JOURNAL OF APPLIED MEASUREMENT, *25*(1/2), 26-47

Exploratory Differential Item Functioning Assessment With Multiple Grouping Variables: The Application of the DIF-Free-Then-DIF Strategy

Jyun-Hong Chen
National Cheng Kung University

Chi-Chen Chen
National Academy for Educational Research, Taiwan

Hsiu-Yi Chao
Soochow University

To ensure test fairness and validity, it is crucial for test practitioners to assess differential item functioning (DIF) simultaneously for all grouping variables to avoid omitted variable bias (OVB; Chao et al., 2018). In testing practice, however, we often face challenges due to insufficient information, such as the absence of DIF-free anchor items, while conducting DIF assessment. This scenario, referred to as exploratory DIF assessment involving multiple grouping variables, has received limited attention, highlighting the importance of accurately identifying DIF-free items as anchors for all grouping variables. To address this issue, this study proposed the parallel DIF-free-then-DIF (p-DFTD) strategy, which selects DIF-free items simultaneously for each grouping variable and utilizes them as anchors in the constant item method for DIF assessment. A comprehensive simulation study was conducted to evaluate the performance of the p-DFTD strategy. The findings revealed that the conventional approach of assessing DIF with one grouping variable at a time was vulnerable to OVB, leading to an inflation of Type I error rates. In contrast, the p-DFTD strategy successfully identified DIF-free anchor items and effectively controlled Type I errors while maintaining satisfactory statistical power in most conditions. The empirical analysis further supported these findings, showing that the p-DFTD strategy provided more accurate and consistent DIF detection compared to methods that do not account for all grouping variables simultaneously. In conclusion, the p-DFTD strategy, which demonstrated a robust performance in this study, holds promise as a reliable approach for test developers to conduct exploratory DIF assessments involving multiple grouping variables, thereby ensuring fairness and validity in testing practices.

Keywords: Differential item functioning, DIF-free-then-DIF strategy, scale purification, omitted variable bias

Requests for reprints should be sent to Hsiu-Yi Chao, No.70, Linhsi Road, Shihlin District, Taipei City, 111002, Taiwan; email: psyhyc@scu.edu.tw

Introduction

Differential item functioning (DIF) assessment has been widely conducted for decades to ensure test fairness in routine item analysis. To assess DIF, test practitioners must establish a common metric for scale linking, in which a set of DIF-free anchors (i.e., items that do not contain DIF) is usually required to ensure its validity. Based on the built common metric, DIF effects (e.g., the difference in item difficulties between groups) can be estimated and tested to determine whether the studied item exhibits DIF. Notably, a reliable common metric is the necessary condition for a valid DIF assessment (Chen et al., 2014; Wang, 2008; Wang et al., 2012).

In practice, test practitioners often face the challenge of insufficient information while conducting DIF assessment (e.g., Kopf et al., 2015a). Specifically, we usually do not know which items are DIF-free, so the validity of a common metric can be reduced if DIF items are mistakenly recruited as anchors, resulting in the inflation of Type I error rates (e.g., Wang, 2004). Likewise, we also have little or no information about which grouping variables cause DIF. Conventional DIF methods (e.g., the Mantel–Haenszel [MH] method; Holland & Thayer, 1988; Mantel & Haenszel, 1959) that assess DIF regarding a single grouping variable at a time can thus be risky. Specifically, given that there is another grouping variable (e.g., college major) that not only induces DIF but also correlates to the assessed grouping variable (e.g., gender), omitted variable bias (OVB) may occur while assessing DIF if we leave out the relevant grouping variable, potentially causing the estimation of the DIF effect to be severely biased (Chao et al., 2018; Chao et al., 2024).

In building a common metric, researchers have proposed several strategies to identify DIF-free anchors. Among those strategies, some are assumption-free and are performed by multiple-step procedures, of which the scale purification (SP) and DIF-free-then-DIF (DFTD) strategies are two of the best known (Lord, 1980; Wang, 2008; Wang et al., 2012). Regarding how to avoid OVB, several methods that have been proposed to assess DIF simultaneously for multiple grouping variables are helpful. For example, the factorial ANOVA approach decomposes the DIF effects of multiple grouping variables into the main effects and their interaction effects (Jin et al., 2015; Wang, 2000). In addition, Chun et al. (2016) proposed the use of the multiple indicators multiple causes (MIMIC) model (Hauser & Goldberger, 1971; Oort, 1998) for simultaneously assessing DIF across multiple grouping variables. However, few studies have investigated DIF assessment considering both exploratory conditions and multiple grouping variables. Hence, test practitioners may encounter unprecedented problems while assessing DIF under this situation.

In testing practice, there can be many grouping variables related to a test. For example, the Student Questionnaire from the Trends in International Mathematics and Science Study (TIMSS) has more than 10 questions regarding examinees' background information, yielding more than 10 grouping variables. Under these circumstances, assessing DIF in exploratory conditions can become unapproachable. Taking the SP procedure as an example, DIF assessment in each iteration requires assessing DIF for all items against all grouping variables to update the anchor sets for scale linking in the next iteration. It can take as many times as the test length (TL) multiplied by the number of grouping variables (G) to assess DIF (or conduct hypothesis testing). To further consider the number of iterations required to reach convergence, the number of times in assessing DIF grows accordingly (e.g., $TL \times G \times$ number of iterations). Consequently, DIF assessment can be too complex and time-consuming to be implemented by test practitioners.

Therefore, this study extends the DFTD strategy to assess DIF for multiple grouping variables under exploratory conditions using a parallel approach (denoted as the p-DFTD strategy). Specifically, the p-DFTD strategy

first applies the iterative constant item (ICI) method (Wang, 2004) to all grouping variables concurrently, followed by using the ICI outcomes to parallelly select the required number of anchors for each grouping variable. Afterward, the selected anchors deemed DIF-free are used to build common metrics for the simultaneous DIF assessment of all multiple grouping variables. As a result, DIF assessment using the p-DFTD strategy requires only $TL + 1$ times of assessing DIF and is immune to OVB, as it simultaneously considers all relevant grouping variables, provided that all have been collected. A simulation study was conducted to investigate the performance of the p-DFTD strategy in DIF assessment.

The remainder of this article is as follows: We first introduce the common metric methods for DIF assessment with a single grouping variable. Next, the p-DFTD strategy for DIF assessment with multiple grouping variables is illustrated. Subsequently, the simulation study and empirical analysis for evaluating the performance of the p-DFTD strategy is presented. Finally, the results and conclusions are reported.

Common Metric Methods for Single Grouping Variable

Establishing a common metric is crucial in DIF assessment, where the common metric is used to characterize individuals and items along a latent continuum for the measured construct (e.g., math ability) shared by the studied groups (e.g., focal and reference groups). Based on the common metric, test practitioners can link the scales of focal and reference groups, testing DIF effects (e.g., the difficulty difference between the studied groups) to determine whether a particular item exhibits DIF. Basically, there are three ways to build a common metric: the equal-mean-difficulty (EMD), all-other-item (AOI), and constant item (CI) methods (Wang, 2004).

Basic Strategies for Building a Common Metric

DIF assessment can be considered a multiple-group analysis, with a focus on whether there is a between-group difference in the studied item's parameters (e.g., item difficulty). When conducting a multiple-group analysis, it is essential to link the scales between the studied groups in advance. Conventionally, in DIF assessment, scale linking is performed by setting the equality between the studied groups' average item difficulties, which is known as the EMD method. Alternatively, researchers may select a subset of items, excluding the studied item, to establish a common metric, which is called the CI method. If the subset includes all items except the studied one, it is referred to as the AOI method.

While these methods have been widely applied in testing practice and implemented in software for test analysis (e.g., *ConQuest*; Adams et al., 2017), they are feasible only under certain assumptions. For example, the EMD method produces unbiased DIF estimates only when the average difficulty of the test is the same between the reference and focal groups, while the CI method requires that all selected anchor items are DIF-free. These assumptions are not necessarily held in test practice, especially in exploratory DIF studies without prior information. To address this issue, researchers have proposed multiple-step procedures to improve exploratory DIF assessment, in which the SP and DFTD strategies are two of the most well-known (Lord, 1980; Wang, 2008; Wang et al., 2012).

The SP Procedure

The SP procedure is applied to DIF assessment to keep the common metric as clean as possible by iteratively removing DIF items from the anchor set and rebuilding the common metric with all items considered DIF-free in previous iterations. By implementing the SP procedure, the common metric built for the studied groups becomes more reliable. DIF assessment methods (e.g., the MH method) incorporated with the SP procedure were found to better control the Type I error rates of a DIF assessment than the non-iterative counterparts (Clauser et al., 1993; Donoghue et al., 1993; Wang & Su, 2004). Due to its iterative characteristics, the SP procedure is also considered a kind of stepwise procedure in DIF assessment scenarios that aims to select the most likely DIF-free items for establishing common metrics (Tay et al., 2013).

The following steps provide a guideline for implementing the SP procedure (Wang et al., 2009): (a) initially assessing DIF for all items, (b) removing items assessed as exhibiting DIF from the anchor set, (c) reassessing DIF for all items, and (d) repeating steps (b) and (c) until the same set of items is assessed as exhibiting DIF in two consecutive iterations.

Although the SP procedure outperforms the above-mentioned basic strategies (e.g., the EMD method), the Type I error rates were inflated when multiple DIF items were present in the test (e.g., over 20%; French & Maller, 2007; Khalid & Glas, 2014; Wang & Shih, 2010; Wang & Su, 2004). To address this issue, the DFTD strategy (Wang, 2008; Wang et al., 2012) was introduced.

The DFTD Strategy

Compared to the EMD and AOI methods, the assumptions required for the CI method of finding several DIF-free items (e.g., four items) are relatively easy to satisfy (Wang, 2004; Wang & Yeh, 2003). In view of this, the DFTD strategy is introduced as a two-step DIF assessment method that first selects the likeliest DIF-free items as anchors and then assesses DIF for the other items in the test (Wang, 2008). In other words, the DFTD strategy divides the DIF assessment procedure into two stages: locating a set of DIF-free items (first stage) and carrying out a DIF assessment with the CI method using the pre-located DIF-free items as anchors (second stage). To locate DIF-free items, the ICI method (Wang, 2004) is proposed, consisting of the following steps:

1. Set Item 1 as the anchor, assess DIF for all other items in the test using the CI method, and obtain the estimate of the DIF effect for each item except the anchor item.
2. Set the next item as the anchor and assess all other items in the test as in Step 1.
3. Repeat Step 2 until the last item is set as the anchor.
4. Compute the mean absolute values of the DIF estimates for each item across all iterations. The item with the lowest mean absolute value is considered most likely to be DIF-free.
5. Select the required number of items that are most likely to be DIF-free as the anchors.

Within each iteration of the ICI method, one item is constrained as having no DIF effect (i.e., DIF-free), so the DIF model can be identified accordingly. Given that most items in the test are DIF-free, most of the iterations will generate unbiased DIF estimates. Based on the results produced by these iterations, DIF-free items can be selected with high accuracy (e.g., Chen et al., 2014). Then, DIF assessment can be carried out via the CI method using the pre-located DIF-free items as anchors. The DFTD strategy has been applied to many DIF assessment methods, including the MIMIC method, yielding satisfying results in assessing DIF (e.g., Chen et al., 2014; Shih et al., 2014; Wang & Shih, 2010).

The DFTD strategy that explores a test first and then utilizes the findings to establish a common metric for DIF assessment does not require any arbitrary constraints or prior information and is appropriate for exploratory DIF assessment. Furthermore, the DFTD strategy also offers a flexible framework to integrate different procedures to improve DIF assessment performances. For example, Chen and Hwu (2018) employed the SP procedure, rather than the CI method, to assess DIF in the second stage of the DFTD strategy, which is called the dual-scale purification (DSP) method. However, currently, the DFTD strategy is applied only to DIF assessment with a single grouping variable. Considering the need for conducting exploratory DIF assessment with

multiple grouping variables in testing practices, the DFTD method is extended in the next section.

The p-DFTD Strategy for Multiple Grouping Variables

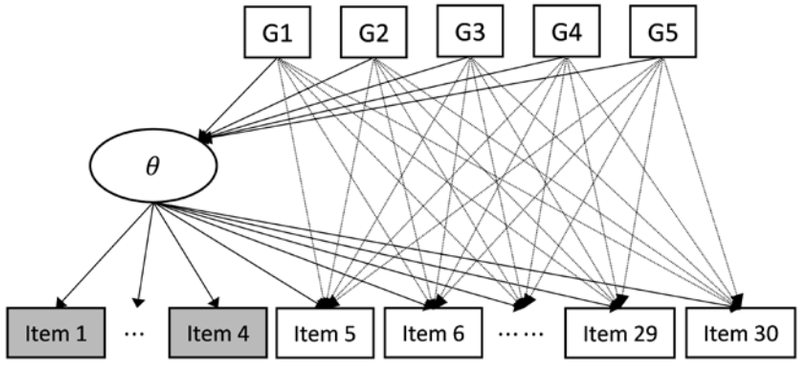
To simultaneously assess DIF for multiple grouping variables, Chao et al. (2018) demonstrated that the CI method with pure anchors can yield satisfactory outcomes. This approach, also known as the controlled method, is visually illustrated in Figure 1, where Item 1 to Item 4 are DIF-free anchors by design, and G1 to G5 are studied grouping variables. Given the efficacy of the controlled method in assessing DIF across multiple grouping variables, it appears feasible to directly extend the DFTD strategy to exploratory DIF assessment involving multiple grouping variables. This extended strategy, known as p-DFTD, involves initially identifying the most likely DIF-free items for each grouping variable and subsequently using those identified DIF-free items as anchors in the controlled method. Crucially, the success of the p-DFTD strategy lies in effectively determining the DIF-free anchors for all grouping variables.

Before delving into the extension, it is beneficial to differentiate between the bias on DIF estimates caused by anchor contamination and OVB. Anchor contamination refers to the incorrect selection of DIF items as anchors, which invalidates the common metric and leads to inflated Type I error rates (e.g., Huelmann et al., 2020; Wang, 2004). To prevent anchor contamination, researchers can conduct auxiliary DIF tests (e.g., by the ICI method) to locate DIF-free anchor items (e.g., Kopf et al., 2015a). On the other hand, OVB occurs when we omit relevant variables while assessing DIF, where the relevant variables are grouping variables that simultaneously induce DIF in test items and are correlated with the studied grouping variables. To prevent OVB, all relevant variables need to be included in the DIF model to control their effects.

To locate DIF-free items for multiple grouping variables, conducting the auxiliary DIF test sequentially, one grouping variable at a time, is inevitably susceptible to OVB caused by relevant variables. To address this issue, the ICI method should be extended to select anchor items simultaneously for all grouping variables. In each iteration, specifically, one anchor item per studied grouping variable is chosen to prevent OVB while estimating the DIF effects of the remaining items. This procedure is iteratively repeated, considering all possible combinations of anchor items for each grouping variable. When there are relatively few DIF items within each grouping variable, the majority of anchor combinations used in the iterations are expected to consist of DIF-free items, generating unbiased DIF estimates.

Figure 1

The Controlled Method



The impact caused by anchor contamination can thus be mitigated, leading to accurate DIF assessments. Based on the results, we should be able to accurately identify anchors for each grouping variable.

However, the number of possible anchor combinations grows with the number of grouping variables, potentially leading to an unmanageable increase in the number of required iterations. Specifically, for one grouping variable, the ICI method requires *TL* iterations to determine the anchor set. In a full crossover iterative procedure with, for example, 10 grouping variables, the number of anchor combinations (also corresponding to the number of iterations) can grow exponentially (e.g., TL^{10}). Under these circumstances, the ICI method becomes extremely complex and computationally time-consuming.

Considering that including all grouping variables in the DIF model is sufficient to prevent OVB, it is not necessary to iterate through all possible anchor combinations across all grouping variables in the auxiliary DIF test. We need only to obtain the iteration results produced by all possible anchor conditions within each grouping variable. This means that, in the simplest case, we can use the same item (e.g., Item 1) as the anchor for all grouping variables in each iteration. By doing so, only *TL* iterations are needed to complete the auxiliary DIF test for all grouping variables. Specifically, the following steps provide a precise summary of the extended ICI method:

1. Set Item 1 as the anchor and assess DIF for all other items in the test using the controlled method, obtaining DIF estimates of each item (except the anchor item) across all grouping variables.
2. Set the next item as the anchor and assess all other items in the test as in Step 1.
3. Repeat Step 2 until the last item is set as the anchor.
4. For the first grouping variable, calculate the mean absolute value of the DIF estimates for each item over all iterations. The item with the lowest mean absolute value is considered most likely to be DIF-free.
5. Select the required number of items that are most likely to be DIF-free as the anchors.
6. Repeat Steps 4 and 5 for each subsequent grouping variable, determining the anchors until the anchor selection process is completed for all grouping variables.

In the aforementioned steps, it is possible that the anchor combinations may not be entirely correct for all grouping variables in any single iteration; however, this only leads to anchor contamination rather than OVB in the DIF estimates. As long as the majority of test items within each grouping variable are DIF-free, most iterations will still yield accurate DIF assessments within each individual grouping variable. By integrating all the results, for each grouping variable, its corresponding most likely DIF-free items can be selected as anchors simultaneously, and DIF assessment can be conducted using the selected anchors with the controlled method. The DFTD strategy, incorporating this extended ICI method (i.e., the p-DFTD strategy), is immune to OVB and is expected to efficiently identify DIF-free anchors, yielding satisfactory outcomes for exploratory DIF assessment involving multiple grouping variables, given that all relevant grouping variables have been collected.

Simulation Study

The purpose of the simulation study was to investigate the performance of the p-DFTD strategy in exploratory DIF assessment with multiple grouping variables. A thorough study with various manipulations was conducted to provide information for test practitioners in applying the p-DFTD strategy to DIF assessment in testing practice.

Design

This study aimed to compare the performances of four methods in DIF

assessment: the single-group method with pure anchors (referred to as SG.PA); the multiple-group method with the p-DFTD strategy (referred to as MG.DFTD); the multiple-group method with the SP procedure (referred to as MG.SP); and the multiple-group method with pure anchors (referred to as MG.PA). Furthermore, the true model (referred to as TM), which utilizes the data-generating model for DIF assessment, was also employed to serve as a baseline for comparison. The purpose of including these methods/models (e.g., SG.PA) was to investigate their different approaches in addressing the challenges associated with exploratory DIF assessment involving multiple grouping variables and their corresponding DIF assessment outcomes. Specifically, based on Table 1, we can categorize the five methods/models into three groups based on two dimensions: single/multiple grouping variables and known/unknown anchors.

Table 1

Methods Compared in the Simulation Study

		Anchors	
		Known	**Unknown**
No. of grouping variables	Single	Category 1 SG.PA	
	Multiple	Category 2 MG.PA TM	Category 3 MG.DFTD MG.SP

In the first category, which involves known anchors for a single grouping variable, SG.PA represents the optimal approach for conducting DIF assessment in the single-grouping-variable context, allowing us to understand the extent of the bias caused when the correlations between multiple grouping variables are ignored (e.g., OVB). Alternatively, MG.PA and TM, belonging to the second category, provide insights into the optimal performance of DIF assessment in the context of multiple grouping variables. Since pure anchors are not available in practical scenarios, the third category, represented by MG.DFTD and MG.SP, reflects the performance of DIF assessment in real-world testing practice. The differences in DIF assessment performances between methods/models in the first category and the second category demonstrate the effectiveness of the methods/models considering the inclusion of multiple grouping variables. By contrast, the differences between the second category and the third category highlight the potential for further improvement of the methods proposed in this study. That is, the methods/models in the second category (i.e., MG.PA and TM) are used as theoretical upper bounds for DIF assessment performance to evaluate how closely the methods proposed in this study (i.e., MG.DFTD and MG.SP in the third category) can approach the optimal performance.

Three independent variables were manipulated in this study: (1) the correlations among grouping variables: 0, .1, .2, .3, .4, .5, and .6; (2) the number of grouping variables: 5 and 10; and (3) the magnitude of DIF effect: 0.4 and 0.6. The group correlation is one of the factors that determine the magnitude of the OVB effect, as highlighted in previous studies (e.g., Chao et al., 2018). By manipulating the correlation, the study aimed to investigate the impact of ignoring the context of multiple grouping variables on DIF assessment. Regarding the number of grouping variables, selecting 5 or 10 grouping variables reflects the typical number of variables that may be collected in testing practice (Chao et al., 2018). While the DIF effects of 0.6 logit are considered moderate in many DIF studies (e.g., Chen et al., 2014; Finch, 2016; Kopf et al., 2015a, 2015b), the DIF effect of 0.4 logit was used to represent a more pervasive small DIF effect.

Regarding DIF-related settings, we divided the grouping variables into five groups, with DIF percentages set at 0%, 10%, 20%, 30%, and 40%, respectively. The manipulation falls within the common range of DIF percentages (typically less than 50%), as observed in previous studies (e.g., Chen et al., 2014; Kopf et al., 2015a, 2015b), allowing us to examine the impact of DIF percentages on DIF assessment. It is worth noting that all DIF effects caused by the grouping variables favored the reference group, representing the constant condition (Wang et al., 2012). The grouping variables randomly caused the pre-specified DIF percentages of items in the test to exhibit DIF. To preliminarily explore the performance of the p-DFTD strategy, this study only considered uniform DIF, meaning the DIF effect appeared in the difficulty of the items. However, the p-DFTD strategy is ready to be generalized to the nonuniform DIF context when additionally considering discrimination parameters in the DIF assessment model.

In terms of the other simulation settings, this study employed the Rasch model (Rasch, 1960) to generate examinees' responses. The five methods (e.g., MG.SP) to be compared all incorporated the MIMIC model for DIF assessment. The *TL* was set to 30 items to simulate a medium-length test commonly used in practice. The number of anchor items was set to four, where a four-item anchor set is sufficient and is generally recommended for DIF assessment (Wang & Yeh, 2003). It is worth noting that in the context of multiple grouping variables, the items most likely to be DIF-free under each grouping variable might differ, and thus, the four anchor items selected for each grouping variable may also vary. The item parameters (i.e., difficulty) and examinees' abilities were randomly sampled from the standard normal distribution. For each grouping variable, the sample size was 500 examinees each for the reference group and focal group. That is, the total sample size was 1,000, a common size for test construction. Specifically, we randomly drew 1,000 samples from a 5- or 10-dimensional binomial distribution, with each dimension having a distribution of $B(1, 0.5)$.

For DIF item generation, a small example with a 10-item test and five grouping variables is illustrated. According to Table 2, grouping variables G1-G5 generated 0%, 10%, 20%, 30%, and 40% DIF items for the 10 test items, respectively. For example, Item 3 simultaneously exhibits DIF for G2, G4, and G5, represented by $I_3 = (0, 1, 0, 1, 1)^T$. Given that the DIF amount is 0.4, if Examinee A's group membership is in the reference group for G1-G3 and in the focal group for G4-G5, represented by the $E_A = (0, 0, 0, 1, 1)^T$, then the difficulty of this item for Examinee A would be $I_3^T \times E_A \times 0.4 + 0.6$, which equals 1.4. Likewise, if Examinee B's group membership is $E_B = (1, 1, 0, 0, 0)^T$, then the difficulty of this item for Examinee B would be calculated as $I_3^T \times E_B \times 0.4 + 0.6$, which equals 1.0.

Table 2

A Small Example Illustrating DIF Item Generation

Item	Difficulty	G1	G2	G3	G4	G5
1	1.3	0	0	0	1	0
2	0.4	0	0	0	0	0
3	0.6	0	1	0	1	1
4	0.1	0	0	1	0	1
5	−2.2	0	0	0	0	0
6	2.5	0	0	0	0	0
7	1.5	0	0	1	0	1
8	2.1	0	0	0	1	0
9	−0.7	0	0	0	0	0
10	−1.3	0	0	0	0	1

The simulation data were generated using the R (R Core Team, 2022) program, written by the authors. The parameter estimation of MIMIC model for DIF assessment was conducted using R with the *lavaan* package (Rosseel, 2012). There were 200 replications for each condition.

The Implementations

For the MG.SP method, which applies the SP procedure to DIF assessment involving multiple grouping variables, the parallel approach employed by the p-DFTD strategy was used. The approach is called the parallel SP (p-SP) procedure here, and the following steps provide a detailed explanation:

1. Set Item 1 as the studied item and all other items as anchors. Conduct a DIF assessment using the controlled method.
2. Set the next item as the studied item and all other items as anchors. Conduct the DIF assessment as in Step 1.

3. Repeat Step 2 until the last item is designated as the studied item.
4. Remove items that exhibit DIF from the anchor set for the next iteration.
5. Repeat Steps 1 to 4 until the DIF assessment outcomes in two consecutive iterations are consistent or reach the predetermined maximum number of iterations (5 in this study).

Additionally, the SG.PA method sequentially assesses each grouping variable; for each grouping variable's DIF assessment, four DIF-free items are randomly selected as anchors for applying the CI method, as mentioned in Shih and Wang (2009). In contrast, the MG.PA method assesses all grouping variables simultaneously; its procedure can be referenced in the controlled method described by Chao et al. (2018). For the implementations of the MG.DFTD method, please refer to the section on "The p-DFTD Strategy for Multiple Grouping Variables."

Analysis

Two dependent variables were used, namely Type I error rates and the power rates of DIF assessment. Since power rates are meaningless when Type I error rates are out of control, power rates corresponding to conditions with inflated Type I error rates were excluded from further examination. Additionally, the accuracy of anchor selection in the p-DFTD strategy was also evaluated. For instance, if out of the four selected anchor items, three were correctly identified as DIF-free, the accuracy of anchor selection would be 0.75.

Results of the Simulation Study

Conditions With DIF Effect = 0.6

Accuracy of Anchor Item Selection

Table 3 shows the accuracy of DIF-free anchor item selection for the MG.DFTD method. In general, the accuracy decreased as the number of grouping variables and the correlation among grouping variables increased. When there were five grouping variables, the accuracy of selecting DIF-free items was consistently high (> .95) for grouping variables with a DIF percentage less than 40% (ranging from .954 to .999). By contrast, for a grouping variable with 40% DIF items, the accuracy can drop below .9 when the correlation exceeded .4. Specifically, when the correlation was .6, the accuracy was .839, indicating a less desirable performance.

When there were 10 grouping variables, the accuracy was slightly lower than that with five grouping variables, but its pattern remained similar. For grouping variables with a DIF percentage less than 30%, the accuracy was generally high, exceeding .94. By contrast, for grouping variables with 30% and 40% DIF items, the accuracy was lower and can fall below .9 when the correlation was high. In particular, for the grouping variable with 30% DIF items, the accuracy was below .9 in conditions with a correlation of .6. Similarly, for the grouping variable with 40% DIF, the accuracy was below .9 for correlations larger than .3. The lowest accuracy, of .782, was observed in conditions with a correlation of .6. In summary, the MG.DFTD method generally demonstrated high accuracy in selecting anchor items, except for grouping variables with a high DIF percentage (e.g., > 30%) and high correlation (e.g., > .3).

Table 3

Accuracy of Anchor Item Selection With the MG.DFTD Method

No. of grouping variables	DIF percentage	Correlation						
		0	0.1	0.2	0.3	0.4	0.5	0.6
5	10%	0.999	0.998	0.999	0.996	0.996	0.990	0.990
	20%	0.996	0.994	0.996	0.993	0.993	0.994	0.973
	30%	0.996	0.996	0.993	0.988	0.978	0.966	0.954
	40%	0.958	0.969	0.958	0.951	0.904	0.876	0.839
10	10%	0.998	0.999	0.996	0.997	0.988	0.988	0.973
	20%	0.991	0.994	0.989	0.987	0.981	0.969	0.943
	30%	0.988	0.983	0.979	0.973	0.955	0.931	0.893
	40%	0.949	0.943	0.944	0.906	0.888	0.848	0.782

Type I Error Rates of DIF Assessment

Figure 2 shows the Type I error rates of DIF assessment for grouping variables with different DIF percentages under each condition. When there were five grouping variables, the MG.DFTD and MG.PA methods effectively controlled the Type I error rates across all correlation levels among grouping variables (ranging from .025 to .076). However, the MG.SP method generated inflated Type I error rates for the grouping variable with 40% DIF items under conditions with correlations higher than .2 (ranging from .104 to .136). In contrast, the SG.PA method demonstrated out-of-control Type I error rates under conditions with correlations higher than .2 across grouping variables with all DIF percentages. The type I error rates increased as the correlation increased, ranging from .128 to .371.

When there were 10 grouping variables, the Type I error rates for the four methods were slightly higher compared to the case with five grouping variables, but the patterns remained similar. Specifically, both the MG.DFTD and MG.PA methods effectively controlled the Type I error rates (ranging from .027 to .088), but the MG.SP method generated inflated Type I error rates for grouping variables with 40% DIF items when the correlation was higher than .1 (ranging from .102 to .153). Alternatively, the SG.PA method showed a dramatic increase in Type I error rates as the correlation increased under conditions with correlations higher than .1 (ranging from .123 to .447).

Power Rates of DIF Assessment

Figure 3 shows the power rates of DIF assessment for grouping variables with different DIF percentages under each condition. Generally, the power decreased as the correlation among grouping variables and the number of grouping variables increased. Higher DIF percentages in grouping variables resulted in lower power in detecting DIF items. The TM generated the highest power rates of DIF assessment across all kinds of grouping variables under all conditions (ranging from .782 to .963). When there were five grouping variables, for grouping variables with DIF percentages less than 40%, the differences among the MG.DFTD, MG.SP, and MG.PA methods were not obvious. The power rates of the MG.DFTD method (ranging from .601 to .930) were slightly lower than those of the other two methods (ranging from .649 to .960). For the grouping variable with 40% DIF items, the power rates of the MG.DFTD methods (ranging from .490 to .881) were lower than those of the MG.PA method (ranging from .691 to .927), and these differences increased as the correlation increased. The MG.SP method yielded comparable power to the MG.DFTD method; however, its Type I error rates are inflated for conditions with correlations higher than .2. Regarding the SG.PA method, its power rates are not further examined due to the fact that its Type I error rates were out of control across all kinds of grouping variables under most of the conditions.

For conditions with 10 grouping variables, the patterns of power rates were like those observed under conditions with five grouping variables. For grouping variables with DIF percentages less than 40%, the power rates of the MG.DFTD method (ranging from .491 to .891) were slightly lower than those of the MG.PA and MG.SP methods (ranging from .548 to .929). For the grouping variable with 40% DIF items, the power rates of the MG.DFTD methods (ranging from .396 to .854) were lower than those of the MG.PA method (ranging from .638 to .907), and the differences increased as

Figure 2

Type I Error Rates of DIF Assessment Under Each Condition

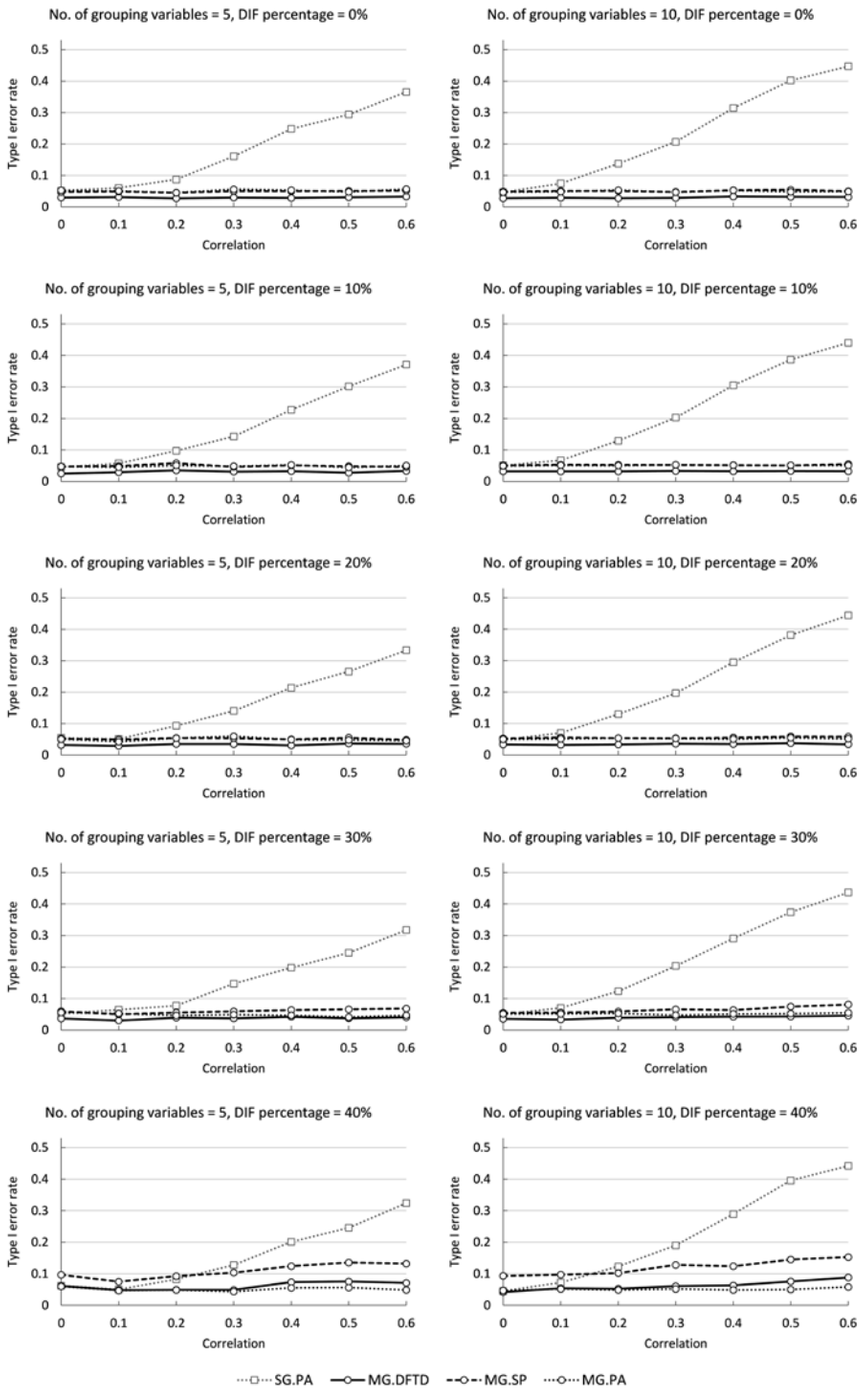
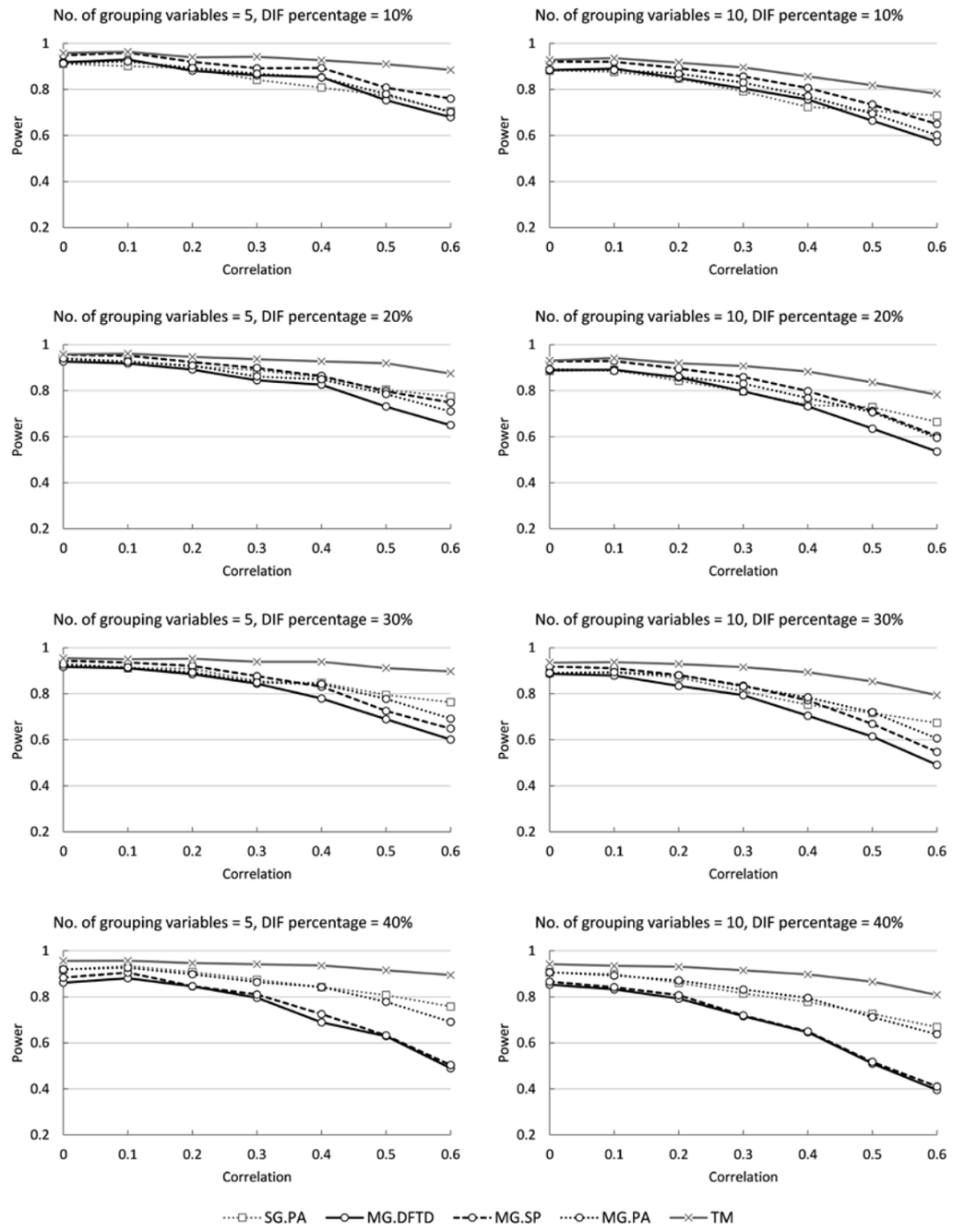


Figure 3

Power Rates of DIF Assessment Under Each Condition



the correlation increased. Since the power rates are meaningless when the Type I error rates are out of control, the power rates for the MG.SP method were not further examined here. Please refer to Tables A1 and A2 in the Supplemental Appendix for detailed results.

Conditions With DIF Effect = 0.4

For brevity, the results for the conditions with the DIF effect of 0.4 logit are not shown here. However, the results were similar to those observed under the conditions with the DIF effects of 0.6 logit, although the power was lower under the 0.4 logit condition. Please refer to Tables A3 to A5 in the Supplemental Appendix for detailed results.

Empirical Study: A TIMSS Example

This study examined the performance of the p-DFTD procedure using empirical data. The findings from the empirical analysis enabled us to generalize the research results to practical testing applications. Specifically, data from the mathematics assessment in the TIMSS 2019 cycle were utilized. The assessment included 211 items, comprising both binary and polytomous items. The research sample consisted of 4,915 8th-grade Taiwanese students. For test administration, TIMSS employed an equating design, dividing the 211 items into 14 booklets to reduce the burden on respondents. This study focused on two of these booklets (i.e., Booklets 3 and 4), analyzing a dataset of 721 participants who completed 29 common binary items across both booklets.

Regarding the grouping variables for DIF assessment, we referred to the Student Questionnaires and identified five variables with non-ignorable correlations (> 0.2) among them, as listed in Table 4. Among these, parental education level, expected education level, and the frequency of working alone were coded as binary variables to facilitate interpretation.

Given that the DIF-free anchor for each grouping variable is unknown in the empirical analysis, the PA approach (e.g., MG.PA) and TM cannot be applied in this context. Therefore, only the MG.DFTD and MG.SP methods were applied to the empirical data. Additionally, the SG.DFTD method, which uses the DFTD strategy to sequentially assess DIF for one grouping variable at a time, was employed. Although SG.DFTD was not included in the simulation study, its performance is expected to be close to, but slightly lower than, that of SG.PA—since SG.PA is designed with a DIF-free anchor, whereas the accuracy of anchor item selection in SG.DFTD is not guaranteed to

Table 4

Correlation Matrix and Sample Size for the Grouping Variables in the Empirical Study

Grouping variable	Correlation matrix					*N*	*N* (Response = 1)
	V1	V2	V3	V4	V5		
V1. Having a computer or tablet at home	-					719	649
V2. Having an internet connection at home	0.32	-				719	622
V3. Parental education level	−0.12	−0.11	-			637	359
V4. Expected education level	−0.10	−0.10	0.31	-		718	536
V5. Frequency of working alone	0.18	0.12	−0.23	−0.30	-	720	334

Note.
V1. Having a computer or tablet at home: 1 = Yes; 2 = No.
V2. Having an internet connection at home: 1 = Yes; 2 = No.
V3. Parental education level: 1 = Below Bachelor's level; 2 = Bachelor's level or above
V4. Expected education level: 1 = Bachelor's level or below; 2 = Above Bachelor's level
V5. Frequency of working alone: 1 = Every or almost every lesson; 2 = Less than every or almost every lesson

be 100%. The comparison between MG.DFTD and MG.SP highlights the differences between the DFTD and SP strategies, while the comparison between MG.DFTD and SG.DFTD reveals the impact of considering multiple grouping variables simultaneously in DIF assessment. The empirical analysis was conducted using R (R Core Team, 2022) with the *lavaan* package (Rosseel, 2012) for coefficient estimation in the MIMIC model.

Results of the Empirical Study

The analysis results indicated that the highest number of items flagged for DIF across all grouping variables and methods was no more than seven items (< 25%), suggesting that the common items in these booklets did not exhibit serious DIF issues across the grouping variables. For complete analysis results, including the coefficient estimates and *p*-values for DIF detection in common items across Booklets 3 and 4 across the three methods, please refer to Tables A6 and A7 in the Supplemental Appendix. Considering the moderate to low correlations among the variables, our simulation findings suggest that MG.DFTD provides well-controlled Type I error rates, while MG.SP also performs effectively under these conditions. Therefore, consolidating the results from both methods should yield robust DIF item detection outcomes. Specifically, MG.DFTD and MG.SP showed considerable consistency in identifying DIF items. Although some items were flagged only by MG.SP (e.g., Item 1 in V5), these items also showed marginal significance under MG.DFTD, aligning with our simulation study findings that, under conditions of lower DIF percentages and moderate to low correlations, MG.SP demonstrated slightly higher power than MG.DFTD. Consequently, we used the union of DIF items detected by MG.DFTD and MG.SP as the final detection results for comparison with SG.DFTD.

As shown in Table 5, there are some differences between the items detected by SG.DFTD and the union of items detected by MG.DFTD and MG.SP. Specifically, SG.DFTD identified additional Items 4, 5, and 6 under V3, and Item 28 under V4, but missed Item 3 under V2, Item 20 under V3, and Item 11 under V5. These differences are not surprising, as Type I and II errors are potential consequences of OVB. For instance, in the case of a potential Type I error, consider Item 5 under V3. The coefficient estimate under SG.DFTD was −0.11, while under MG.DFTD it was −0.09. Given that the correlation between V3 and V5 is −0.23, and the coefficient of V5 on Item 5 in MG.DFTD is 0.19, the OVB formula (Greene, 2003) suggests that not accounting for V5 results in a negative bias in the DIF estimate for V3. This leads to a more negative effect of −0.11 under SG.DFTD compared to −0.09 under MG.DFTD. Similarly, Item 3 under V2 illustrates the potential for Type II error. The coefficient estimate under SG.DFTD was −0.06, whereas under MG.DFTD it was −0.11. Considering that the correlation between V1 and V2 is 0.32, and the coefficient of V1 on Item 3 in MG.DFTD is 0.25, the OVB formula indicates that not accounting for V1 results in a positive bias in the DIF estimate for V2. This

Table 5

Item Numbers Identified With DIF Under Each Grouping Variable in the Empirical Study

Grouping variable	MG.DFTD ∪ MG.SP	SG.DFTD
V1. Having a computer or tablet at home	Items 1, 2, 3, 4, 5, 6, 7	Items 1, 2, 3, 4, 5, 6, 7
V2. Having an internet connection at home	Items 3, 22	Item 22
V3. Parental education level	Items 20, 26	Items 4, 5, 6, 26
V4. Expected education level	Items 12, 18, 27	Items 12, 18, 27, 28
V5. Frequency of working alone	Items 1, 2, 3, 4, 5, 6, 11,	Items 1, 2, 3, 4, 5, 6

leads to a more positive effect of −0.06 under SG.DFTD compared to −0.11 under MG.DFTD. The mechanism for the inconsistent results between SG.DFTD and MG.DFTD & MG.SP (e.g., DIF detection for Items 4 and 6 under V3) can be explained similarly.

In summary, the analysis of this empirical data illustrates that the SG.DFTD approach, which does not account for other grouping variables, may lead to Type I errors when there is a non-negligible correlation (e.g., > .2) between grouping variables. This can, in turn, affect subsequent revisions and interpretations related to the sources of DIF items.

Conclusion and Discussion

To ensure test fairness and validity, test practitioners have to assess DIF for all collected grouping variables under exploratory conditions. In testing practice, however, we often face the challenges of insufficient information while conducting exploratory DIF assessments. To address this issue, this study proposed the p-DFTD strategy, which utilizes the ICI method with a parallel approach to select anchor items for each grouping variable, followed by using the located anchor items with the controlled method (i.e., the CI method for multiple grouping variables) for DIF assessment. A thorough simulation study was conducted to evaluate the performance of the p-DFTD strategy in comparison to several important methods (e.g., the p-SP procedure). The results demonstrated that the conventional method (i.e., SG.PA) was prone to OVB, resulting in a severe inflation of Type I errors. In contrast, the p-DFTD strategy (i.e., MG.DFTD) successfully identified DIF-free anchor items and effectively controlled Type I errors while maintaining satisfactory statistical power in most scenarios.

The findings from the empirical analysis further support these results. Specifically, the comparison between the DIF items detected by MG.DFTD, MG.SP, and SG.DFTD highlighted the robustness of the p-DFTD strategy. While SG.DFTD, which does not account for all grouping variables simultaneously, showed some discrepancies in DIF detection—such as identifying additional DIF items and missing some items compared to MG.DFTD and MG.SP—these differences are consistent with the potential Type I and II errors predicted by the OVB framework. This empirical evidence underscores the importance of considering multiple grouping variables simultaneously in DIF assessment to minimize bias and improve the accuracy of the results.

However, it is worth noting that the p-DFTD strategy exhibited a noticeable decrease in statistical power for grouping variables with high DIF percentages (e.g., larger than 30%) under situations with high group correlations (e.g., higher than .5). This power decrease was even more pronounced when a large number of grouping variables were involved (e.g., 10). Regarding the p-SP procedure (i.e., MG.SP), it performed well for grouping variables with DIF percentages less than 40%. However, for grouping variables with a DIF percentage of 40%, the p-SP procedure exhibited an inflation of Type I error rates. Overall, the p-DFTD strategy demonstrated a robust performance in this study, providing test developers with a reliable approach to conducting exploratory DIF assessment involving multiple grouping variables in most scenarios.

Although the proposed method offers a robust solution to exploratory DIF assessment, there are several directions for future studies that are in line with the findings and address the limitations of the current study. First, in contexts with high group correlations, the power of the p-DFTD strategy may decrease to around .4. Further studies are required to incorporate sophisticated strategies into the p-DFTD strategy to improve the power of the p-DFTD strategy. For example, the DSP strategy, which increases the number of anchor items and enhances the accuracy of DIF-free item selection, has been shown to effectively improve the power of the DFTD strategy (Chen & Hwu, 2018). Additionally, regularization estimation techniques, such as the lasso estimator (Tibshirani, 1996), are suitable for exploratory DIF models with multiple grouping variables, especially when the DIF model involves a large and sparse design matrix for DIF assessment.

Second, several important variables that are often considered in DIF studies, such as group mean differences (i.e., impact), sample sizes, and proportions between reference and focal groups, were not manipulated in this study. Additionally, the robustness of the p-DFTD strategy in DIF assessment with polytomous models (e.g., rating scale model; Andrich, 1978) and multi-parameter models (e.g., two-parameter logistic model; Birnbaum, 1968) requires further investigation to assess its applicability across different testing contexts and models. Particularly, in addition to uniform DIF, the context with nonuniform DIF should be further explored to more comprehensively understand how different groups might be affected by certain test items.

Third, this study assumed that all relevant variables were known and measured. However, in practical situations, relevant variables may not be collected completely. Conducting DIF assessment using a model that includes incomplete relevant variables is of great importance for future research to address its impact on DIF assessment. Furthermore, future research could also explore the incorporation of more advanced statistical methods, such as multiple imputation, a widely applied technique in social science, into the p-DFTD strategy to address the issue of omitted relevant variables in exploratory DIF assessment. Furthermore, more empirical data analysis to validate the effectiveness of the p-DFTD strategy in practical applications is of great importance for future research.

Finally, considering that exploratory DIF assessment with multiple grouping variables involves multistage analyses (e.g., p-DFTD strategy), there is currently no corresponding procedure available in the existing software (e.g., R). To facilitate DIF assessment and alleviate the burden on test practitioners, future research could develop dedicated software or packages based on the findings of this study. By addressing these issues, we can continue to enhance the effectiveness and applicability of the p-DFTD strategy to exploratory DIF assessment involving multiple grouping variables, ensuring fair and valid assessments across diverse testing contexts.

References

Adams, R. J., Wu, M. L., & Wilson, M. R. (2017). ACER conquest. In *Handbook of item response theory* (pp. 495–506). Chapman and Hall/CRC.

Andrich, D. (1978). Rating formulation for ordered response category. *Psychometrika*, *43*(4), 561–573.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Chao, H. Y., Chen, C. C., Cheng, C. P., & Chen, J. H. (2018). Omitted variable bias in differential item functioning assessment. *Chinese Journal of Psychology*, *60*(4), 233–250.

Chao, H.-Y., Chen, J.-H. & Chen, C.-C. (2024). Mitigating omitted variable bias in exploratory differential item functioning assessment: A propensity score adjustment approach. *Journal of Educational and Behavioral Statistics*.

Chen, C.-T., & Hwu, B.-S. (2018). Improving the assessment of differential item functioning in large-scale programs with dual-scale purification of Rasch models: The PISA example. *Applied Psychological Measurement*, *42*(3), 206–220.

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

Chun, S., Stark, S., Kim, E. S., &

Chernyshenko, O. S. (2016). MIMIC methods for detecting DIF among multiple groups: Exploring a new sequential-free baseline procedure. *Applied Psychological Measurement*, *40*(7), 486–499.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, *6*(4), 269–279.

Donoghue, J. R., Holland, P. W., & Thayer, D. T. (1993). A Monte Carlo study of factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 137–166). Routledge.

Finch, W. H. (2016). Detection of differential item functioning for more than two groups: A Monte Carlo comparison of methods. *Applied Measurement in Education*, *29*(1), 30–45.

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.

Greene. W. H. (2003). *Econometric analysis*. Pearson Education.

Hauser, R. M., & Goldberger, A. S. (1971). The treatment of unobservable variables in path analysis. *Sociological Methodology*, *3*, 81–117.

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.

Huelmann, T., Debelak, R., & Strobl, C. (2020). A comparison of aggregation rules for selecting anchor items in multigroup DIF analysis. *Journal of Educational Measurement*, *57*(2), 185–215.

Jin, K.-Y., Chen, H.-F., & Wang, W.-C. (2015). Assessing differential item functioning in multiple grouping variables with factorial logistic regression. In R. Millsap, D. Bolt, L. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research: The 78th annual meeting of the Psychometric Society* (pp. 243–259). Springer.

Khalid, M. N., & Glas, C. A. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197.

Kopf, J., Zeileis, A., & Strobl, C. (2015a). Anchor selection strategies for DIF analysis: Review, assessment, and new approaches. *Educational and Psychological Measurement*, *75*(1), 22–56.

Kopf, J., Zeileis, A., & Strobl, C. (2015b). A framework for anchor methods and an iterative forward approach for DIF detection. *Applied Psychological Measurement*, *39*(2), 83–103.

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

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.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling*, *5*(2), 107–124

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. University of Chicago Press.

Rosseel, Y. (2012). lavaan: An R Package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. http://www.jstatsoft.org/v48/i02/

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

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.

Tay, L., Vermunt, J. K., & Wang, C. (2013). Assessing the item response theory with covariate (IRT-C) procedure for ascertaining differential item functioning. *International Journal of Testing*, *13*(3), 201–222.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Wang, W.-C. (2000). The simultaneous factorial analysis of differential item functioning. *Methods of Psychological Research Online*, *5*(1), 57–75.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, *72*(3), 221–261.

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

Wang, W.-C., & Shih, C.-L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, *34*(3), 166–180.

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

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

Wang, W.-C., & Su, Y.-H. (2004). 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.

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.

Appendix

Table A1

Type I Error Rates of DIF Assessment Under Each Condition (DIF Magnitude = 0.6)

		No. of grouping variables = 5							No. of grouping variables = 10						
DIF%	Correlation	0	0.1	0.2	0.3	0.4	0.5	0.6	0	0.1	0.2	0.3	0.4	0.5	0.6
0%	SG.PA	0.051	0.060	0.087	0.161	0.248	0.294	0.366	0.046	0.075	0.138	0.207	0.314	0.402	0.447
	MG.DFTD	0.030	0.031	0.027	0.030	0.029	0.031	0.033	0.028	0.029	0.027	0.029	0.033	0.032	0.032
	MG.SP	0.047	0.049	0.045	0.050	0.050	0.051	0.052	0.049	0.051	0.051	0.048	0.053	0.055	0.050
	MG.PA	0.053	0.051	0.045	0.056	0.053	0.048	0.057	0.047	0.048	0.053	0.047	0.053	0.048	0.049
10%	SG.PA	0.046	0.057	0.097	0.143	0.227	0.302	0.371	0.051	0.067	0.129	0.203	0.305	0.386	0.440
	MG.DFTD	0.025	0.029	0.035	0.031	0.033	0.028	0.033	0.032	0.032	0.032	0.033	0.032	0.033	0.032
	MG.SP	0.047	0.049	0.058	0.046	0.051	0.048	0.046	0.051	0.053	0.053	0.052	0.052	0.051	0.055
	MG.PA	0.047	0.045	0.051	0.048	0.053	0.044	0.051	0.049	0.051	0.050	0.053	0.052	0.051	0.051
20%	SG.PA	0.055	0.051	0.094	0.141	0.214	0.266	0.334	0.048	0.070	0.130	0.197	0.295	0.381	0.444
	MG.DFTD	0.032	0.029	0.035	0.035	0.031	0.037	0.036	0.033	0.032	0.033	0.036	0.035	0.037	0.034
	MG.SP	0.052	0.049	0.054	0.053	0.050	0.055	0.049	0.053	0.056	0.054	0.053	0.055	0.059	0.058
	MG.PA	0.051	0.043	0.054	0.060	0.049	0.048	0.046	0.050	0.051	0.054	0.053	0.050	0.055	0.052
30%	SG.PA	0.053	0.065	0.078	0.147	0.198	0.245	0.318	0.050	0.071	0.123	0.204	0.291	0.374	0.437
	MG.DFTD	0.037	0.031	0.040	0.037	0.042	0.038	0.041	0.036	0.033	0.039	0.041	0.043	0.043	0.046
	MG.SP	0.060	0.050	0.056	0.060	0.064	0.066	0.069	0.055	0.056	0.059	0.066	0.064	0.075	0.081
	MG.PA	0.056	0.053	0.046	0.049	0.047	0.041	0.047	0.051	0.052	0.053	0.048	0.051	0.052	0.055
40%	SG.PA	0.062	0.050	0.083	0.128	0.201	0.246	0.324	0.045	0.073	0.123	0.190	0.289	0.396	0.442
	MG.DFTD	0.061	0.048	0.049	0.049	0.074	0.076	0.071	0.042	0.054	0.053	0.061	0.063	0.076	0.088
	MG.SP	0.097	0.075	0.093	0.104	0.124	0.136	0.132	0.093	0.097	0.102	0.128	0.124	0.145	0.153
	MG.PA	0.060	0.047	0.049	0.044	0.055	0.056	0.049	0.047	0.052	0.049	0.052	0.049	0.050	0.058

Table A2

Power Rates of DIF Assessment Under Each Condition (DIF Magnitude = 0.6)

		No. of grouping variables = 5							No. of grouping variables = 10						
DIF%	Correlation	0	0.1	0.2	0.3	0.4	0.5	0.6	0	0.1	0.2	0.3	0.4	0.5	0.6
10%	SG.PA	0.912	0.902	0.892	0.842	0.808	0.773	0.705	0.883	0.877	0.847	0.792	0.724	0.708	0.687
	MG.DFTD	0.918	0.930	0.882	0.862	0.853	0.753	0.680	0.883	0.890	0.850	0.804	0.757	0.664	0.573
	MG.SP	0.948	0.960	0.920	0.892	0.893	0.808	0.760	0.921	0.920	0.893	0.857	0.807	0.734	0.650
	MG.PA	0.913	0.922	0.893	0.868	0.852	0.780	0.702	0.885	0.884	0.868	0.830	0.771	0.695	0.602
	TM	0.958	0.963	0.940	0.942	0.927	0.910	0.885	0.928	0.936	0.917	0.896	0.857	0.818	0.782
20%	SG.PA	0.937	0.926	0.905	0.890	0.857	0.804	0.774	0.886	0.890	0.843	0.797	0.737	0.729	0.665
	MG.DFTD	0.927	0.918	0.892	0.845	0.826	0.731	0.650	0.889	0.891	0.858	0.797	0.733	0.635	0.536
	MG.SP	0.956	0.953	0.924	0.898	0.864	0.798	0.748	0.928	0.929	0.895	0.860	0.799	0.712	0.604
	MG.PA	0.940	0.927	0.909	0.862	0.849	0.785	0.710	0.894	0.887	0.861	0.831	0.768	0.707	0.595
	TM	0.958	0.961	0.947	0.937	0.928	0.919	0.874	0.931	0.942	0.920	0.908	0.883	0.836	0.783
30%	SG.PA	0.924	0.911	0.909	0.854	0.847	0.795	0.763	0.890	0.892	0.872	0.810	0.752	0.717	0.674
	MG.DFTD	0.917	0.911	0.885	0.844	0.779	0.689	0.601	0.887	0.879	0.834	0.794	0.704	0.614	0.491
	MG.SP	0.943	0.936	0.921	0.877	0.831	0.726	0.649	0.918	0.911	0.881	0.836	0.772	0.669	0.548
	MG.PA	0.928	0.914	0.893	0.849	0.842	0.778	0.692	0.891	0.894	0.881	0.832	0.784	0.720	0.606
	TM	0.954	0.950	0.953	0.939	0.939	0.912	0.897	0.936	0.937	0.929	0.916	0.894	0.853	0.794
40%	SG.PA	0.918	0.935	0.908	0.875	0.843	0.808	0.758	0.907	0.899	0.861	0.815	0.778	0.726	0.669
	MG.DFTD	0.862	0.881	0.845	0.796	0.690	0.630	0.490	0.854	0.833	0.792	0.716	0.646	0.510	0.396
	MG.SP	0.884	0.905	0.847	0.810	0.725	0.633	0.504	0.866	0.842	0.807	0.720	0.650	0.518	0.411
	MG.PA	0.920	0.927	0.898	0.864	0.843	0.779	0.691	0.907	0.893	0.871	0.832	0.796	0.712	0.638
	TM	0.957	0.957	0.947	0.942	0.936	0.915	0.895	0.942	0.935	0.931	0.915	0.898	0.866	0.808

Table A3

Accuracy of Anchor Item Selection with the MG.DFTD Method (DIF Magnitude = 0.4)

No. of grouping variables	DIF percentage	Correlation						
		0	0.1	0.2	0.3	0.4	0.5	0.6
5	10%	0.991	0.990	0.984	0.990	0.980	0.981	0.960
	20%	0.966	0.966	0.965	0.955	0.958	0.949	0.908
	30%	0.949	0.936	0.915	0.918	0.881	0.870	0.833
	40%	0.836	0.815	0.824	0.789	0.773	0.730	0.733
10	10%	0.989	0.988	0.982	0.978	0.975	0.963	0.965
	20%	0.958	0.964	0.954	0.945	0.929	0.906	0.894
	30%	0.929	0.921	0.908	0.896	0.865	0.850	0.826
	40%	0.817	0.795	0.806	0.759	0.722	0.706	0.678

Table A4

Type I Error Rates of DIF Assessment Under Each Condition (DIF Magnitude = 0.4)

		No. of grouping variables = 5							No. of grouping variables = 10						
DIF%	Correlation	0	0.1	0.2	0.3	0.4	0.5	0.6	0	0.1	0.2	0.3	0.4	0.5	0.6
0%	SG.PA	0.052	0.051	0.076	0.110	0.135	0.184	0.241	0.058	0.065	0.095	0.151	0.218	0.273	0.341
	MG.DFTD	0.032	0.029	0.031	0.029	0.030	0.033	0.028	0.031	0.029	0.029	0.032	0.031	0.032	0.030
	MG.SP	0.054	0.052	0.051	0.048	0.048	0.052	0.047	0.054	0.049	0.052	0.054	0.052	0.053	0.051
	MG.PA	0.051	0.046	0.052	0.049	0.045	0.049	0.049	0.056	0.054	0.051	0.054	0.057	0.054	0.055
10%	SG.PA	0.043	0.059	0.070	0.099	0.137	0.190	0.230	0.049	0.068	0.093	0.134	0.206	0.263	0.334
	MG.DFTD	0.027	0.031	0.034	0.037	0.032	0.031	0.035	0.032	0.034	0.032	0.035	0.035	0.033	0.032
	MG.SP	0.048	0.054	0.055	0.057	0.051	0.048	0.056	0.051	0.058	0.054	0.054	0.054	0.057	0.053
	MG.PA	0.044	0.051	0.049	0.060	0.053	0.048	0.056	0.051	0.057	0.050	0.055	0.054	0.057	0.047
20%	SG.PA	0.046	0.051	0.073	0.097	0.118	0.156	0.208	0.054	0.054	0.089	0.133	0.196	0.250	0.322
	MG.DFTD	0.029	0.038	0.036	0.030	0.035	0.040	0.035	0.034	0.032	0.040	0.038	0.037	0.040	0.039
	MG.SP	0.051	0.065	0.063	0.052	0.056	0.058	0.055	0.059	0.055	0.066	0.062	0.060	0.062	0.059
	MG.PA	0.047	0.052	0.062	0.045	0.050	0.050	0.047	0.055	0.045	0.056	0.053	0.051	0.055	0.051
30%	SG.PA	0.050	0.051	0.071	0.085	0.103	0.151	0.203	0.045	0.058	0.088	0.136	0.194	0.250	0.325
	MG.DFTD	0.044	0.045	0.043	0.047	0.043	0.050	0.043	0.040	0.045	0.048	0.043	0.048	0.042	0.044
	MG.SP	0.071	0.076	0.069	0.072	0.068	0.070	0.069	0.066	0.072	0.076	0.079	0.080	0.074	0.079
	MG.PA	0.053	0.046	0.051	0.048	0.044	0.046	0.053	0.046	0.051	0.054	0.050	0.049	0.052	0.055
40%	SG.PA	0.046	0.046	0.066	0.093	0.120	0.154	0.176	0.048	0.058	0.085	0.131	0.193	0.249	0.316
	MG.DFTD	0.075	0.076	0.071	0.074	0.074	0.087	0.069	0.077	0.076	0.076	0.080	0.078	0.069	0.068
	MG.SP	0.124	0.121	0.127	0.132	0.134	0.119	0.105	0.121	0.126	0.128	0.127	0.117	0.115	0.109
	MG.PA	0.048	0.049	0.050	0.053	0.056	0.051	0.049	0.050	0.049	0.049	0.052	0.051	0.048	0.052

Table A5

Power Rates of DIF Assessment Under Each Condition (DIF Magnitude = 0.4)

		No. of grouping variables = 5							No. of grouping variables = 10						
DIF%	Correlation	0	0.1	0.2	0.3	0.4	0.5	0.6	0	0.1	0.2	0.3	0.4	0.5	0.6
10%	SG.PA	0.657	0.638	0.628	0.630	0.610	0.580	0.587	0.641	0.626	0.613	0.617	0.592	0.576	0.536
	MG.DFTD	0.623	0.580	0.582	0.550	0.482	0.422	0.312	0.613	0.591	0.538	0.478	0.437	0.368	0.303
	MG.SP	0.720	0.670	0.640	0.620	0.568	0.482	0.382	0.690	0.676	0.628	0.583	0.517	0.456	0.393
	MG.PA	0.657	0.620	0.603	0.572	0.533	0.482	0.367	0.640	0.627	0.576	0.548	0.494	0.432	0.348
	TM	0.740	0.693	0.712	0.692	0.688	0.647	0.588	0.708	0.713	0.690	0.683	0.628	0.563	0.521
20%	SG.PA	0.668	0.638	0.647	0.682	0.612	0.628	0.595	0.651	0.629	0.618	0.600	0.565	0.558	0.553
	MG.DFTD	0.603	0.564	0.535	0.531	0.433	0.371	0.293	0.575	0.550	0.496	0.429	0.371	0.323	0.258
	MG.SP	0.668	0.623	0.625	0.636	0.522	0.433	0.343	0.659	0.638	0.579	0.521	0.462	0.389	0.333
	MG.PA	0.668	0.623	0.623	0.615	0.511	0.496	0.364	0.658	0.619	0.583	0.543	0.463	0.411	0.354
	TM	0.730	0.705	0.716	0.727	0.675	0.648	0.586	0.718	0.706	0.692	0.660	0.625	0.576	0.513
30%	SG.PA	0.622	0.648	0.606	0.642	0.615	0.603	0.565	0.656	0.616	0.613	0.588	0.586	0.569	0.564
	MG.DFTD	0.530	0.530	0.473	0.441	0.376	0.290	0.251	0.530	0.484	0.433	0.397	0.321	0.277	0.219
	MG.SP	0.593	0.597	0.558	0.513	0.443	0.351	0.283	0.613	0.553	0.504	0.451	0.396	0.334	0.262
	MG.PA	0.621	0.639	0.574	0.581	0.535	0.453	0.372	0.660	0.609	0.578	0.541	0.471	0.422	0.367
	TM	0.723	0.738	0.704	0.695	0.679	0.642	0.580	0.721	0.699	0.687	0.669	0.635	0.574	0.519
40%	SG.PA	0.684	0.650	0.625	0.631	0.619	0.583	0.589	0.644	0.621	0.609	0.586	0.565	0.529	0.567
	MG.DFTD	0.445	0.418	0.381	0.327	0.304	0.238	0.188	0.423	0.385	0.346	0.293	0.237	0.206	0.158
	MG.SP	0.481	0.479	0.421	0.365	0.322	0.299	0.224	0.465	0.436	0.390	0.341	0.280	0.250	0.194
	MG.PA	0.682	0.647	0.594	0.579	0.538	0.466	0.391	0.646	0.598	0.579	0.535	0.453	0.419	0.356
	TM	0.738	0.723	0.705	0.704	0.710	0.677	0.628	0.713	0.696	0.700	0.672	0.615	0.576	0.523

Table A6

Coefficient Estimates for DIF Effect Under Each Grouping Variable in the Empirical Study

	MG.DFTD					MG.SP					SG.DFTD				
Item	V1	V2	V3	V4	V5	V1	V2	V3	V4	V5	V1	V2	V3	V4	V5
1	0.208	-0.034	-0.068	-0.029	0.151	0.164	-0.014	0.003	-0.024	0.132	0.175	0.005	-0.081	-0.011	0.164
2	0.119	-0.024	-0.050	-0.007	0.146	0.089	-0.011	-0.027	0.003	0.133	0.103	0.005	-0.060	-0.005	0.155
3	0.247	-0.112	-0.063	-0.047	0.185	0.195	-0.087	0.012	-0.041	0.162	0.183	-0.062	-0.078	-0.017	0.197
4	0.153	0.007	-0.051	-0.017	0.114	0.124	0.021	-0.011	-0.011	0.102	0.144	0.040	-0.063	-0.013	0.125
5	0.232	-0.046	-0.094	-0.017	0.192	0.179	-0.023	-0.015	-0.008	0.168	0.190	-0.003	-0.106	0.002	0.202
6	0.227	-0.048	-0.113	-0.039	0.224	0.163	-0.024	-0.036	-0.034	0.196	0.178	-0.008	-0.125	-0.011	0.238
7	0.159	-0.083	-0.012	0.047	0.028	0.133	-0.071	0.012	0.058	0.022	0.125	-0.058	-0.010	0.077	0.038
8	0.017	-0.039	0.001	-0.022	0.041	0.002	-0.034	-0.011	-0.013	0.039	1.000	-0.036	1.000	-0.010	0.049
9	-0.009	0.037	0.007	0.005	0.032	-0.029	0.047	0.017	0.013	0.028	0.000	0.026	1.000	0.025	0.038
10	0.068	0.006	0.008	0.004	0.018	0.051	0.017	0.016	0.012	0.014	0.069	0.015	0.005	0.026	0.028
11	-0.025	0.049	1.000	1.000	0.071	-0.043	0.059	0.003	0.001	0.066	-0.007	0.041	1.000	1.000	0.078
12	-0.028	-0.045	0.025	0.064	-0.017	-0.053	-0.035	0.037	0.072	-0.022	-0.048	-0.072	0.038	0.105	-0.019
13	0.091	0.029	-0.021	0.052	0.010	0.069	0.048	-0.005	0.064	0.010	0.112	0.037	-0.024	0.077	0.032
14	0.051	0.005	-0.020	0.046	-0.028	0.026	0.022	-0.001	0.056	-0.031	0.062	0.000	-0.022	0.076	-0.009
15	0.023	0.024	-0.019	0.058	0.002	-0.008	0.045	0.003	0.070	0.000	0.041	0.013	-0.020	0.088	0.019
16	0.031	-0.045	-0.004	-0.001	0.034	0.012	-0.036	0.004	0.007	0.030	0.012	-0.044	-0.005	0.021	0.041
17	-0.012	0.036	0.041	-0.024	0.064	-0.037	0.049	0.049	-0.014	0.062	-0.001	0.023	0.030	0.005	0.075
18	0.008	0.014	1.000	0.083	-0.016	-0.014	0.026	0.019	0.089	-0.022	1.000	1.000	0.014	0.117	-0.017
19	0.041	0.024	0.011	1.000	0.051	0.023	0.035	0.029	0.001	0.048	0.049	0.028	1.000	1.000	0.062
20	0.055	1.000	0.061	-0.017	-0.042	0.033	0.021	0.083	-0.011	-0.046	0.051	1.000	0.052	0.034	-0.029
21	0.003	1.000	0.015	0.001	0.010	-0.024	0.020	0.029	0.011	0.005	-0.005	-0.008	0.018	0.037	0.018
22	1.000	-0.082	0.031	1.000	1.000	-0.014	-0.076	0.028	0.020	-0.006	-0.031	-0.090	0.031	0.038	1.000
23	-0.063	0.042	0.066	-0.033	1.000	-0.067	0.045	0.060	-0.028	0.023	-0.043	1.000	0.052	1.000	1.000
24	-0.058	0.040	0.026	-0.014	0.043	-0.064	0.044	0.018	-0.010	0.042	-0.041	0.024	0.018	1.000	0.046
25	-0.042	-0.032	1.000	0.065	-0.051	-0.050	-0.027	0.011	0.067	-0.053	-0.056	-0.055	0.030	0.088	-0.056
26	1.000	1.000	0.111	1.000	1.000	-0.002	-0.015	0.109	0.018	-0.004	1.000	-0.031	0.110	0.050	1.000
27	1.000	1.000	0.053	0.071	1.000	-0.024	-0.007	0.052	0.076	0.002	1.000	-0.031	0.066	0.098	1.000
28	1.000	0.017	1.000	0.069	-0.060	0.000	0.024	0.005	0.078	-0.063	0.025	1.000	0.005	0.103	-0.052
29	-0.016	-0.012	-0.029	0.059	-0.063	-0.039	0.001	-0.010	0.065	-0.068	-0.021	-0.034	-0.017	0.092	-0.058

Note. V1: Having a computer or tablet at home; V2: Having an internet connection at home; V3: Parental education level; V4: Expected education level; V5: Frequency of working alone.

Table A7

P-Values for DIF Detection Under Each Grouping Variable in the Empirical Study

	MG.DFTD					MG.SP					SG.DFTD				
Item	V1	V2	V3	V4	V5	V1	V2	V3	V4	V5	V1	V2	V3	V4	V5
1	0.023	0.583	0.161	0.589	0.089	0.000	0.675	0.903	0.379	0.000	0.003	0.904	0.061	0.820	0.022
2	0.086	0.625	0.166	0.871	0.023	0.030	0.750	0.285	0.899	0.000	0.034	0.904	0.061	0.883	0.004
3	0.017	0.111	0.234	0.443	0.073	0.000	0.005	0.617	0.111	0.000	0.005	0.195	0.086	0.743	0.017
4	0.019	0.887	0.156	0.675	0.060	0.002	0.506	0.656	0.665	0.000	0.002	0.273	0.049	0.713	0.012
5	0.028	0.523	0.084	0.779	0.067	0.000	0.473	0.541	0.766	0.000	0.004	0.958	0.024	0.966	0.015
6	0.077	0.570	0.074	0.599	0.083	0.000	0.420	0.106	0.157	0.000	0.019	0.891	0.019	0.866	0.021
7	0.039	0.144	0.775	0.327	0.685	0.011	0.099	0.715	0.103	0.548	0.047	0.237	0.808	0.095	0.543
8	0.701	0.241	0.969	0.422	0.273	0.958	0.214	0.610	0.569	0.079	1.000	0.216	1.000	0.708	0.151
9	0.881	0.397	0.816	0.888	0.513	0.500	0.181	0.523	0.647	0.331	0.995	0.489	1.000	0.481	0.401
10	0.237	0.882	0.798	0.922	0.715	0.229	0.631	0.542	0.674	0.626	0.149	0.693	0.857	0.463	0.535
11	0.649	0.235	1.000	1.000	0.116	0.310	0.097	0.913	0.983	0.023	0.881	0.260	1.000	1.000	0.064
12	0.685	0.384	0.508	0.142	0.764	0.301	0.415	0.254	0.039	0.523	0.426	0.133	0.309	0.017	0.729
13	0.311	0.665	0.660	0.356	0.898	0.238	0.313	0.886	0.101	0.808	0.129	0.511	0.607	0.167	0.663
14	0.553	0.933	0.667	0.393	0.716	0.655	0.650	0.975	0.164	0.442	0.392	0.999	0.656	0.161	0.898
15	0.800	0.711	0.698	0.304	0.980	0.886	0.337	0.932	0.066	0.991	0.578	0.817	0.678	0.116	0.801
16	0.596	0.309	0.904	0.974	0.512	0.781	0.301	0.864	0.795	0.301	0.797	0.243	0.875	0.557	0.383
17	0.867	0.513	0.300	0.605	0.313	0.473	0.251	0.129	0.697	0.084	0.985	0.625	0.422	0.905	0.198
18	0.915	0.819	1.000	0.098	0.808	0.816	0.611	0.615	0.035	0.597	1.000	1.000	0.759	0.019	0.784
19	0.568	0.663	0.774	1.000	0.380	0.686	0.466	0.418	0.973	0.224	0.422	0.569	1.000	1.000	0.252
20	0.466	1.000	0.163	0.731	0.516	0.588	0.681	0.026	0.783	0.262	0.457	1.000	0.228	0.493	0.634
21	0.966	1.000	0.730	0.985	0.878	0.706	0.705	0.455	0.796	0.909	0.946	0.879	0.686	0.463	0.779
22	1.000	0.023	0.263	1.000	1.000	0.739	0.034	0.289	0.501	0.837	0.475	0.010	0.262	0.209	1.000
23	0.336	0.434	0.111	0.454	1.000	0.293	0.399	0.136	0.518	0.599	0.487	1.000	0.200	1.000	1.000
24	0.142	0.209	0.310	0.597	0.129	0.084	0.160	0.467	0.706	0.094	0.256	0.422	0.461	1.000	0.109
25	0.551	0.579	1.000	0.164	0.310	0.454	0.631	0.788	0.149	0.249	0.397	0.314	0.484	0.056	0.269
26	1.000	1.000	0.010	1.000	1.000	0.976	0.790	0.011	0.705	0.931	1.000	0.573	0.010	0.270	1.000
27	1.000	1.000	0.144	0.079	1.000	0.658	0.884	0.130	0.046	0.952	1.000	0.507	0.070	0.014	1.000
28	1.000	0.768	1.000	0.149	0.328	1.000	0.632	0.902	0.057	0.126	0.718	1.000	0.912	0.035	0.378
29	0.835	0.845	0.506	0.238	0.321	0.521	0.985	0.788	0.115	0.101	0.757	0.531	0.702	0.065	0.338

Note. V1: Having a computer or tablet at home; V2: Having an internet connection at home; V3: Parental education level; V4: Expected education level; V5: Frequency of working alone.