An Exploratory Strategy to Identify and Define Sources of Differential Item Functioning

Applied Psychological Measurement
2020, Vol. 44(7-8) 548–560
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0146621620931190
journals.sagepub.com/home/apm

Chung-Ping Cheng^{1,2}, Chi-Chen Chen³ and Ching-Lin Shih^{3,4}

Abstract

The sources of differential item functioning (DIF) items are usually identified through a qualitative content review by a panel of experts. However, the differential functioning for some DIF items might have been caused by reasons outside of the experts' experiences, leading to the sources for these DIF items possibly being misidentified. Quantitative methods can help to provide useful information, such as the DIF status and the number of sources of the DIF, which in turn help the item review and revision process to be more efficient and precise. However, the current quantitative methods assume all possible sources should be known in advance and collected to accompany the item response data, which is not always the case in reality. To this end, an exploratory strategy, combined with the MIMIC (multiple-indicator multiple-cause) method, that can be used to identify and name new sources of DIF is proposed in this study. The performance of this strategy was investigated through simulation. The results showed that when a set of DIF-free items can be correctly identified to define the main dimension, the proposed exploratory MIMIC method can accurately recover a number of possible sources of DIF and the items that belong to each. A real data analysis was also implemented to demonstrate how this strategy can be used in reality. The results and findings of this study are further discussed.

Keywords

differential item functioning, factor analysis, source of DIF

The existence of differential item functioning (DIF) can highly influence test validity, test fairness, and score comparability for various examinee groups. The influence of DIF increases as the number of DIF items in the test increases. To ensure test validity and fairness, a DIF assessment has become a routine item analysis procedure during test construction. Developing statistical methods to assess DIF items is an important task, and their initial use was viewed as the beginning of the second generation of DIF analysis (Zumbo, 2007). Although DIF assessment methods have advanced over the years, and their results have become more accurate, test

¹National Cheng Kung University, Tainan, Taiwan

²College Entrance Examination Center, Taipei, Taiwan

³National Sun Yat-sen University, Kaohsiung, Taiwan

⁴University of British Columbia, Vancouver, BC, Canada

Corresponding Author:

Ching-Lin Shih, Associate Professor, Institute of Education and Center for Teacher Education, Assessment Research Center, National Sun Yat-sen University, 70 Lienhai Rd., Kaohsiung 804, Taiwan.

Email: educls@g-mail.nsysu.edu.tw

developers and item writers are often confronted by a predicament, namely that “no amount of deliberation seems to help explain why some perfectly reasonable items have large DIF value” (Angoff, 1993, p. 19). Therefore, solely providing the results of a DIF assessment without including qualitative information about the sources of DIF fails to guide the item revision process for item writers. In contrast, when the sources of DIF are known, an explanation of why the items function differently for various groups of examinees is more likely to be revealed. In turn, the item revision process can be carefully monitored, and the test validity and fairness can be preserved during the test construction.

To identify sources for the DIF items, both qualitative and quantitative analysis procedures can be used; this combined approach marks the third generation of DIF analysis (Zumbo, 2007). When applying qualitative methods to identify these sources, a panel of experts is typically asked to execute a judgmental review of DIF items and explain why these items pose higher difficulty for one group of examinees than another (Camilli & Shepard, 1994). For example, when analyzing item response data from the Israeli Psychometrics Entrance Test in both the Hebrew and Russian versions, Allalouf et al. (1999) found that 34% of the items exhibited DIF. An eight-member committee was then asked to identify possible sources for 42 DIF items. There were seven items for which the committee could not reach an agreement regarding a source; the remaining 35 DIF items were deemed caused by the following four possible sources of DIF: change in word difficulty (16 items), change in item format (five items), difference in cultural relevance (six items), and change in content (eight items). Moreover, adaptation effects and curricular differences were found as two additional potential sources of DIF in the multilanguage versions of the assessments. Within these identified DIF items, six mathematics and nine science items were interpreted to be related to adaptation differences (Ercikan, 2002). As to the effectiveness of using qualitative methods to identify sources of DIF, more than 80% of the sources of DIF items were identified by experts for the Hebrew and Russian versions of the Israeli Psychometrics Entrance Test (Allalouf et al., 1999). In addition, for mathematics items on the English and French versions of Canadian tests, bilingual reviewers identified that 36% and 38% of the DIF items were caused by adaptation differences for the 13-year-old and 16-year-old age groups, respectively (Ercikan et al., 2004).

Some researchers use a statistical approach to explore sources of DIF, such as the hierarchical logistic regression (HLR) model (Swanson et al., 2002) and the mediated multiple-indicator multiple-cause (MIMIC) method (Cheng et al., 2015). When using the HLR model to analyze DIF for a data set of U.S. medical licensing examinations, Swanson et al. (2002) predicted gender DIF with several variables; two of these variables, “patient gender” and “medical discipline,” significantly explained the variance of DIF coefficients. In another study, Cheng et al. (2015) analyzed eight items that measured the enjoyment of science for eighth-grade students in a 2007 Trends in International Mathematics and Science Study (TIMSS) data set. To investigate the sources for three DIF items, all the variables included in the data set were examined. The researchers found that the total score of the self-confidence in learning science and math scale was the complete mediator and partial mediator, or sources of DIF, for Item 5 (“Science is not one of my strengths”) and Item 6 (“I learn things quickly in science”), respectively.

While previous research identified the sources of up to 87.5% of the DIF items, sources for the other DIF items remained unclear. This may be because the review of experts was based on their professional experience. However, the differential functioning for some DIF items might have been caused by reasons outside of the experts’ experiences. Furthermore, the experts might have performed a subjective review of the items for the DIF, and their impressions may have been inconsistent with the results of the statistical DIF analyses (Sireci et al., 1998). In addition, for the HLR and the mediated MIMIC methods, the variables deemed as possible sources of DIF must be collected with the item responses. Because the sources for some DIF items might

not exist in the collected variables, such a confirmatory approach is therefore not available. Consequently, a more exploratory approach that can help researchers self-define the possible sources for DIF items is needed and must be developed. In addition, as previously described, Allalouf et al. (1999) and Ercikan (2002) both showed that several DIF items within a test might share a common cause, that is, a secondary dimension other than the latent trait θ (which will be formally defined later in this article). In other words, these studies provided evidence that the DIF might be due to the presence of secondary dimensions, which fits the assumptions of the multidimensional model for DIF (MMD; Shealy & Stout, 1993). Therefore, the goal of this study was to propose a statistical strategy that was based on the MMD model and could identify the possible sources of DIF items that might not exist in the collected variables. By combining such a strategy with DIF assessment methods, valuable information could be provided to item writers to revise DIF items and ensure item quality and test validity.

The article is organized as follows. First, the literature dealing with the statistical method of identifying and explaining sources of DIF is reviewed, followed by a proposed strategy that is combined with the MIMIC method to identify possible sources of DIF. In addition, the performance of this method under various conditions was investigated through a series of simulation studies and a real data analysis. Finally, the conclusions and implications of this strategy are explained.

The Mediated and Exploratory MIMIC Method to Identify Sources of DIF

The current statistical methods for identifying sources of DIF consider the variables that are collected with item responses as the possible sources of DIF. These methods were deemed as confirmatory approaches in this study and will be introduced first. Assume a test measures a latent trait θ , and a grouping variable z (e.g., gender) has indirect effects on items through θ . For example, examinees of different genders might have different levels of latent trait θ and therefore yield different performances on the latent response variable y_i^* . In the measurement component of the MIMIC method, the variables suspected to be related to DIF take the factor-analytic model form as follows (B. O. Muthén et al., 1991):

$$y_i^* = \lambda_i \theta + \beta_i' z + \varepsilon_i, \quad (1)$$

where λ_i is the factor loading of θ on item i , and ε_i is the residual and has a standard normal distribution. β_i represents the direct effect from the grouping variables z (gender) to y_i^* . In the MIMIC model, a DIF assessment for item i is implemented by testing whether the direct effect β_i exists or not. If $\beta_i \neq 0$, a direct effect exists from z to y_i^* so that a uniform DIF is found in item i (Shih & Wang, 2009; Wang & Shih, 2010).

As y_i^* cannot be observed directly, test items are adopted to measure y_i^* such that y_i^* is transformed to observed ordinal responses y_i via a threshold model as follows:

$$y_i = \begin{cases} 0 & \text{if } y_i^* \leq \tau_{i1} \\ 1 & \text{if } \tau_{i1} \leq y_i^* \leq \tau_{i2} \\ \dots & \dots \\ J & \text{if } \tau_{iJ} < y_i^* \end{cases}, \quad (2)$$

where τ_{ij} is the threshold parameter of step j in item i . Given the residual ε_i follows a standard normal distribution, $p(y_i \leq k | \theta)$ will be a normal cumulative function of θ .

To explore the possible sources of DIF within the MIMIC model, Cheng et al. (2015) included variables and tested whether they could partially or fully mediate the DIF effect. In their research, they used both simulation study and real data analysis to investigate the performance of the mediated MIMIC model on DIF assessment. The results indicated the mediated MIMIC method could successfully detect the mediation effect when it occurred in the context of DIF. Within the mediated MIMIC method, however, the mediator should be an existing variable that is collected before the DIF analysis, which is not always the case. Therefore, after a preferred confirmatory method is used, several to many DIF items whose sources remain unclear may still exist for which a following exploratory approach is then expected to identify the sources.

Within the framework of an MMD, the concept that the test intended to measure is called the primary dimension. Other dimensions that produce DIF are referred to as the secondary dimensions. The secondary dimensions are further determined as auxiliary or nuisance dimensions, depending on whether they were assessed intentionally or unintentionally as a part of the concepts that were measured by the test. Therefore, identifying sources of DIF items is equivalent to determining the test's secondary dimensions. To develop an exploratory method that can be used to identify possible sources of DIF outside the existing variables, the most important procedures include finding and, more critically, naming the possible sources. To do that, a well-developed technique known as exploratory factor analysis (EFA; Spearman, 1904) can be used. As the exploratory method proposed here was combined with the MIMIC method, it will be referred to as the exploratory MIMIC method hereafter.

Applying EFA to Investigate Sources of DIF

In routine item analysis, DIF assessment may be carried out for several grouping variables, such as gender and ethnicity, which can result in many DIF items. The sources of all these DIF items can be identified simultaneously to help the item writers better revise the DIF items and ensure the test quality, no matter what the grouping variable is. Using the exploratory MIMIC method to identify the possible sources for DIF items, the subjects' responses are assumed to be determined by the two dimensions: the primary dimension and a secondary dimension. To better depict the exploratory MIMIC method and its rationale, the framework of this method is shown in Figure 1. Three grouping variables and two sources of DIF are included in the figure, indicating that the sources of the DIF items assessed for the three grouping variables were identified simultaneously. Two sources of DIF were represented: one for the auxiliary dimension and the other for the nuisance dimension. In addition, the DIF-free items (Items 1 to k) measured the primary dimension (θ) only, whereas the DIF items (Items p to s) measured both the primary dimension and one of the secondary dimensions; Items p to q measured the primary dimension and the first secondary dimension, while Items r to s measured the primary dimension and the second secondary dimension. The arrows connecting the grouping variables to the primary dimension in the figure indicate the mean ability difference between the groups (or impact). The double arrow between the factors or indicators denotes an existing correlation that is being freely estimated.

The exploratory MIMIC method identifies the possible source of DIF through the following steps:

1. *Assessing the DIF for items with the DIF-free-then-DIF (DFTD) strategy.* To assess DIF with a well-controlled Type I error rate, the DFTD strategy (Wang et al., 2012) was used in this study. The MIMIC method with iterative constant item (CI) procedure (Shih & Wang, 2009) was used to identify four DIF-free items, which in turn were used as the anchor of the

Figure 1. The framework of test items, sources of DIFs, main ability, and grouping variables in this study.
 Note. DIF = differential item functioning.

CI method. The CI method was then used to assess the DIF for every item, except for the anchors. This procedure was implemented for each of the three grouping variables to determine all the possible DIF items. The detailed procedures of the DFTD strategy, the CI procedure and the iterative CI procedure, as well as their performance, are introduced in Online Appendix A.

2. *Determining the primary dimension with the DIF-free items.* After a set of DIF-free items were selected across the various grouping variables from Step 1, an MIMIC analysis that only contains these DIF-free items was then performed to obtain the estimates for the model parameters.

3. *Estimating the correlations between the residuals of the DIF items.* All the DIF items were added back to the test, and another MIMIC analysis was then carried out to obtain the parameter estimates for these DIF items by fixing the parameters of the DIF-free items. After modeling the primary dimension, the correlations between the residuals for each pair of DIF items were freely estimated and then used as the correlation matrix that being analyzed in the following procedure. By doing this, the variance of the primary dimension was excluded from the total variance of the item responses.

4. *Performing an EFA for the correlation matrix of residuals.* An EFA using unweighted least squares with an oblique rotation was implemented on the correlation matrix of residuals

of all the DIF items to extract factors. The number of factors contained in the residuals were then ascertained by a parallel analysis.

5. *Determining and naming the sources of the DIF.* The estimated factor loadings on the extracted factors for each DIF item were used to decide which factor was more determinant on the item than the other factors. Items that measured identical factors or had a higher factor loading on the same factor were deemed as collectively forming a factor and then underwent a qualitative review. The reviewers then named the factor, yielding a potential new source of DIF for further investigation.

Two crucial components of the exploratory MIMIC method were the qualitative review and the subsequent naming procedure of the EFA. Although the performance of EFA has been investigated thoroughly in the literature, this study was the first one using EFA to identify the possible sources of DIF. Therefore, the performance of using EFA to identify the sources for DIF items was evaluated through a series of simulation studies and a real data analysis.

Simulation Study I

Design

Seven independent variables were manipulated in this simulation study: (a) number of grouping variables (1 and 3); (b) scoring points of the items (2, 3, 4, and 5); (c) sample size (R288/F144, R216/F216, R720/F360, and R540/F540, where R and F represent the reference and focal group, respectively); (d) magnitude of impact, that is, difference in the groups' theta means (0.0, 0.5, and 1.0); (e) percentage of DIF items in the test (40% and 80%); (f) item assignment for each source of DIF (equal numbers of items and nonequal numbers of items); and (g) correlation between sources of DIF (0.0, 0.3, and 0.6, which indicted zero, low, and medium-related, respectively). Two dependent variables were used: accuracy in correctly identifying the number of factors (accuracy) and per element accuracy (PEA). The reasons for selecting these variables, and the levels, are explained in Online Appendix B.

Other Settings

The responses for DIF items were generated according to a multiple group bi-factor model, and the structural relationship between the dimensions and items is shown in Figure 1. The means of the primary and secondary dimensions were set at 0, whereas the variances were 1.00 and 0.25, respectively. In addition, the factor loadings of the items on the primary and secondary dimensions (i.e., the sources of the DIF) were generated from $N(0.7, 0.05^2)$. This study used only two sources of DIF to depict the scenario that both the auxiliary dimension and nuisance dimension might exist simultaneously in reality. Too many sources might have complicated the conditions, which would be beyond the scope of this study. The test contained 40 items, and the difference in threshold between the groups (i.e., the magnitude of the DIF) was set at 0.4. The authors wrote a set of MATLAB 2012 computer programs for the data generation and performed 500 replications under each condition. All the estimation procedures for the exploratory MIMIC model were carried out with Mplus 7 (L. K. Muthén & Muthén, 1998–2012).

The two following dependent variables were used: accuracy in correctly identifying the number of factors (called accuracy hereafter) and PEA. Accuracy was used because knowing the “number of factors as sources of DIF” might affect the degree to which expert review could correctly identify sources of DIF for items (Ercikan, 2002). It was defined as the frequency with

which the method correctly identified two sources of DIF in the data over replications and can be calculated as follows:

$$\text{Accuracy} = \frac{\text{Number correct}}{\text{Number of replications}}.$$

The accuracies ranged from 0 to 1, and higher accuracy indicated better performance of the method. The proportion of items that loaded correctly in each of the two secondary dimensions could be evaluated after the number of DIF sources were correctly identified, which is the formal definition of PEA (Hogarty et al., 2005). In this study, the definition of PEA was modified as follows:

$$\text{PEA} = 1 - \frac{\text{Steps to perfect classification}}{\text{Number of total DIF items}},$$

where the numerator represents the minimum number of steps needed to move all wrongly classified items to the correct categories, given the number of factors is correctly identified as two. For example, of 20 DIF items in the test, the first half of the items were caused by the first source of DIF and the other half were caused by the second source. If the sources of the first seven and the other 13 DIF items were identified as the first and second secondary dimensions, respectively, Items 8 to 10 were moved and therefore resulted in the steps to perfect classification equal to three and $\text{PEA} = (1 - (3/20)) = 0.85$. The PEA fell between 0.5 and 1.0, with a higher value indicating a better method performance. Replications that did not correctly identify the number of factors were excluded from the calculation of PEA.

Results of Simulation Study I

The results of the accuracies and PEAs of the simulation study are listed in Tables 1 and 2, respectively. Because the results under different levels of impact were quite similar, only results under impact = 0.5 are presented here. According to the results in Table 1, the exploratory MIMIC method could correctly identify two secondary dimensions almost perfectly when scoring points greater than two and correlation between sources of DIF no greater than 0.3. The accuracies became lower when the test was scored dichotomously and the sample size was as small as R216/F216 and R288/F144, especially when the two sources of DIF were highly correlated. This may have been because the two factors were quite likely to be identified as one when they were highly correlated. The number of grouping variables showed little effect on the accuracies. Under the 40% DIF and two sources with highly correlated conditions, accuracies were higher for the three grouping variable than for the single grouping variable conditions.

Given that the method correctly identified two sources of DIF, the results shown in Table 2 further indicated that the PEA for correctly classifying sources for DIF items was nearly perfect. A slightly lower PEA occurred when the items were scored dichotomously and when the sample size was small. According to the results in Tables 1 and 2, the EFA correctly recovered the number of DIF sources and identified the items that belonged to each of the sources with high accuracy; these results in turn indicated that the exploratory MIMIC method was recommended to identify the possible sources of DIF for items exhibiting DIF over different grouping variables. In addition to investigating the exploratory MIMIC method under the former scenarios, the performance of this method when part of the DIF-free items were misidentified as DIF and included into the EFA was further explored through Simulation Study 2, which is described in Online Appendix C.

Table 1. The Accuracies of Correctly Identify the Number of Secondary Dimensions in Simulation Study (Impact = 0.5).

Correlation	Sample size	Item assignment	DIF percentage = 40%										DIF percentage = 80%									
			2-point		3-point		4-point		5-point				2-point		3-point		4-point		5-point			
			IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV
.0	R216/F216	Equal	.97	.91	—	—	—	—	—	—	—	—	.99	.86	—	.99	—	—	—	—	—	—
		Unequal	.99	.93	—	—	—	—	—	—	—	—	.99	.83	—	.99	—	—	—	—	—	—
	R288/F144	Equal	.99	.98	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—
		Unequal	—	.99	—	—	—	—	—	—	—	—	—	.98	—	—	—	—	—	—	—	—
	R540/F540	Equal	—	—	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—
.3	R720/F360	Unequal	—	—	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—
		Equal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	R216/F216	Equal	.99	.92	—	—	—	—	—	—	—	—	.99	.86	—	—	—	—	—	—	—	—
		Unequal	.99	.94	—	—	—	—	—	—	—	—	.83	—	—	—	—	—	—	—	—	—
	R288/F144	Equal	.99	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—	—
.6	R540/F540	Unequal	.99	.99	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—
		Equal	—	—	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—
	R720/F360	Unequal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		Equal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	R216/F216	Equal	.69	.77	.58	.72	.49	.72	.43	.66	.99	.87	.99	.87	.99	.99	.96	.99	.98	.98	.98	.98
	R288/F144	Unequal	.62	.75	.46	.63	.35	.55	.29	.45	.99	.86	.97	.99	.99	.99	.96	.99	.98	.98	.98	.98
		Equal	.68	.79	.58	.74	.50	.67	.45	.62	—	.98	.99	.98	.99	.99	.99	.99	.98	.98	.98	.98
	R540/F540	Unequal	.56	.66	.47	.56	.30	.49	.26	.45	.97	.99	.97	.99	.98	.99	.98	.99	.98	.98	.98	.98
		Equal	.98	.99	—	—	.99	—	.99	—	—	—	—	—	—	—	—	—	—	—	—	—
	R720/F360	Unequal	.97	.99	.97	.99	.97	—	.99	—	—	.99	—	.99	—	—	—	—	—	—	—	—
	R720/F360	Equal	.99	—	.99	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		Unequal	.96	.98	.98	—	.98	—	.98	—	—	—	—	—	—	—	—	—	—	—	—	—

Note: DIF = differential item functioning; R = reference group; F = focal group; 2-, 3-, 4-, 5-point indicates the item scores in 2 to 5 points, respectively; GV = grouping variable; equal and unequal mean the number of DIF items assigned to each of two sources of DIF is equal and unequal, respectively; “—” = 1.00.

Table 2. The Per Element Accuracy for Classifying DIF Item to Its Source of DIF (Impact = 0.5).

Correlation	Sample size	Item assignment	DIF percentage = 40%										DIF percentage = 80%									
			2-point		3-point		4-point		5-point		2-point		3-point		4-point		5-point					
			IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV	IGV	3GV
.0	R216/F216	Equal	.99	.95	—	—	—	—	—	—	.97	.87	.99	.96	—	—	—	—	—	—	—	—
		Unequal	.99	.96	—	—	—	—	—	—	.97	.91	.99	.97	—	—	—	—	—	—	—	—
	R288/F144	Equal	.99	.98	—	—	—	—	—	—	.97	.93	.99	.97	—	—	—	—	—	—	—	—
		Unequal	.99	.99	—	—	—	—	—	—	.98	.94	.99	.98	—	—	—	—	—	—	—	—
	R540/F540	Equal	—	—	—	—	—	—	—	—	—	.97	—	—	—	—	—	—	—	—	—	—
.3	R720/F360	Unequal	—	—	—	—	—	—	—	—	—	.98	—	—	—	—	—	—	—	—	—	—
		Equal	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—	—	—
	R216/F216	Unequal	—	—	—	—	—	—	—	—	—	.99	—	—	—	—	—	—	—	—	—	—
		Equal	.99	.98	—	—	—	—	—	—	.99	.95	—	.99	—	—	—	—	—	—	—	—
	R288/F144	Unequal	—	.98	—	—	—	—	—	—	.99	.96	—	—	—	—	—	—	—	—	—	—
.6	R216/F216	Equal	.99	.99	—	—	—	—	—	—	.99	.97	—	—	—	—	—	—	—	—	—	—
		Unequal	.99	.99	—	—	—	—	—	—	.99	.98	—	—	—	—	—	—	—	—	—	—
	R540/F540	Equal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		Unequal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
	R720/F360	Equal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
.6	R216/F216	Equal	.92	.89	.97	.98	.99	.99	—	—	.97	.94	.99	.99	—	—	—	—	—	—	—	—
		Unequal	.94	.93	.98	.99	.99	.99	—	—	.97	.95	.99	.99	—	—	—	—	—	—	—	—
	R288/F144	Equal	.92	.94	.97	.99	.99	.99	—	—	.97	.97	.99	.99	—	—	—	—	—	—	—	—
		Unequal	.94	.96	.98	.99	.99	.99	—	—	.97	.97	.99	.99	—	—	—	—	—	—	—	—
	R540/F540	Equal	.99	.99	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
.6	R720/F360	Equal	.99	.99	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
		Unequal	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

Note. DIF = differential item functioning; R = reference group; F = focal group; 2-, 3-, 4-, 5-point indicates the item scores in 2 to 5 points, respectively; GV = grouping variable; equal and unequal mean the number of DIF items that assigned to each of two sources of DIF is equal and unequal, respectively; “—” = 1.00.

Real Data Analysis

After the possible sources of DIF were identified by the exploratory MIMIC method, the results were then reviewed by content experts. In the real data analysis, the exploratory MIMIC method and the following expert review process were fully implemented.

TIMSS Data

The expert review procedure of the exploratory MIMIC method was illustrated using the 2011 math data from the TIMSS. The Eighth-grade Math Scale Form 1 response data as well as three background variables of the student questionnaire were analyzed (the data set is available at <https://timssandpirls.bc.edu/timss2011/international-database.html>). In the data set, 26 items of Scale Form 1 were administered to 20,457 participants recruited from 50 countries. The three background variables included in the analysis were (a) the amount of books in your home, (b) home educational resources, and (c) parents' highest education level. All the background variables were scored dichotomously; a high score indicated more books at home, more educational resources, and a higher education level of the parents. More than 25 books in the home, many resources in the home, and university or higher as the parents' highest education level were each scored as 1. All other responses were scored as 0.

Analysis

To apply the exploratory MIMIC method that allowed multiple grouping variables to be included, the three background variables were taken one at a time as grouping variables in the DIF assessment. All the estimation procedures for the exploratory MIMIC model were implemented with Mplus 7 (L. K. Muthén & Muthén, 1998–2012). The items deemed DIF for at least one grouping variable were included in the following analysis for sources of DIF.

Results

After assessing the DIF for all 26 items, the data from the Republic of Macedonia showed more DIF items than any other country and were therefore selected for use in this real data analysis. In the TIMSS 2011 Macedonia sample, six items were deemed as DIF (the Item IDs are listed below), where Item M052173 was shown as DIF for all three grouping variables and the other five items were identified as DIF for just one grouping variable. After removing the primary dimension, the residuals were left for the analysis of the possible sources of DIF. The results of the EFA showed two factors existed. The estimated factor loadings of Items M032166, M032760A, M052173, M052302, M052503A, and M052503B were 0.10, 1.00, -0.07, -0.22, -0.14, and 0.00 for Factor 1, respectively, and 0.33, 0.00, -0.31, 0.18, 0.58, and 0.78 for Factor 2, respectively. The factor loadings 1.00 and 0.00 for Item 2 were estimated by the software, rather than the constraints for model identification. For Factor 1, only Item M032760A showed a higher factor loading, whereas Items M052503A and M052503B that were bundled into a testlet both loaded strongly on Factor 2. These two items are shown in Online Appendix D. To explore the possible sources of DIF, these items were reviewed by two high school mathematics teachers who had 5 and 15 years of teaching experience, respectively. Both reviewers deemed that the statistical figures concerning the population pyramid, as well as its explanation, were difficult concepts for eighth-grade students. However, the students who owned more educational resources or had more books at home would be more likely to read or learn about these concepts than other students, which might have caused these items to exhibit DIF. Therefore,

the concept of the population pyramid, which was additional to the mathematics ability measured by Items M052503A and M052503B, was identified as a likely source of these two DIF items.

Discussion

For the past several decades, researchers have attempted to formulate methods to identify sources of DIF in both qualitative and quantitative ways. When the sources of DIF are known, the revision of those items can be more guided; in turn, the items' quality can be carefully monitored, and test validity and fairness can be maintained. For qualitative methods, a panel of experts is usually asked to identify the possible sources for items that have been assessed and deemed as DIF. However, although the chosen experts could identify possible sources of DIF, they could not be expected to determine all possible sources. By providing quantitative information, such as a number of sources of DIF for qualitative review, the experts could identify sources of DIF with a higher accuracy (Ercikan, 2002).

Most quantitative methods take a confirmatory approach, meaning that the variables that are deemed as possible sources should be collected before exploring the sources for DIF items. Although this approach is preferred as it is more theory driven and allows for more thorough interpretations and explanations of DIF (Ferne & Rupp, 2007), it cannot realistically identify all possible sources of DIF. Therefore, the main goal of this study was to propose an exploratory method that could supplement the confirmatory approach method to determine other possible sources for a set of DIF items. The proposed strategy was based on the MMD, indicating that DIF is caused by secondary dimensions that might or might not be measured intentionally by the test. The identification of sources of DIF is therefore equivalent to the identification of secondary dimensions. Before identifying the secondary dimensions, the primary dimension had to be clearly defined by a set of DIF-free items to form a prerequisite condition of this strategy. Next, the residuals were freely estimated, and the correlation matrix was then analyzed. By combining the factor analysis and factor-naming procedures of EFA, the new sources of DIF could be identified by the exploratory MIMIC method. Given the number of factors known as possible sources of DIF through this method, the following expert review process was expected to identify the true sources more accurately. However, if many DIF items in the test are commonly caused by one factor, then the factor structure of this test needs to be further validated carefully.

Other methods also explore the possible sources of DIF based on the assumption of a multidimensionality of the data, such as the mixed dimensionality method (De Boeck et al., 2011). The mixed dimensionality method assumes that DIF exists for a latent class. A second dimension exists in the DIF latent class but not in the non-DIF latent class. Such a method is useful in explaining latent class *DIF*. Given that the strategy of this study was also based on the MMD, the exploratory MIMIC method was different from the mixed dimensionality method in at least three ways. First, the mixed dimensionality approach is used to explain latent class *DIF*, whereas the exploratory MIMIC method is used to explain manifest DIF. Second, the common assumption of multidimensionality of data in both approaches differs in that the mixed dimensionality method assumes that the secondary dimension is limited to the DIF latent class, whereas the exploratory MIMIC method assumes that the secondary dimension exists in both manifest groups (i.e., the reference group and the focal group; De Boeck et al., 2011). Third, although both approaches are meant for DIF explanation, the mixed dimensionality method needs a set of possible sources of DIF that can be investigated, whereas the method proposed by this study could be used to supplement the confirmatory approach and to identify new sources of DIF.

Understanding the secondary dimensions and their effects on DIF provides the “potential for a more accurate interpretation of the test score, more control over the influence of relevant secondary dimensions, and the reduction of influence by unintended and irrelevant secondary dimensions” (Roussos & Stout, 2004, p. 114). This study focused on the condition that DIF items measured the primary dimension and a secondary dimension; future studies can extend this to items that measure the primary dimension and multiple secondary dimensions. Furthermore, as measuring both the primary dimension and a secondary dimension can be taken as a property of an item, other approaches that use item properties as possible sources of DIF could be an alternative way of explaining DIF and worth further investigation.

Several limitations exist for the newly proposed strategy. For example, this strategy may not be useful when only one or a few items are influenced by each secondary dimension, because a factor that is only measured by one or two items through EFA is not expected to be identified. In real conditions, this method should be used to identify possible sources for DIF items that belong to an item bank or a long questionnaire that might consist of many subscales in which many DIF items might exist. Besides, this method is most valuable when multiple DIF items share a common source of DIF. Although such a scenario is stringent, it is not rare in reality. For example, Allalouf et al. (1999) found four possible sources for 35 DIF items and each of these possible sources caused DIF for a number of items, which would have made the exploratory MIMIC method valuable in that case. In addition, the current version of this strategy is used to identify possible sources for uniform DIF items. Furthermore, because a confirmatory approach method is more theory-driven, it is recommended the confirmatory approach should be used to identify sources for DIF items before this strategy is used. Finally, although seven independent variables were manipulated in the simulation study, the results should be extended to other conditions with caution. For example, when more than two grouping variables exist or when the number of sources of DIF increases, the resulting conditions become more complex, and it is hard to explore all conditions thoroughly. Researchers are encouraged to investigate this method further.

Acknowledgments

The authors expressed their deeply thanks to Editor, Associate Editor, and three anonymous reviewers for their valuable comments.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Science and Technology of Taiwan (Grant Number MOST 104-2410-H-110 -051 -).

ORCID iD

Ching-Lin Shih <https://orcid.org/0000-0002-0655-849X>

Supplemental Material

Supplemental material for this article is available online.

References

- Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement*, 36(3), 185–198.
- Angoff, W. H. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Lawrence Erlbaum.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. SAGE.
- Cheng, Y., Shao, C., & Lathrop, Q. N. (2015). The mediated MIMIC model for understanding the underlying mechanism of DIF. *Educational and Psychological Measurement*, 76, 43–63.
- De Boeck, P., Cho, S. J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, 35(8), 583–603.
- Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing*, 2, 199–215.
- Ercikan, K., Gierl, M. J., McCreith, T., Puhon, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education*, 17(3), 301–321.
- Ferne, T., & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly*, 4(2), 113–148.
- Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: the influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, 65, 202–226.
- Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, 28(1), 1–22.
- Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.).
- Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107–116). SAGE.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, 58, 159–194.
- 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, 184–199.
- Sireci, S. G., Fitzgerald, C., & Xing, D. (1998). *Adapting credentialing examinations for international uses* [Paper presentation]. Annual meeting of the American Educational Research Association, San Diego, CA.
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292.
- Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, 27, 53–75.
- Wang, W. C., & Shih, C. L. (2010). MIMIC methods for assessing differential item functioning in polytomous items. *Applied Psychological Measurement*, 34, 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, 687–708.
- Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223–233.