution influenced the DIF-s and DIF-g estimations. Equal ability distribution lowered DIF-s RMSE compared to unequal ability distribution, and vice versa for DIF-g. This result is logical since the estimation of DIF-s and DIF-g is tightly connected to Table 1's item parameter estimation. Moreover, longer test lengths showed marginally better parameter recovery than shorter ones across six different DIF scenarios.

Table 3 presents the attribute-level classification accuracy (ACA) and the overall PCA for DINA and DINO models across all conditions. It was observed that the test length had a minor impact on the recovery of ACA and PCA. As the test length increased, the correct classification rates also increased. The data indicated that the overall consistency of the examinees' attribute vector also increased with test length, particularly transitioning from the short to the median test length condition. This might be because the extended test length provided more information, subsequently improving the estimation of attribute mastery. Furthermore, the ACA rates for the DINA model were higher than .92 across all conditions and higher than .80 for the DINO model, suggesting accurate estimates of the examinee attribute profile score estimates using the two proposed models under these conditions. Similar to findings from previous CDM studies, the PCA rates have typically been much lower than ACA rates (Liu et al., 2017; Templin & Bradshaw, 2014). Paulsen et al. (2020) discovered that a DIF magnitude of 0.10 could result in inadequate ACA and PCA rates of the DINA model (i.e., ACA rates lower than .70 with the same ability distribution and below .50 with unequal ability distribution). It is worth noting that when the LCDM-DIF was implemented, it could provide satisfactory attribute-level classification rates and the entire profile classification rates under similar DIF conditions (i.e., the ACA rates were higher than .97 for equal ability distribution and higher than .95 for unequal ability distribution using the DINA example).

As anticipated, the model-based DIF method excelled in parameter recovery across all manipulated conditions, particularly when the simplest DINA model was exhibited. Due to the characteristics of the DINO-constrained model, a longer test length provides ample information on items, benefiting the estimation of the ACA.

Comparison of LCDM-DIF Method and Traditional Method with Purification Procedure

Type I Error Study

The Type I error rate was the proportion of detections for all items across all conditions not simulated to exhibit DIF. In this study, each condition entailed 25 replications, each having 20 items. Under the "No DIF" condition, all 20 items were simulated without any DIF-g or DIF-s. The Type I error rate for DIF-g was measured as the percentage of detected DIF-g out of 500 (25 replications × 20 items) no-DIF counts. The same was done for the DIF-s Type I error. This study simulated varying DIF percentages, leading to the empirical Type I error rate for DIF-g and DIF-s depending on the simulated DIF percentages. The ensuing results illustrate the empirical Type I error rate for each of the 80 conditions, calculated as the percentage of DIF items for non-DIF items across 25 replications. To estimate the power of DIF-g or DIF-s, the Type I error for DIF-g or DIF-s must be controlled. The nominal level for these Type I error rates was set at α = .05. Accepted Type I error rates fell within the range of Bradley's (1978) criterion of .025 to .075. Table 3 shows that the model-based method could keep Type I error rates within a suitable range when detecting DIF under manipulated conditions. The proposed model-based DIF detection method consistently controlled the Type I error rate tolerably, no matter the sample size, ability distribution, DIF percentages, or magnitudes. Compared with the MH and LR techniques, the parametric approach sustained a better Type I error control, regardless of the ability distribution and DIF percentage. Comparable outcomes were found with data generated from the DINO model.

Table 4 illustrates a comparison between the Type I error rates of DIF detection using the MH and LR methods, based on total score matching both with and without purification procedures. When the data set was formed using the DINA model, the MH method successfully controlled Type I errors only when groups came from equivalent ability distributions. However, under different conditions, both non-parametric DIF detection methods failed to maintain Type I error control. As the magnitude of DIF augmented, both methods caused the inflation of Type I errors. Broadly speaking, the LR method preserved slightly better Type I error control than the MH method, regardless of the ability distribution and DIF percentage. Nevertheless, differences in ability affected both detection methods. Therefore, unequal abilities lessen the performance of both the MH and LR methods.

One explanation centers on the MH and LR methods' dependence on the observed raw score or number-correct, X. When X inadequately represents θ, and coupled with group impact, the expected mean of θ given X can differ across groups. Stated differently, examinees matched based on observed scores may not necessarily align with the latent score (DeMars, 2010). Thus, higher rates of Type I errors can arise when there are significant group disparities in true proficiency. In the current study, for instance, the odds of the reference group were higher than those of the focal group

Table 3

Attribute Classification Accuracy Rates Across Conditions: DINA and DINO Model

	DINA model																	
Test length	**20**						**40**						**60**					
	Equal			**Unequal**			**Equal**			**Unequal**			**Equal**			**Unequal**		
Classification	BA	ON	NO	BA	ON	NO	BA	ON	NO	BA	ON	NO	BA	ON	NO	BA	ON	NO
Attribute 1	0.93	0.93	0.93	0.89	0.89	0.89	0.97	0.97	0.97	0.95	0.95	0.95	0.98	0.98	0.96	0.97	0.97	0.97
Attribute 2	0.94	0.95	0.95	0.92	0.93	0.92	0.98	0.98	0.98	0.97	0.97	0.97	0.99	0.99	0.99	0.97	0.98	0.98
Attribute 3	0.93	0.93	0.93	0.91	0.91	0.90	0.97	0.98	0.98	0.96	0.96	0.96	0.99	0.99	0.99	0.98	0.98	0.98
Attribute 4	0.92	0.92	0.92	0.90	0.90	0.90	0.97	0.97	0.97	0.96	0.96	0.96	0.99	0.99	0.99	0.98	0.98	0.98
Attribute 5	0.95	0.95	0.95	0.94	0.94	0.94	0.98	0.98	0.98	0.98	0.97	0.98	0.99	0.99	0.99	0.99	0.99	0.99
Overall consistency	0.76	0.77	0.76	0.68	0.68	0.68	0.90	0.90	0.90	0.85	0.85	0.85	0.95	0.95	0.95	0.91	0.91	0.91
	DINO model																	
	Equal			**Unequal**			**Equal**			**Unequal**			**Equal**			**Unequal**		
	BA	ON	NO	BA	ON	NO	BA	ON	NO	BA	ON	NO	BA	ON	NO	BA	ON	NO
Attribute 1	0.93	0.93	0.93	0.93	0.93	0.93	0.95	0.93	0.96	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98
Attribute 2	0.90	0.90	0.90	0.91	0.91	0.91	0.93	0.92	0.95	0.96	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.97
Attribute 3	0.88	0.88	0.88	0.89	0.89	0.89	0.92	0.91	0.94	0.95	0.95	0.95	0.97	0.97	0.97	0.97	0.97	0.97
Attribute 4	0.80	0.80	0.80	0.82	0.82	0.82	0.85	0.86	0.88	0.90	0.90	0.90	0.93	0.92	0.92	0.94	0.94	0.94
Attribute 5	0.89	0.89	0.89	0.90	0.90	0.90	0.90	0.91	0.94	0.95	0.95	0.94	0.96	0.96	0.96	0.97	0.97	0.96
Overall consistency	0.55	0.56	0.55	0.59	0.59	0.59	0.68	0.70	0.73	0.77	0.77	0.77	0.82	0.82	0.82	0.85	0.85	0.85

Table 4

Marginal Mean Type I Error Rates for Model-Based and Two Nonparametric DIF Methods

		Data derived from the DINA model						**Data derived from the DINO model**					
		Model-based		**MH**	**MH-P**	**LR**	**LR-P**	**Model-based**		**MH**	**MH-P**	**LR**	**LR-P**
		DIF-s	**DIF-g**	**Mean**	**Mean**	**Mean**	**Mean**	**DIF-s**	**DIF-g**	**Mean**	**Mean**	**Mean**	**Mean**
Sample size	F500/R500	0.06	0.04	0.12	0.07	0.12	0.09	0.06	0.03	0.14	0.09	0.15	0.12
	F500/R1000	0.05	0.05	0.15	0.07	0.17	0.11	0.05	0.04	0.18	0.09	0.20	0.16
	F1000/R1000	0.05	0.05	0.20	0.08	0.23	0.14	0.05	0.04	0.22	0.10	0.26	0.18
	F1000/R2000	0.05	0.05	0.23	0.10	0.27	0.18	0.05	0.04	0.26	0.11	0.31	0.23
Ability distribution	Equal	0.05	0.04	0.18	0.05	0.14	0.05	0.05	0.04	0.20	0.06	0.15	0.06
	Unequal	0.05	0.05	0.17	0.11	0.25	0.21	0.05	0.04	0.20	0.14	0.31	0.28
DIF percentage	0%	0.05	0.04	0.07	0.07	0.12	0.12	0.05	0.04	0.09	0.10	0.13	0.15
	10%	0.05	0.04	0.09	0.08	0.14	0.13	0.05	0.04	0.12	0.09	0.16	0.16
	20%	0.05	0.05	0.19	0.08	0.20	0.13	0.05	0.04	0.20	0.09	0.23	0.16
	30%	0.05	0.05	0.29	0.09	0.27	0.13	0.05	0.04	0.32	0.10	0.33	0.20
DIF magnitude	Large	0.05	0.05	0.27	0.08	0.26	0.12	0.05	0.04	0.30	0.09	0.30	0.17
	Median	0.05	0.05	0.18	0.08	0.20	0.14	0.05	0.04	0.21	0.10	0.23	0.17
	Small	0.05	0.05	0.11	0.09	0.15	0.13	0.05	0.04	0.14	0.10	0.18	0.18

due to distinct group distributions, leading to the rejection of the null hypothesis with the MH method. However, this decision might be misguided as the substantial disparities in the ability of subgroups could have influenced the computational results. Furthermore, the prominent group differences might have inflated Type I errors due to the misuse of the parametric form in modeling the item responses (DeMars, 2010).

Power Study

In the power analysis, we empirically estimated the power of the statistic by calculating the percentage of identified DIF-g or DIF-s when simulating DIF-g or DIF-s, respectively. The study involved manipulating three levels of DIF item percentages. Specifically, in the 10 percent test condition, DIF-g or DIF-s was simulated for Items 6 and 14. In the 20 percent test condition, DIF-g or DIF-s was simulated for Items 1, 6, 14, and 19. In the 30 percent test condition, DIF-g or DIF-s was simulated for Items 1, 6, 12, 14, 18, and 19. It is worth noting that DIF-g consistently exhibited lower power than DIF-s across all simulated conditions. Table 5 presents the marginal means and power ranges for both DIF-g and DIF-s based on the two sub-models.

DIF-g exhibited higher power when a substantial sample size and DIF magnitude were combined, mirroring the same pattern for DIF-s. These findings highlight the influence of sample size and DIF magnitude on power rates. Notably, power rates for DIF-g generally stayed below .70, except for conditions featuring a sample size of F1000/R2000 and substantial DIF magnitude. This underscores that the model-based approach yields sufficient power rates only in scenarios characterized by ample sample size (e.g., F500/R1000 with the DINA model or F1000/R2000 with the DINO model) and appropriate DIF magnitude.

Moreover, a reverse effect was observed for ability distributions regarding DIF-g and DIF-s. Unequal ability distribution led to higher power for DIF-g, while an equal ability distribution led to higher power for DIF-s. The power rates for DIF-g and DIF-s seemed to correlate with the sample sizes of non-masters and masters. Conditions with larger sample sizes and unequal ability distributions produced more non-masters, enhancing the precision of DIF-g estimation and, therefore, increasing its power. On the other hand, conditions with more masters led to a more precise DIF-s estimation and a higher power for DIF-s. This pattern was maintained when subgroup ratios were unequal (e.g., F500/R1000 and F1000/R2000) and explained the suboptimal power rates observed for DIF-g with the DINO model. In the DINO model, examinees who mastered only a single attribute required by an item could boost correction rates, resulting in more masters and fewer non-masters.

Table 5

Marginal Mean Power Rates for Model-Based and Two Nonparametric DIF Methods

	DINA	Data derived from the DINA model						Data derived from the DINO model					
		Model-based		MH	MH-P	LR	LR-P	Model-based		MH	MH-P	LR	LR-P
		DIF-s	DIF-g	Mean	Mean	Mean	Mean	DIF-s	DIF-g	Mean	Mean	Mean	Mean
Sample size	F500/R500	0.60	0.48	0.71	0.78	0.66	0.73	0.82	0.15	0.70	0.78	0.65	0.71
	F500/R1000	0.72	0.60	0.79	0.85	0.77	0.83	0.88	0.22	0.80	0.86	0.74	0.78
	F1000/R1000	0.83	0.72	0.89	0.94	0.87	0.92	0.96	0.30	0.89	0.92	0.84	0.87
	F1000/R2000	0.88	0.83	0.94	0.96	0.92	0.93	0.98	0.44	0.93	0.96	0.90	0.92
Ability distribution	Equal	0.83	0.61	0.85	0.90	0.82	0.88	0.93	0.20	0.89	0.93	0.86	0.90
	Unequal	0.68	0.71	0.81	0.86	0.79	0.82	0.89	0.35	0.77	0.83	0.71	0.74
DIF percentage	10%	0.75	0.67	0.90	0.91	0.88	0.89	0.92	0.27	0.89	0.91	0.84	0.85
	20%	0.78	0.64	0.86	0.91	0.84	0.87	0.91	0.30	0.84	0.89	0.78	0.83
	30%	0.74	0.67	0.74	0.83	0.71	0.79	0.90	0.26	0.76	0.85	0.72	0.78
DIF magnitude	Large	0.93	0.84	0.96	0.99	0.95	0.99	0.99	0.40	0.97	0.99	0.96	0.98
	Median	0.79	0.69	0.89	0.93	0.86	0.89	0.95	0.27	0.88	0.95	0.84	0.88
	Small	0.55	0.45	0.64	0.72	0.61	0.67	0.79	0.14	0.63	0.70	0.55	0.59

Additionally, the number of DIF items appeared to influence power rates in DIF detection. Fewer DIF items were less detectable, resulting in lower power rates compared to scenarios with more DIF items. This observation is consistent with prior research suggesting that power rates rise with increased DIF magnitude (Zhang, 2006).

In conclusion, the model-based DIF detection method demonstrated the ability of the DINA model to maintain sufficient power in detecting both DIF-s and DIF-g across the majority of experimental conditions. Nonetheless, power rates for detecting DIF-g, when applying the DINO model, consistently fell below .70 in most scenarios. Ultimately, the success of the LCDM-based DIF detection method depends on the balance between DIF magnitude and sample size.

The non-parametric analysis revealed that the power rates for the MH method were influenced by sample size, DIF magnitude, and DIF percentage. A significant trend emerged: as the DIF magnitude increased from small to large, the power rate also increased correspondingly. Additionally, sample size had a noticeable effect, with power rates rising as the sample size increased. On the contrary, introducing a higher percentage of DIF items resulted in decreased power rates for both MH and MH-P methods.

Additionally, following the application of the purification procedure, power rates under the same conditions exceeded those of the MH method. Similar outcomes were noted in the LR method. Both MH and LR methods sustained acceptable power rates only in conditions with a DIF percentage less than 20% and at least a moderate DIF magnitude. Notably, the implementation of the purification procedure resulted in increased power rates under these conditions. However, it is essential to highlight that both methods showed inflated Type I errors, potentially leading to an overestimation of power rates (as indicated in Table 4).

The LCDM-based approach demonstrated superior Type I error control and was suitable for most conditions, effectively highlighting the benefits of the two proposed models in the context of CDM. However, it is crucial to acknowledge a limitation of the model-based approach, specifically, its demand for a comparatively larger sample size than the MH and LR methods. In the present study, when employing the DINA model, power rates dropped below .70 in scenarios involving a blend of a sample size of F500/R500 and either medium or small DIF magnitudes. Yet, as the sample size expanded to F1000/R1000, the power rates enhanced and fell within a reasonable range.

DIF Detection Using Example TIMSS Data

This section presents a practical application of DIF detection procedures, employing a model-based approach alongside two conventional methods: LR and MH methods, with purification steps within the framework of cognitive diagnostic measurement (CDM). The item parameters and examinee parameters for the TIMSS dataset were calibrated using the same software employed in the simulation study.

The MCMC method, utilizing Metropolis-Hastings within the Gibbs Sampler iterations, was employed with a total of 25,000 runs. The initial 5,000 runs were designated as burn-in, stabilizing subsequent iterations. This extensive iteration process was chosen to ensure the robustness and accuracy of the estimated item and examinee parameters.

Results for Real Data Example

We chose the male group as the reference. We evaluated the model fit by comparing the DINO and DINA models. The comparison using the information criteria-based statistics, reveals that the DINA model, with lower AIC (21,140) and BIC values (21,260) than the DINO model (AIC = 21,850; BIC = 21,970), provides a better fit for the dataset. Therefore, it is the preferred choice. This result aligns with a previous study using the same dataset (Lee et al., 2011). The improved fit of the DINA model provided a solid basis for further examination and served as a benchmark for comparing results obtained using the MH and LR methods.

Table 6 shows parameter estimates for slip and guessing. Unlike the ranges used in the simulation study, the slip parameter estimates for this dataset ranged from .001 to .597, while guessing parameter estimates ranged from .024 to .792. These estimates were reasonable, albeit the higher guessing parameters indicated that non-masters had a greater likelihood of a correct response compared to the simulated guessing parameters. The 1-s column showed p-values for masters in both subgroups, ranging from .40 to .99. This pointed to potential skill gaps in the Q-matrix if tiny values were found.

The MH, MH-P, LR, LR-P, and Model-based approaches were utilized to identify DIF items within gender groups. The DIF detection results are presented in Table 7. Notably, a significant χ^2 result in the MH method indicates uniform DIF. Table 6, meanwhile, provides estimates for DIF-g and DIF-s from the model-based method, along with 95% confidence intervals. If the 95% CI for DIF-g and DIF-s does not encompass 0, it signifies significant DIF-g or DIF-s.

Table 6

Item Parameters of the TIMSS Example

	s	**S.E.**	**1-s**	**g**	**S.E.**	**delta**
item1	0.064	0.005	0.936	0.676	0.003	0.259
item2	0.199	0.011	0.801	0.177	0.019	0.624
item3	0.348	0.005	0.652	0.272	0.004	0.380
item4	0.041	0.032	0.959	0.639	0.004	0.321
item5	0.029	0.025	0.971	0.464	0.004	0.506
item6	0.068	0.008	0.932	0.770	0.006	0.161
item7	0.115	0.044	0.885	0.380	0.008	0.504
item8	0.209	0.004	0.791	0.289	0.004	0.503
item9	0.132	0.008	0.868	0.207	0.132	0.661
item10	0.250	0.007	0.750	0.191	0.050	0.559
item11	0.600	0.002	0.400	0.271	0.003	0.129
item12	0.008	0.048	0.992	0.830	0.004	0.162
item13	0.182	0.003	0.818	0.553	0.003	0.265
item14	0.315	0.006	0.685	0.238	0.003	0.447
item15	0.049	0.009	0.951	0.551	0.003	0.400
item16	0.110	0.005	0.890	0.263	0.005	0.627
item17	0.156	0.082	0.844	0.102	0.006	0.742
item18	0.100	0.012	0.900	0.397	0.004	0.503
item19	0.229	0.006	0.771	0.205	0.005	0.566
item20	0.106	0.011	0.894	0.495	0.003	0.399
item21	0.536	0.005	0.464	0.194	0.007	0.270
item22	0.299	0.006	0.701	0.421	0.005	0.280
item23	0.087	0.006	0.913	0.337	0.004	0.576
item24	0.416	0.003	0.584	0.257	0.003	0.327
item25	0.016	0.050	0.984	0.629	0.004	0.356

Table 7

Summary of DIF Detection Results in the TIMSS Example

	MH		*MH-P*		*LR*		*LR-P*		*Model based method*	
	Stat.	ΔMH	*Stat.*	ΔMH	*Stat.*	ΔR^2	*Stat.*	ΔR^2	*DIF-g(95% CI)*	*DIF-s(95% CI)*
i1	5.412 (0.02)	-1.179^{b}								
i2	5.583 (0.02)	-0.939^{a}								
i4	4.014 (0.05)	-0.895^{a}								
i5			4.676 (0.03)	1.060^{b}						
i6	10.140 (0.00)	1.537^{c}	12.600 (0.00)	1.736^{c}	11.260 (0.00)	0.078^{c}	13.215 (0.00)	0.084^{c}		0.394^{c} (0.135, 0.660)
i8	7.468 (0.01)	1.029^{b}	12.145 (0.00)	1.368^{b}	8.395 (0.02)	0.025^{a}	10.990 (0.01)	0.028^{a}	-0.301^{b} (−0.54, −0.067)	
i9			5.273 (0.02)	0.983^{a}						
i11	3.997 (0.05)	-0.707^{a}								
i14	6.526 (0.01)	0.973^{a}	7.224 (0.01)	1.022^{b}			6.997 (0.03)	0.020^{a}		0.283^{a} (0.053, 0.526)
i15	3.871 (0.049)	-0.926^{a}								
i20									0.466^{c} (0.066, 0.976)	
i25	10.751 (0.00)	1.645^{c}	13.711 (0.00)	1.861^{c}	11.269 (0.01)	0.032^{a}	15.008 (0.00)	0.033^{a}		0.355^{c} (0.143, 0.567)

Note. a denotes negligible DIF; b denotes moderate DIF and c denotes large
DIF values with positive sign means favor focal group

In the real data example, predicting DIF patterns is challenging. To distinguish between uniform and non-uniform DIF patterns, one can inspect the signs of DIF-g and DIF-s. When DIF-g and DIF-s share the same signs, it suggests a uniform DIF. In contrast, if DIF-g and DIF-s have different signs, it implies a non-uniform DIF. To ensure that the results are practical and not just statistically significant, it is important to consider effect sizes, as underlined by French and Maller (2007). In the actual data example, the ETS effect size classification was employed for the MH method.

The $\Delta R^2 measure$ was used for the LR method and adopted the effect size classification proposed by Jodoin and Gierl (2001). In addition, for DIF-g and DIF-s, the study adopted an effect size measure, −2.35 times the difference between item difficulties of the reference group and the focal group (Penfield & Camilli, 2007, p. 138). This effect size is similar to $|\Delta\alpha\mathrm{MH}|$ and applied the ETS classification criterion.

Table 7 demonstrates that MH identified nine DIF items, while MH with purification detected six DIF items. Initially, LR found three DIF items, but with the application of purification, it identified four. Moreover, five items displayed either DIF-g or DIF-s. This dataset mirrors the condition in the simulation study, with a similar sample size of 500 per group and equal ability distribution. The mean ability for the reference group is N (0, 1), whereas, for the focal group, it is N (0.04, 1.65).

When comparing these results to the

Type I error rate analysis in a condition with a smaller sample and equal ability distribution, it becomes evident that Type I error rates remain within an acceptable range when data derived from the DINA model is analyzed using the MH and LR methods. This is especially true in conditions with 10% DIF items and small or median magnitudes. The Type I error rates for the MH method inflate under other conditions, as displayed in Table 4. For the LR method, the Type I error rates are just barely maintained within an acceptable range in conditions with small DIF magnitude, with inflation occurring in other conditions. Nevertheless, by applying the purification procedure, we managed to bring these conditions with inflated Type I error rates back within the acceptable range. Furthermore, LR-P achieves a power rate that exceeds 0.70 in conditions with 20% DIF items and small magnitudes, as demonstrated in Table 5. Hence, the detection results for MH-P and LR-P prove more reliable than MH and LR.

In summary, the detection results from MH-P, LR-P, and the model-based methods prove to be more reliable than those from MH and LR. When compared under the same conditions, the model-based method demonstrates greater accuracy, yielding lower Type I error rates in DIF detection. Consequently, the findings from MH-P, LR-P, and the model-based methods were used to summarize similarities, differences, and interpretations.

While several items were identified as having DIF, it is important to note that some of them had negligible or moderate DIF according to the ETS standards and ΔR^2 measure used in this study, when $\left|\Delta\alpha\text{MH}\right| \geq 1.5$ or $\Delta R^2 \geq .070$. Both MH-P and LR-P methods agreed that Item 6 had significance for DIF-g and DIF-s, the study adopted an effect size measure, which is similar to $\left|\Delta\alpha\text{MH}\right|$ and applied the ETS classification criterion. Using this approach, all three methods detected the same item with significant or moderate DIF (Items 6, 8, 25). These items pertained to the content domains of “Geometric Shapes & Measurement,” “Number,” and “Data Display,” respectively.

Items 7 and 25 indicated a bias towards male students, with female students more likely to make errors on these items. Item 8, too, showed a favor for male students, particularly those who were non-masters; they had a higher probability of guessing incorrectly compared to females. An analysis of the item context revealed that items favoring males typically involved figures or geometric problems, whereas items favoring females were placed in real-world contexts and conveyed content via text. This result aligns with previous gender DIF studies on mathematics performance (Lane et al., 1996). There might exist a secondary ability disparity between female and male groups, not assessed by the test, influencing examinee responses when DIF was apparent.

This real data application was designed to illustrate two frequently used DIF detection methods, along with a model-based approach within the cognitive diagnostic framework. In line with the simulation study outcomes, MH-P decreased the number of identified DIF items compared to the MH method. On the other hand, the results of the LR and LR-P methods differed. Just as in the simulation study, the LR method was found to be sensitive to DIF magnitude. However, LR-P, even in scenarios comprising 20% DIF items and a small DIF magnitude, maintained a power rate exceeding 0.70. This could be why Item 14 was only detected by LR-P and not by LR.

Although all three methods singled out the same items as having large DIF magnitudes, the model-based approach stood out in offering a more intricate interpretation of the results compared to other DIF methods. With the model-based method for DIF detection within the cognitive diagnostic framework, practitioners can discern whether DIF ensues due to slipping, guessing, or both. This can spur further investigation into the root causes of DIF.

Conclusions and Discussion

This study conducted a critical evaluation of various DIF detection methods in the context of cognitive diagnostic modeling. Its goal was to bridge the gap between traditional DIF techniques and modern model-based approaches by providing a comprehensive analysis under diverse testing conditions. A significant emphasis was placed on the challenge of addressing contaminated matching criteria, a crucial concern frequently ignored in previous studies. Our findings provide insights into key aspects of DIF detection, thereby illuminating how these methods can be applied in real-world scenarios.

Our findings offer a detailed understanding of DIF detection in cognitive diagnostic assessments. Notably, the LCDM-DIF method demonstrated its effectiveness by accurately identifying DIF patterns, regardless of their complexity, highlighting both one-sided and balanced DIF cases. We found that the simplicity of the Q-matrix design significantly contributes to parameter estimation, with implications for the accuracy of slip and guessing parameters. This insight, drawn from the interaction between the Q-matrix structure and item parameters, serves as valuable guidance for practitioners and emphasizes the need for a well-defined Q-matrix.

Furthermore, our study explored the effects of various testing conditions on DIF detection. A significant observation emphasized the importance of adequate sample size when using the model-based approach, ensuring sturdy Type I error control and power rates. The simulation study results suggest that if the dataset fits the DINA model, then F500/R1000 is ample to reach an acceptable Type I error rate and power. However, a bigger sample size is needed if the dataset fits the DINO model. Observations also underscored how thoughtfully considering differences in ability distribution and test length can critically impact the reliability of attribute mastery estimates. Notably, the DINA and DINO models displayed unique responses to uneven ability distribution, underscoring the significance of model-data fit evaluations in practical applications.

Addressing the issue of contaminated matching criteria, we introduced purification procedures to improve the reliability of traditional methods such as MH and LR. Notably, these methods exhibited enhanced efficacy under conditions of equal ability distribution when purified. However, challenges persisted under conditions of unequal ability distribution, underlining the need for rigorous selection of matching criteria.

The results from our real data example indicated that with the implementation of the purification procedure, the number of DIF items was decreased when using the MH method. The MH-P, LR-P, and model-based methods all detected three typical items as DIF items. However, the model-based method possesses a more significant advantage in isolating the DIF item that occurred in slipping, guessing, or both scenarios. This could guide practitioners in considering the causes of DIF.

Future studies in DIF research within cognitive diagnostic contexts hold promising avenues. First, exploring alternative purification strategies beyond those investigated in this study could refine our understanding of contaminated matching parameters. Thus, the issue of item purification within CDMs is a topic that warrants further attention. For instance, if there’s a low correlation between attributes, the total test score will no longer serve as a sufficient statistic. In such instances, using the total test score as a matching criterion can certainly lead to inaccurate results. As such, the purification procedure of the mastery profile score requires additional examination. Second, probing into DIF beyond the item level is critical, specifically in understanding its effect on higher-order cognitive abilities. Multidimensional CDMs could offer valuable insights here, presenting a more inclusive perspective on DIF sources. Furthermore, integrating external predictors into the proposed model could augment our grasp of potential DIF causes, offering a more nuanced analysis. Lastly, since the use of effect size is crucial in application, more simulations relating to differing effect size measures should be explored. A consistency analysis regarding the classification of effect size should also be conducted in future studies. This study employs MCMC for parameter estimation, prioritizing the more accurate results it provides despite its longer computational time. Future research could further explore the development of more efficient estimation methods.

In summary, this study provides an in-depth exploration of DIF detection methods within cognitive diagnostic frameworks. Our findings stress the significance of meticulous consideration of model-data fit, Q-matrix design, and contaminated matching criteria. Addressing these complexities and considering approaches such as the DFTD method proposed by Wang (2008), future applications of DIF analysis in cognitive diagnostic assessments can see substantial enhancement. This, in turn, fosters fair and unbiased evaluations in educational settings.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*(3), 253–260. https://doi.org/10.1177/014662168801200304

de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika*, *74*(4), 595–624. https://doi.org/10.1007/s11336-008-9063-2

DeCarlo, L. T. (2011). On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-Matrix. *Applied Psychological Measuremen*t, *35*(1), 8–26. https://doi.org/10.1177/0146621610377081

DeMars, C. E. (2010). Type I error inflation for detecting DIF in the presence of impact. *Educational and Psychological Measurement*, *70*(6), 961–972. https://doi.org/10.1177/0013164410366691

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, *14*(1), 76–94. https://doi.org/10.21031/epod.1218144

Finch, W. H., & French, B. F. (2007). Detecting of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement*, *67*(4), 565–582. https://doi.org/10.1177/0013164406296975

Foy, P., & Olson, J. F. (Eds.). (2008). *TIMSS 2007 user guide for the international database*. TIMSS & PIRLS International Study Center, Boston College.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement*, *67*(3), 373–393. https://doi.org/10.1177/0013164406294781

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4) 457–472. https://doi.org/10.1214/ss/1177011136

Gierl, M. J., Zheng, Y., & Cui, Y. (2008). Using the attribute hierarchy method to identify and interpret cognitive skills that produce group differences. *Journal of Educational Measurement*, *45*(1), 65–89. https://doi.org/10.1111/j.1745-3984.2007.00052.x

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement*, *26*(4), 333–352. http://dx.doi.org/10.1111/j.1745-3984.1989.tb00336.x

Harris, A. M., & Carlton, S. T. (1993). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education*, *6*(2), 137–151. https://doi.org/10.1207/s15324818ame0602_3

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191–210. https://doi.org/10.1007/s11336-008-9089-5

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.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Lawrence Erlbaum.

Hou, L., de la Torre, J., Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnosis modeling: Application of the Wald test to investigate DIF for DINA model. *Journal of Educational Measurement*, *51*(1), 98–125. https://doi.org/10.1111/jedm.12036

Huang, H.-Y. (2014). Effects of the common scale setting in the assessment of differential item functioning. *Psychological Reports*, *114*, 104–125. https://doi.org/10.2466/03.PR0.114k11w0

Huang, H.-Y. (2018). Effects of item calibration errors on computerized adaptive testing under cognitive diagnosis models. *Journal of Classification*, *35*, 437–465. https://doi.org/10.1007/s00357-018-9265-y

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

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Lane, S., Liu, M., Ankenmann, R. D., & Stone, C. A. (1996). Generalizability and validity of a mathematics performance assessment. *Journal of Educational Measurement*, *33*(1), 71–92. https://doi.org/10.1111/j.1745-3984.1996.tb00480.x

Lee, Y.-S., Park, Y. S., & Taylan, D. (2011). A cognitive diagnostic modeling of attribute mastery in Massachusetts, Minnesota, and the U.S. National sample using the TIMSS 2007. *International Journal of Testing*, *11*(2), 144–177. https://doi.org/10.1080/15305058.2010.534571

Li, F. (2008). *A modified higher-order DINA Model for detecting differential item functioning and differential attribute functioning* [Unpublished doctoral dissertation]. University of Georgia.

Li, H.-H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika*, *61*(4), 647–677. https://doi.org/10.1007/BF02294041

Li, X., & Wang, W.-C. (2015). Assessment of differential item functioning under cognitive diagnosis models: The DINA model example. *Journal of Educational Measurement*, *52*(1), 28–54. https://doi.org/10.1111/jedm.12061

Lord, F. M. (1980). Applications of item response theory to practical testing problems (1st ed.). Routledge. https://doi.

org/10.4324/9780203056615

Magis, D., Béland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*(3), 847–862. https://doi.org/10.3758/BRM.42.3.847

Magis, D., Raîche, G., Béland, S., & Gérard, P. (2011). A generalized logistic regression procedure to detect differential item functioning among multiple groups. *International Journal of Testing*, *11*(4), 365–386. https://doi.org/10.1080/15305058.2011.602810

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel- Haenszel procedure. *Journal of the American Statistical Association*, *58*(303), 690–700. https://doi.org/10.1080/01621459.1963.10500879

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*(2), 131–144. https://doi.org/10.1111/j.1745-3984.1995.tb00459.x

Miller, T. R., & Spray, J. (1993). Logistic discrimination function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, *30*(2), 107–122. https://doi.org/10.1111/j.1745-3984.1993.tb01069.x

Mullis, I., Martin, M., Ruddock, G., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2005). *TIMSS 2007 assessment frameworks*. International Association for the Evaluation of Educational Achievement.

Park, Y. S., Xing, K., & Lee, Y.-S. (2018). Explanatory cognitive diagnostic models: Incorporating latent and observed predictors. *Applied Psychological Measurement*, *42*(5), 376–392. https://doi.org/10.1177/0146621617738012

Paulsen, J., Svetina, D., Feng, Y., & Valdivia, M. (2020). Examining the impact of differential item functioning on classification accuracy in cognitive diagnostic models. *Applied Psychological Measurement*, *44*(4), 267–281. https://doi.org/10.1177/0146621619858675

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167). North-Holland Publishing.

Qiu, X.-L., Li, X., & Wang, W.-C. (2019). Differential item functioning in diagnostic classification models. In M. von Davier & Y.-S. Lee (Eds.), *Handbook of diagnostic classification models, methodology of educational measurement and assessment* (pp. 379–393). Springer. https://doi.org/10.1007/978-3-030-05584-4_18

R Development Core Team. (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Roussos, L. A., & Stout, W. F. (1996). Simulation studies of the effects of small sample and studied item parameters on SIBTEST and Mantel-Haenzel type I error performance. *Journal of Educational Measurement*, *33*(2), 215–230. https://doi.org/10.1111/j.1745-3984.1996.tb00490.x

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement*, *44*(4), 293–311. https://doi.org/10.1111/j.1745-3984.2007.00040.x

Rupp, A. A., & Templin, J. L. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, *6*(4), 219–262. https://doi.org/10.1080/15366360802490866

Rupp, A. A., Templin, J. L., & Henson, R. A. (2010). *Diagnostic assessment: Theory, methods, and applications*. Guilford Press.

Ryan, K. E., & Chiu S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education*, *14*(1), 73–90. https://doi.org/10.1207/S15324818AME1401_06

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. http://www.jstor.org/stable/2958889

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*(2), 159–194. https://doi.org/10.1007/BF02294572

Shih, C. L., & Wang, W. C. (2009). Differential item functioning detection using the multiple indicators, multiple causes method with a pure short anchor. *Applied Psychological Measurement*, *33*(3), 184–199. https://doi.org/10.1177/0146621608321758

Spiegelhalter, D. J., Thomas, A., & Best, N. (2003). *WinBUGS* (version 1.4) [Computer program]. MRC Biostatistics Unit. https://www.mrc-bsu.cam.ac.uk/software/bugs-project

Svetina, D., Feng, Y., Paulsen, J., Valdivia, M., Valdivia, A., & Dai, S. (2018). Examining DIF in the context of CDMs when the Q-matrix is misspecified. *Frontiers in Psychology*, *9*, Article 696. https://doi.org/10.3389/fpsyg.2018.00696

Swaminathan, H., & Rogers, J. H. (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

Templin, J. L., Henson, R. A., Templin, S. E., & Roussos, L. (2008). Robustness of hierarchical modeling of skill association in cognitive diagnosis models. *Applied Psychological Measurement*, *32*(7), 559–574. https://doi.org/10.1177/0146621607300286

Templin, J., & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

Tennant, A., & Pallant, J. F. (2007). DIF matters: A practical approach to test if differential item functioning makes a difference. *Rasch Measurement Transactions*, *20*, 1082–1084.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Erlbaum.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (ETS Research Rep. No. RR-05-16). Educational Testing Service.

Wang, W.-C. (2004). Effects of anchor item methods on detecting differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*(3), 221–261. https://doi.org/10.3200/JEXE.72.3.221-261

Wang, W.-C. (2008). Assessment of differential item functioning. *Journal of Applied Measurement*, *9*(4), 387–408.

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

Wang, W.-C., & Su, Y.-H. (2004b). Factors influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement*, *28*(6), 450–480. https://doi.org/10.1177/0146621604269792

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, *27*(6), 479–498. https://doi.org/10.1177/0146621603259902

Wu, M. L., Adams R. J., Wilson M. R., Haldane S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Australia Council for Educational Research Press.

Zhang, W. (2006). *Detecting differential item functioning using the DINA model* [Unpublished doctoral dissertation]. University of North Carolina.