MULTIVARIATE BEHAVIORAL RESEARCH
2024, VOL. 59, NO. 1, 46–61

Combining Item Purification and Multiple Comparison Adjustment Methods in Detection of Differential Item Functioning

Adéla Hladká[a,b], Patrícia Martinková[a,c], and David Magis[d]

[a]Institute of Computer Science of the Czech Academy of Sciences; [b]Faculty of Mathematics and Physics, Charles University; [c]Faculty of Education, Charles University; [d]IQVIA Belux

ABSTRACT

Many of the differential item functioning (DIF) detection methods rely on a principle of testing for DIF item by item, while considering the rest of the items or at least some of them being DIF-free. Computational algorithms of these DIF detection methods involve the selection of DIF-free items in an iterative procedure called *item purification*. Another aspect is the need to correct for multiple comparisons, which can be done with a number of existing *multiple comparison adjustment* methods. In this article, we demonstrate that implementation of these two controlling procedures together may have an impact on which items are detected as DIF items. We propose an iterative algorithm combining item purification and adjustment for multiple comparisons. Pleasant properties of the newly proposed algorithm are shown with a simulation study. The method is demonstrated on a real data example.

KEYWORDS

Differential item functioning; item purification; multiple comparison adjustments

Introduction

The Differential Item Functioning (DIF) is a well-known phenomenon that can arise in various contexts of multi-item scales, including psychological and educational measurement, admission tests, or health-related inventories (Martinková et al., 2017; Osterlind & Everson, 2009; Penfield & Camilli, 2007). An item is said to function differently (or, in short, to be a DIF item) when test takers from different subgroups of the target population, comparable on their level of ability or other underlying latent trait, have different probabilities of answering an item correctly or endorsing the item. DIF is a potential threat to fairness and validity of measurement and DIF analysis should, therefore, be a routine part of test validation. Many detection methods were developed to flag DIF items, using either score-based techniques or Item Response Theory (IRT) modeling, and these as well as new methods are still being studied intensively (Belzak, 2020; Berger & Tutz, 2016; Cho et al., 2016; Drabinová & Martinková, 2017; Hladká & Martinková, 2020; Magis et al., 2010; Maij-de Meij et al., 2010; Martinková & Hladká, 2023; Penfield et al., 2009; Schneider et al., 2021).

Most traditional DIF detection methods rely on the basic principle of testing for DIF one item after another, with the remaining items being considered as anchor (DIF-free) items. This process is known to have at least two drawbacks. First, when DIF items are truly present in the data, gradual DIF testing implies that DIF items are included in the matching variable (for instance the test score), which is known to be a source for a potentially serious bias and misidentification of DIF and non-DIF items (Jodoin & Gierl, 2001; Kopf et al., 2015a, 2015b; Woods, 2009). Second, testing each item one after another usually yields inflated type I error rates (i.e., proportion of falsely detected items) because traditional methods do not adjust for multiple comparisons involved in this repeated, item-by-item process.

Each issue was to some extent addressed in the DIF literature in different ways. To correctly identify a set of anchor items and to reduce the impact of DIF items on the matching variable, an *item purification* process was proposed (Lord, 1980); first suggested by Marco (1977) and later extended and improved by many authors including Candell and Drasgow (1988), Clauser et al. (1993), and French and Maller (2007). Item purification consists of the iterative removal of

items flagged as DIF from the set of anchor items Candell and Drasgow (1988). This algorithm was shown to improve the results of most DIF detection methods (Clauser et al., 1993; French & Maller, 2007; Navas-Ara & Gómez-Benito, 2002; Wang & Su, 2004), with the notable exception of Angoff's delta plot method (Magis & Facon, 2013). The other issue, inflated type I error rates due to multiple comparisons can, on the other hand, be accurately controlled with adequate *multiple comparison adjustment* procedures. Adjustments for multiple comparisons are easy to implement, non-iterative, and were also shown to improve the accuracy of DIF identification (i.e., non-inflated type I errors and larger power; see Kim & Oshima, 2013).

Though conceptually different and with different purposes, both item purification and adjustments for multiple comparisons share the same objective, that is, improvement for the classification of items into DIF and non-DIF groups. While both controlling procedures are still being studied intensively (Chen & Hwu, 2018; Fikis & Oshima, 2017; Khalid & Glas, 2014; Kim & Oshima, 2013), surprisingly, to the best of our knowledge, performance of these approaches has not yet been jointly evaluated in a comprehensive study, and, moreover, the various combinations have not yet been fully explored. This represents a potential gap in DIF literature, as both approaches have been shown to improve DIF detection to a certain extent.

In this work we propose an iterative combination of item purification and multiple comparison adjustment, and we evaluate their properties in a simulation study under various scenarios for three selected DIF detection methods: The Mantel-Haenszel test Mantel and Haenszel (1959), the logistic regression method Swaminathan and Rogers (1990), and the Simultaneous Item Bias Test (SIBTEST) method Shealy and Stout (1993). We finally offer a practical illustration using a real dataset from lower secondary education. The general goal of this paper is to assess the effect of controlling procedures in the improvement of DIF identification which includes the task of best identifying anchor (DIF-free) items and items truly affected by DIF.

The paper proceeds as follows: Section "Methods" introduces the proposed method and design of the simulation study, including the data generation process, the considered DIF detection methods and settings, and simulation evaluation. It also describes the real data example and the implementation of the methods in R. Section "Results" contains results of the simulation study separately for the three DIF detection methods considered, and provides results of the real data analysis. Section "Discussion" offers discussion and concluding remarks.

Methods

Controlling procedures in DIF detection

Item purification

DIF analysis is based upon the principle of comparing item performance of the test takers being matched by their ability. Therefore, defining an appropriate matching criterion is mandatory. For non-IRT DIF detection methods such as the Mantel-Haenszel test (Holland & Thayer, 1988; Mantel & Haenszel, 1959), the logistic regression method (Swaminathan & Rogers, 1990), or the SIBTEST method (Shealy & Stout, 1993), the total test score, i.e., the number of correct responses, is often used as the matching criterion. For IRT-based techniques such as the Lord's test Lord (1980), an estimate of their latent ability level is used instead.

The danger of computing this matching criterion for the set of administered items is that the inclusion of DIF items could seriously impact the results of the identification process. It is then of primary importance to ensure that anchor (i.e., DIF-free) items are available for proper computation of this matching variable. For non-IRT methods, the matching criterion (observed ability) should be computed by using only anchor items. For IRT-based methods, linking the two scales (one for the reference group and one for the focal group) should be based upon only these anchor items.

Because it is oftentimes impossible to predict which items will function differently, Candell and Drasgow (1988) proposed an iterative process that is currently referred to as item purification. In test-score-based DIF detection methods, item purification begins with one run of the DIF detection method per item, all other items being considered as anchor items. All items flagged as DIF are then removed from the set of anchor items, and the method is re-run using this reduced anchor set. These two steps (running DIF analysis and removing flagged items from the anchor set) are repeated until two successive runs yield the same set of items identified as functioning differently (Figure 1A).

To illustrate the item purification algorithm, let's assume an artificial test consisting of 10 items and an arbitrary non-IRT DIF detection method (Table 1). At the initial step, the total test score was calculated based upon all 10 items. Using a DIF detection method and the total test score, items 1, 7, and 8 were detected as DIF items. In the first step of item purification, these items were removed from the calculation of the total score and a DIF detection procedure was then applied

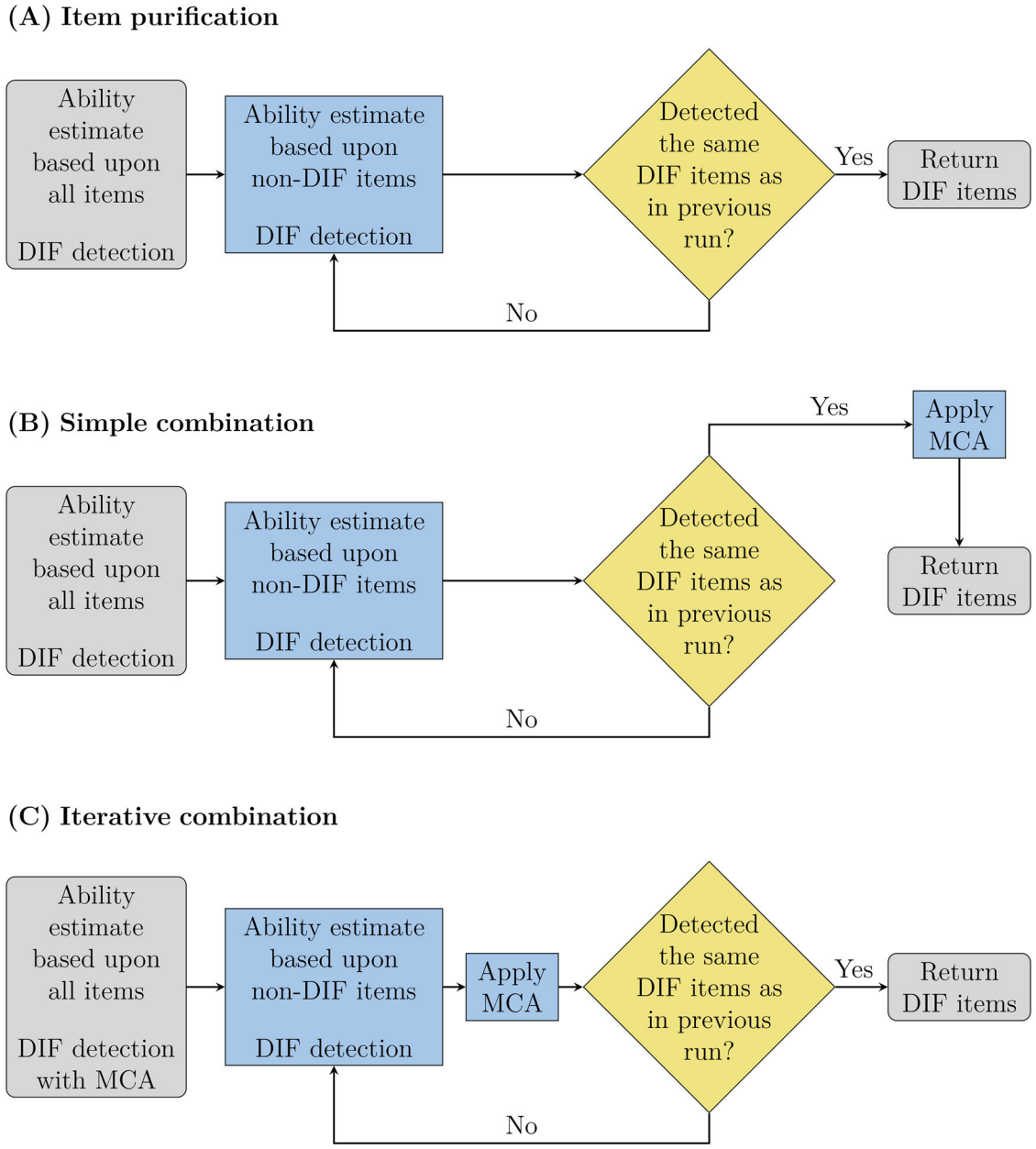

Figure 1. Scheme of (A) Item purification, (B) Simple combination of item purification and multiple comparison adjustment (MCA), (C) Iterative combination of item purification and MCA.

Table 1. Illustration of the item purification algorithm.

Matching criterion	$\sum_{i=1}^{10} Y_i$	$\sum_{i\neq\{1,7,8\}} Y_i$	$\sum_{i\neq\{1,8\}} Y_i$	$\sum_{i\neq\{1,2,8\}} Y_i$
Item	Step 1	Step 2	Step 3	Step 4
1	DIF	DIF	DIF	DIF
2	NON-DIF	NON-DIF	DIF	DIF
3	NON-DIF	NON-DIF	NON-DIF	NON-DIF
4	NON-DIF	NON-DIF	NON-DIF	NON-DIF
5	NON-DIF	NON-DIF	NON-DIF	NON-DIF
6	NON-DIF	NON-DIF	NON-DIF	NON-DIF
7	DIF	NON-DIF	NON-DIF	NON-DIF
8	DIF	DIF	DIF	DIF
9	NON-DIF	NON-DIF	NON-DIF	NON-DIF
10	NON-DIF	NON-DIF	NON-DIF	NON-DIF

using this new matching criterion (total test score without items 1, 7, and 8). In the second step, only items 1 and 8 were detected as functioning differently. The set of DIF items was not the same as in the previous iteration and thus the matching criterion (total test score without items 1 and 8) was recalculated and the DIF detection procedure was run again. In the third step, items 1, 2, and 8 were detected as DIF. Again, current and previous sets of DIF items were not the same and the matching criterion needed to be calculated was once more the total test score without items 1, 2, and 8. Finally, in the fourth step, items 1, 2, and 8 were detected as in the previous iteration and the algorithm stopped.

Item purification is an approach which is intuitively appealing and simple to implement. Though item purification can be done efficiently in most cases,

Table 2. Illustration of the Holm's and BH adjustments for multiple comparisons in DIF detection in comparison with standard decision not using adjustment for multiple comparisons.

Item	Rank	p-value	Standard decision	Holm's boundary	Holm's decision	BH boundary	BH decision
5	1	0.0014	DIF	0.0050	DIF	0.0050	DIF
10	2	0.0039	DIF	0.0056	DIF	0.0100	DIF
9	3	0.0111	DIF	0.0062	NON-DIF	0.0150	DIF
8	4	0.0182	DIF	0.0071	NON-DIF	0.0200	DIF
3	5	0.0209	DIF	0.0083	NON-DIF	0.0250	DIF
6	6	0.0306	DIF	0.0100	NON-DIF	0.0300	NON-DIF
2	7	0.0388	DIF	0.0125	NON-DIF	0.0350	NON-DIF
4	8	0.2430	NON-DIF	0.0167	NON-DIF	0.0400	NON-DIF
7	9	0.3623	NON-DIF	0.0250	NON-DIF	0.0450	NON-DIF
1	10	0.7826	NON-DIF	0.0500	NON-DIF	0.0500	NON-DIF

it can sometimes become time consuming (especially for IRT-based methods), and there is no guarantee that the iterative process will converge (i.e., will provide two successive identical sets of DIF items).

Multiple comparisons adjustments

Another potentially problematic issue, often present in DIF detection though less often investigated, is that each item is being tested individually, while all other items are considered free of DIF. This implies that multiple comparisons among all test items will arise, which is without adjustment to the significance level known to lead to inflated type I error rates. In the DIF framework, Kim and Oshima (2013) compared different methods of adjustment for multiple comparisons. Two such adjustment procedures were shown to be superior in the DIF context: Holm's procedure Holm (1979) and Benjamini-Hochberg (BH) procedure Benjamini and Hochberg (1995). Holm's procedure Holm (1979) is an improvement to Bonferroni's procedure which was shown to be more powerful Holland and Copenhaver (1988). It is intended to control family-wise error, that is, the probability of making one or more type I errors (in a DIF context, the type I error means flagging a non-DIF item as DIF). BH procedure controls the false discovery rate, that is, the expected proportion of type I errors Benjamini and Hochberg (1995). Procedures to control the false discovery rate have greater power at the cost of increased type I error rates (Shaffer, 1995).

These two adjustment methods can be schematically described as follows: First, for each tested item i (say from 1 to I), let p_i be the corresponding p-value for the DIF detection method (obtained when all other items are set as anchor items), and let $p_{(1)}, \ldots, p_{(I)}$ be the I values sorted in increasing order. Then, for a given global significance level α, the index k is defined as

1. the minimal index that satisfies $p_{(k)} > \alpha/(I+1-k)$ for Holm's procedure,
2. the maximal index that satisfies $p_{(k)} \leq \alpha k/I$ for BH procedure.

Eventually, items with corresponding ordered p-values $p_{(1)}$ to $p_{(k-1)}$ (for Holm's procedure) or to $p_{(k)}$ (for BH procedure) are flagged as DIF, while the remaining items are considered non-DIF.

These methods are illustrated using an artificial example of ten items, highlighting how the choice of an adjustment method has a straightforward impact upon which item is detected as a DIF item (Table 2). The Holm's and BH boundaries were calculated by formulae in (1) and (2), and then compared with ordered p-values. With Holm's procedure, index k was equal to three; therefore only the first two listed items (i.e., items 5 and 10) were eventually flagged as DIF. This is a considerable reduction when compared to the original classification (without Holm's adjustment) which led to flagging seven out of the ten items as DIF. With BH procedure, k index was equal to five; therefore the first five items (according to their classification in an increased order of p-values) were flagged as DIF, compared to seven items when no adjustment was considered.

Simple and iterative combination of controlling procedures

We will introduce two different settings for applying both methods together, item purification and multiple comparisons adjustments. We consider first a *simple* combination of both approaches, containing the full item purification process followed by a single multiple comparison adjustment (Figure 1B). This combination of controlling procedures is easily applicable using software providing DIF analysis with an implemented item purification algorithm.

We then propose an *iterative* algorithm which performs item purification followed by an adjustment for a multiple comparison after each run of item purification (Figure 1C). We expect that this algorithm may be more precise and also less time consuming thanks

to effectively lowering the number of DIF items by a computationally undemanding adjustment procedure, already present in the initial run.

Simulation study

Data generation

Six design factors were manipulated to generate the data: (a) sample size, (b) test length, (c) amount of DIF items, (d) type of DIF, (e) size of DIF effect, and (f) distribution of ability for the focal group. The total sample sizes 250 (125 per group), 500 (250 per group), 1,000 (500 per groups), and 2,000 (1,000 per group) were selected, while test lengths of 20, 40, and 80 items were considered. Four different proportions of DIF items (0%, 5%, 15%, and 30%) were considered. Parameters of DIF items were chosen to incorporate both types of DIF (uniform and non-uniform) in two different sizes of DIF effect (0.4 and 0.8) quantified by the area between the characteristic curves (Raju, 1988). DIF effect sizes correspond to small and large DIF magnitudes and were selected following Swaminathan and Rogers (1990) and Narayanan and Swaminathan (1996).

The item responses were generated under a true underlying three-parameter logistic IRT model. In all scenarios, the ability of the reference group was drawn from the standard normal distribution. For the focal group we considered three options for determining ability levels. First, ability levels were the same as those for the reference group—drawn from a standard normal distribution. Second, ability levels for the focal group were drawn from a normal distribution with a mean equal to 1 but with the same standard deviation as used for the reference group. Third, the standard deviation for a normal distribution for the focal group was manipulated and set to 1.5.

Parameters of non-DIF items were selected from problem solving of the Graduate Management Admission Test (Kingston et al., 1985, see Table at p. 47 for parameters according to 3PL IRT model for all 80 non-DIF items) to reflect realistic values. When tests of 20 or 40 items were considered, only the set of parameters for the first 20 or 40 items were used.

The DIF item parameters creation was inspired by Narayanan and Swaminathan (1996). The *c*-parameter was fixed at a value of 0 for all DIF items. The selection of discrimination and difficulty parameter values depends upon the type of DIF effect generated. For uniform DIF, discrimination parameter a was fixed for both groups, and difficulty parameter b varied to gain the desired DIF effect size (either small 0.4, or large 0.8). 24 uniform DIF items were simulated with varying values of b parameter—$b = -1$, $b = 0$, or $b = 1$ for the reference group, and an appropriate shift for the focal group—either 0.4 or 0.8 to perform either a small or large DIF effect size; and varying values of a parameter—from $a = 0.25$ to $a = 2.25$. Table A1 summarizes all of these options and highlights which non-DIF item(s) were replaced by those parameter values.

For non-uniform DIF items, difficulty parameter b was fixed for both groups, and discrimination parameter a varied to gain the desired DIF effect size. 24 non-uniform DIF items were simulated with varying values of the common b parameter—low ($b = -1$), medium ($b = 0$), and high ($b = 1$) and varying values of a parameter—from $a = 0.50$ to $a = 0.85$ for the reference group, and an appropriate shift for the focal group—from 0.20 to 0.82 to perform either a small or large DIF effect size. These combinations are also listed in Table A1.

This simulation design yields 36 settings in the absence of DIF (four sample sizes, three test lengths, and three ability distributions) and 432 settings in the presence of DIF (in addition, three proportions of DIF, two DIF sizes, and two types of DIF effect), thus 468 design settings in total. 1,000 datasets were generated for each setting. Note that given the fact that some DIF detection methods may yield convergence issues, no results were obtained for items in which the algorithm failed to converge and thusly no conclusion about DIF detection could be drawn. To overcome this problem, runs with convergence issues were excluded, and simulations were re-run until 1,000 replications without convergence failures were obtained.

DIF detection

Three non-IRT DIF detection methods were selected: the Mantel-Haenszel test (Holland & Thayer, 1988; Mantel & Haenszel, 1959), the logistic regression procedure (Swaminathan & Rogers, 1990) with the likelihood ratio test accounting for both types of DIF (i.e., uniform and non-uniform), and the SIBTEST (Shealy & Stout, 1993). These three methods were chosen because they are the most commonly used non-IRT procedures for identifying DIF. Moreover, they might benefit the most from multiple comparison adjustments and their combinations since DIF detection is done item by item in contrast to IRT-based methods.

All three DIF detection methods were employed for each generated dataset, together with eight possible procedures to control type I error: (a) item purification, (b) Holm's adjustment method, (c) BH

adjustment method, (d) simple combination of item purification with Holm's adjustment, (e) simple combination of item purification with BH adjustment, (f) iterative algorithm with Holm's adjustment, (g) iterative algorithm with BH adjustment, and (h) no controlling procedure (i.e., the DIF detection method performed without any further controlling procedure; for bench-marking purposes). 24 combinations of DIF detection methods and type I error controlling procedures were applied for each dataset. In a case of using item purification, either alone or in combination with an adjustment method, a maximal number of iterations was set to 50. The significance value was set to 5%.

Summary statistics and simulation evaluation

Three summary statistics (type I error rate, rejection rate, and power rate) were computed across the 1,000 generated datasets per study design, and separately for each of the 24 combinations of DIF detection method and controlling procedures. A type I error rate was estimated as the proportion of falsely detected items when none of the items were generated as DIF. The rejection rate was calculated as the proportion of falsely detected items among all non-DIF items (in the cases when DIF items were present in the simulation scenario). The power rate was calculated as the proportion of correctly detected DIF items among all truly DIF items.

The results were interpreted with respect to the following research questions:

1. Are the DIF detection methods able to control for type I error (i.e., type I error and rejection rates close to the 5% significant level) with sufficient power (i.e., over 80%) even without any controlling procedure?
2. How do the studied controlling procedures (item purification, Holm's adjustment, and BH adjustment) and their combinations (simple combination of item purification with Holm's adjustment, simple combination of item purification with BH adjustment, iterative algorithm with Holm's adjustment, iterative algorithm with BH adjustment) compare in different scenarios in terms of power?
3. Which design factors have significant impact on type I error rate, rejection rates, and power rates?

The first question has been investigated by many authors (see, e.g., van de Water, 2014). In the context of this simulation study, the answer to the first question could help to set the bench-marking values to which other methods are being compared.

To get an initial idea, summarizing figures with observed type I error, rejection rates, and power rates were produced. For simplicity, presented values were averaged by scenarios with the same level of a given factor. Type I error and rejection rates were considered as suitable if they were close to the 5% significance level. Power rates were considered as satisfactory if they achieved a value of at least 80%.

To test for significance in the differences between controlling procedures and an effect of other study factors, beta regression models for type I error, rejection rates, and power rates were fitted with a logit link to cover all values between 0 and 1. All possible double interactions between factors were included into the models. Note that since the beta regression model cannot handle extreme values (i.e., 0 or 1), such type I error rates, rejection rates, and power rates were replaced by values 10^{-6} higher or lower. Interpretation of the parameter effects in the beta regression model is the same as in logistic regression (e.g., Agresti, 2002). It should be noted that interpretation of the results was made primarily with focus on controlling procedures and their possible interactions with other factors (see research questions above). Since we were not interested in the differences between DIF detection methods in this study, three separate models were fitted, one for each method.

Real data example

To demonstrate the impact of the choice for the controlling procedures on DIF detection in practice, we analyze data from the Czech Longitudinal Study in Education (CLoSE) (Greger et al., 2022; Martinková et al., 2020). We focused on the results in a test of reading skills taken in the 6th grade. Two versions of the same test were distributed. We have considered here only version B.

A total of 2,634 students participated in this test, including 1,310 girls and 1,324 boys. The test was comprised of 19 items, some of which were multiple-choice and some were open answer questions. For purpose of this paper, the item responses were dichotomized: 1 point was awarded if the answer was fully correct and 0 if it was not. DIF was investigated across gender using three DIF detection methods and eight scenarios of controlling procedures compared in the simulation study.

Practical implementation

For all analyses, software **R**, version 3.6 (R Core Team, 2019) was used. All DIF detection methods and controlling procedures were fitted using the **difR** package (Magis et al., 2010). The proposed iterative combination of item purification and multiple comparison adjustment was implemented as an extension to difMH(), difLogistic(), and difSIBTEST() functions. Functionalities of the extended functions may also be explored in an interactive application of the **ShinyItemAnalysis** package, version 1.5.0 (Martinková & Drabinová, 2018; Martinková & Hladká, 2023); see also Figure A1. Beta regression models for summary statistics in the simulation study were fitted using the **betareg** package (Cribari-Neto & Zeileis, 2010). The selected **R** codes for browsing the results of the simulation study and for the real-data analysis together with the datasets are provided in the electronic supplemental files available at https://osf.io/jng7y/.

Results

Simulation study

Mantel-Haenszel test

Empirical rates. For small sample sizes, the Mantel-Haenszel test was able to control type I error and rejection rates under all choices of the controlling procedures and in almost all scenarios (Figure 2, top and middle panels). It is a common phenomenon that with an increasing sample size the rejection rate increases which could also be observed for the Mantel-Haenszel test here. When using item purification, the proportion of cases when rejection rate exceeded a significance level of 0.05 was lower than when using no controlling procedure and, moreover, the mean rejection rate remained near the significance level even for large sample sizes (Figure 2, middle panel). On the other hand, as expected, item purification had no effect on type I error rate (Figure 2, top panel). The multiple comparison adjustments and their combinations with item purification yielded rejection rates under the significance level in most of the scenarios. However, when using only the multiple comparison adjustments (without item purification), there was an increase in the proportion of scenarios with rejection rates exceeding 0.15 more often than in scenarios with item purification only. In such cases, BH adjustment yielded an even larger mean rejection rate than item purification (Figure 2, middle panel).

For small sample sizes, there was only a small proportion of scenarios when power was sufficient and mean power of the Mantel-Haenszel using any of the options for controlling procedures remained at a low level. However, power rates were generally increasing with an increasing sample size. While the multiple comparison adjustments and their combinations with item purification gained lower power rates than item purification alone or when using no controlling procedure, this difference was somehow softened when the sample size was large. Item purification seemed to gain the largest power, followed by a scenario of using no controlling procedure and then by the simple combination of BH adjustment and item purification (Figure 2, bottom panel).

Using item purification alone, the mean number of iterations of item purification was decreasing with the increasing sample size and increasing with a larger proportion of DIF items. The mean number of iterations varied from 5.95 to 18.94. As expected, the iterative combinations of item purification and the adjustments for multiple comparison generally yielded a lower mean number of iterations (varied from 0.21 to 2.82 for Holm's adjustment and from 0.39 to 4.33 for BH adjustment). While the effect of increasing the proportion of DIF items was similar to that for item purification, the number of iterations increased with the increasing sample size.

Beta regression model. A beta regression model confirmed increasing rejection and power rates with the increasing sample size. While there was no significant effect of item purification in the power with the increasing sample size, this method significantly improved control of rejection rates compared to the scenario using no correction. The finding was also suggested by the empirical rates (displayed in Figure 2, middle panel, and discussed above). Furthermore, item purification improved control of rejection rates when there were a large amount of DIF items and when the underlying DIF magnitude was large. When there was a large proportion of DIF items, item purification also significantly, but only slightly, increased power rate. Generally, using a multiple comparison adjustment led to a substantial decrease in power which significantly improved with an increased sample size. This was also accompanied by a significant decrease in rejection rates, which was somehow softened by an increased sample size when using BH multiple comparison adjustment alone. All multiple comparison adjustments and their combinations with item purification yielded lower values in all three

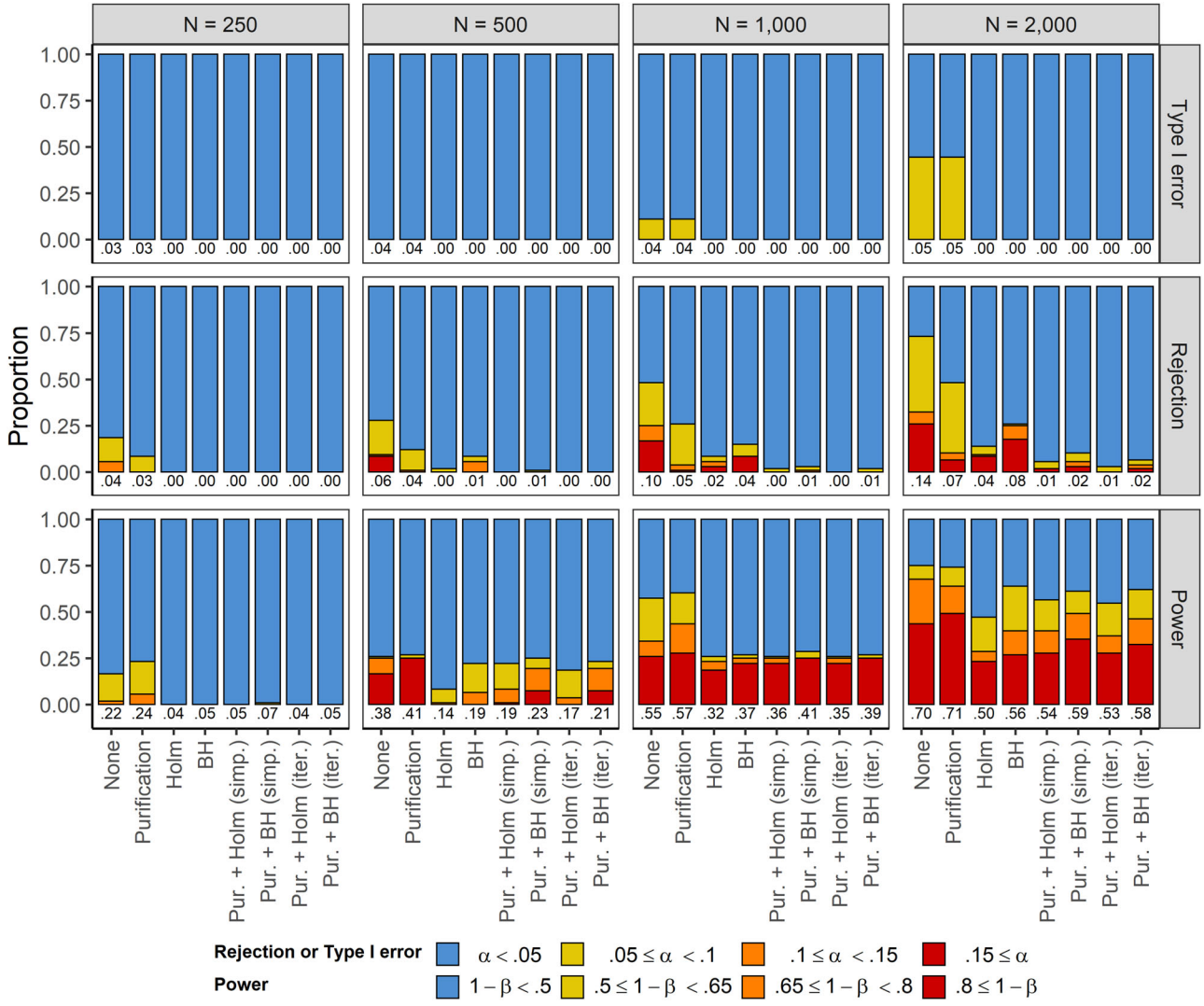

Figure 2. Empirical type I error rates, empirical rejection rates, and empirical power rates for the Mantel-Haenszel test. Plot shows proportions (among 1,000 data sets) of 4 levels of type I error, rejection, or power rates within given controlling procedure and sample size. Values below the bars indicate mean rates.

summary statistics (type I error, rejection rate, and power) when considering a test consisting of 40 or 80 items (Figure 3, Table A2).

Logistic regression method

Empirical rates. When using no controlling procedure or item purification alone, there was a large proportion of scenarios slightly exceeding the significance level of 0.05, i.e., type I error and rejection rates varied mostly between 0.05 and 0.1. In both cases, a proportion of severe overrun of the significance level increased for large sample sizes. Also a mean value of type I error and rejection rates exceeded the significance level of 0.05 especially for large sample sizes (Figure 4, top and middle panels). Item purification exhibited lower control of type I error especially for larger sample sizes which resulted in slightly increased rejection rates compared to a case of using no controlling procedure (Figure 4, top panel). This was also the case in presence of DIF items. However, item purification yielded a larger proportion of scenarios with good control of the rejection rate at the same time (Figure 4, middle panel). All multiple comparison adjustments and their combinations with item purification were able to control for type I error. However, both simple combinations showed increased proportions of severe overrun for a large sample size and thus increased mean type I error and rejection rates (Figure 4, top and middle panel).

None of the controlling procedures were able to gain sufficient power for small sample sizes. However, power rates were increasing with the increasing sample size, while item purification yielded the largest proportion of scenarios with sufficient power (i.e., at

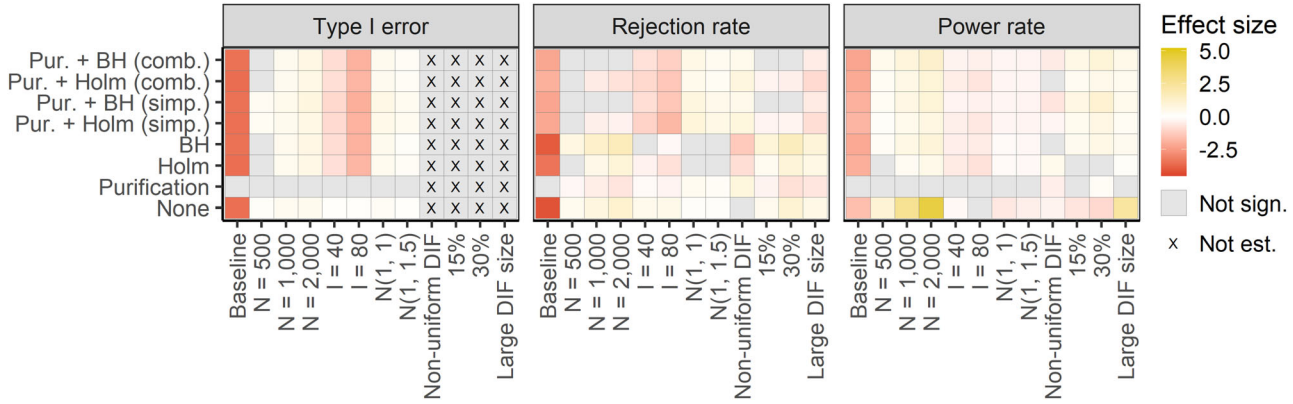

Figure 3. Beta regression coefficients for controlling procedures (in rows) and their interaction with other factors (in columns) on type I error, rejection, and power rates for the Mantel-Haenszel test.

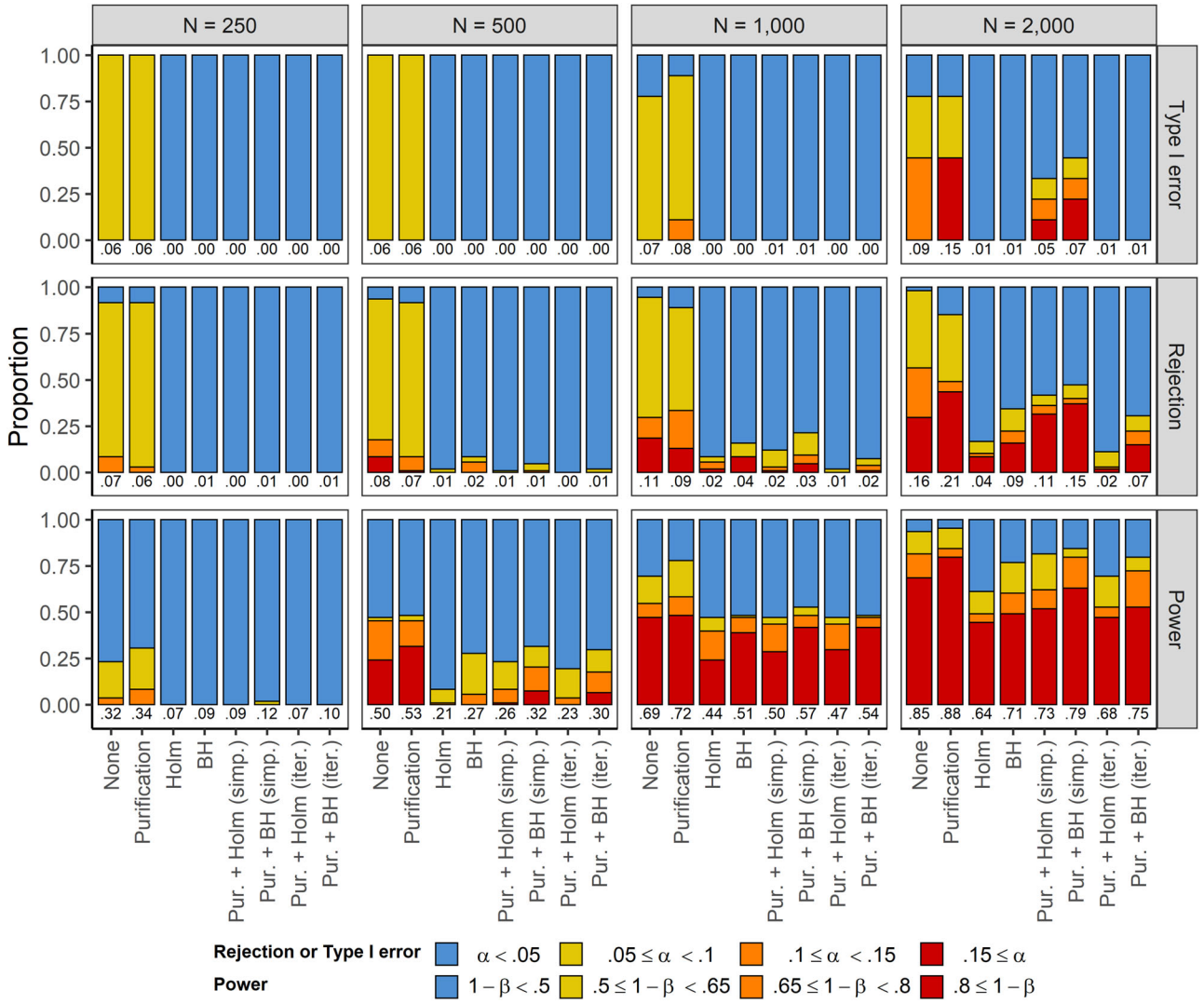

Figure 4. Empirical type I error rates, empirical rejection rates, and empirical power rates for the logistic regression method. Plot shows proportions (among 1,000 data sets) of 4 levels of type I error, rejection, or power rates within given controlling procedure and sample size. Values below the bars indicate mean rates.

least 80%), followed by a setting with no controlling procedure and both combinations of BH adjustment and item purification (Figure 4, bottom panel).

The mean number of iterations for item purification was increasing with the increasing sample size and with the increasing proportion of DIF items in all

settings of controlling procedures. However, as expected, item purification alone (and its simple combinations) yielded a larger mean number of iterations (varied from 2.46 to 9.41) than the proposed iterative combination with multiple comparison adjustments (varied from 0.23 to 2.77 for Holm's adjustment and from 0.36 to 5.72 for BH adjustment).

Beta regression model. Similarly to the Mantel-Haenszel test, item purification in the logistic regression method improved rejection rate control in the case of a large DIF effect size and a large proportion of DIF items. However, unlike in the Mantel-Haenszel test, the power of the logistic regression DIF detection method increased when the sample size increased. On the other hand, the power decreased slightly when the DIF was non-uniform and control of rejection rates worsened when the latent trait of focal groups was drawn from a normal distribution with a different mean and variance. Multiple comparison adjustments and their combinations with item purification generally indicated lower rejection and power rates. While there were no crucial differences between the controlling procedures in terms of power and their interactions with other factors, their control for rejection rates differed. Especially, the iterative combination of item purification with multiple comparison adjustments, either BH or Holm's, demonstrated better control when the sample size increased (Figure 5, Table A3).

SIBTEST

Empirical rates. All controlling procedures improved or at least did not worsen their control of type I error and rejection rates, even in the case of a large sample size, when increased values were observed using no controlling procedure. While item purification and BH adjustment itself slightly overran the significance level, all four combinations of multiple comparison adjustments and item purification kept rejection rates under the significance level for almost all scenarios (Figure 6, top and middle panels).

Sufficient power was gained only for larger sample sizes regardless of whatever controlling procedure was used. In these cases, item purification performed slightly better than other procedures, followed by a scenario using no controlling procedure and then by both combinations, either simple or iterative, of BH multiple comparison adjustment and item purification (Figure 6, bottom panel).

Similar to the logistic regression method, the mean number of iterations for item purification in the SIBTEST method was increasing with the increasing sample size and with the increasing proportion of DIF items in all settings of controlling procedures. Item purification alone (and its simple combinations) yielded again, as expected, a larger mean number of iterations (varied from 8.67 to 25.42) than the proposed iterative combination of item purification and the multiple comparison adjustments (varied between 0.24 and 6.67 for Holm's adjustment and between 0.38 and 9.91 for BH adjustment).

Beta regression model. Analogously to previous DIF detection methods, the SIBTEST also showed increasing power with the increasing sample size. However, it seemed that the SIBTEST struggled when a number of items increased, since the power and rejection rates decreased rapidly. Moreover, it was somehow more difficult to identify DIF when a larger proportion of DIF items were present. It was slightly better when the proposed combination of item purification with BH multiple comparison adjustment (either simple or iterative) was applied. Item purification once again demonstrated better control of the rejection rate in the case of a large proportion of DIF items, large DIF

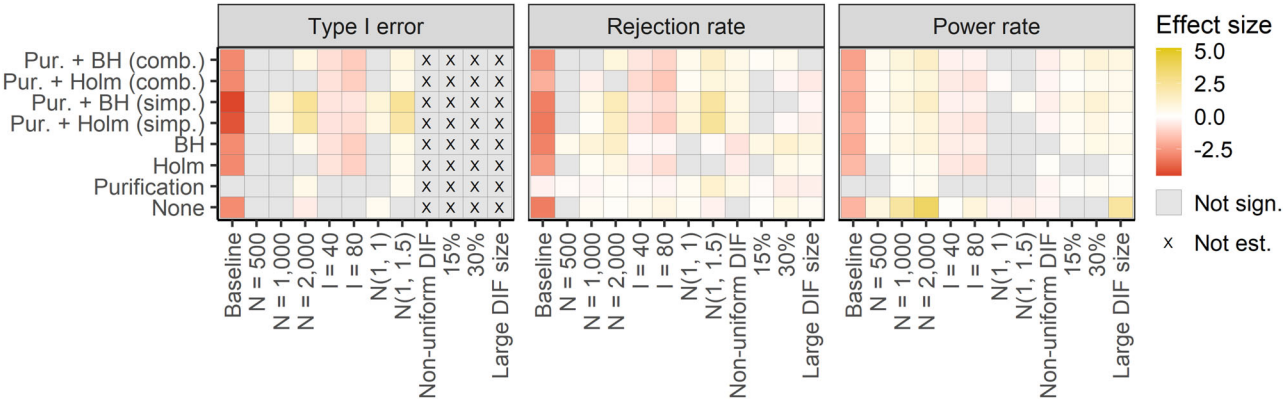

Figure 5. Beta regression coefficients for controlling procedures (in rows) and their interaction with other factors (in columns) on type I error, rejection, and power rates for the logistic regression method.

Table 3. Items from the CLoSE dataset flagged as DIF items by at least one DIF detection method.

Item	1	3	4	8	10	11	13	14	
Mantel-Haenszel test									
None	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	42%
Purification	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Holm	DIF	DIF	DIF	NON-DIF	DIF	DIF	DIF	NON-DIF	32%
BH	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	42%
Pur. + Holm (simple)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	DIF	NON-DIF	26%
Pur. + BH (simple)	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Pur. + Holm (iterative)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	NON-DIF	NON-DIF	21%
Pur. + BH (iterative)	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Logistic regression method									
None	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	42%
Purification	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Holm	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	DIF	NON-DIF	26%
BH	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Pur. + Holm (simple)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	DIF	NON-DIF	26%
Pur. + BH (simple)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	DIF	NON-DIF	26%
Pur. + Holm (iterative)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	NON-DIF	NON-DIF	21%
Pur. + BH (iterative)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	DIF	NON-DIF	26%
SIBTEST									
None	DIF	DIF	DIF	DIF	DIF	DIF	DIF	DIF	42%
Purification	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Holm	DIF	DIF	DIF	NON-DIF	NON-DIF	DIF	DIF	NON-DIF	26%
BH	DIF	DIF	DIF	DIF	DIF	DIF	DIF	NON-DIF	37%
Pur. + Holm (simple)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	NON-DIF	NON-DIF	21%
Pur. + BH (simple)	DIF	DIF	DIF	DIF	NON-DIF	DIF	DIF	NON-DIF	32%
Pur. + Holm (iterative)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	NON-DIF	NON-DIF	21%
Pur. + BH (iterative)	DIF	NON-DIF	DIF	DIF	NON-DIF	DIF	NON-DIF	NON-DIF	21%
Perc. DIF	100%	58%	100%	92%	50%	100%	79%	17%	

effect size, and increased sample size. However, its effect on power was limited (Figure 7, Table A4).

In summary, all three considered DIF detection methods when applied without adjustment for multiple comparisons led to rejection rates somewhat exceeding the nominal significance level, especially in scenarios with large sample sizes. Adjustments for multiple comparisons helped to reduce both the rejection and the type I error rates and in some scenarios, the algorithms combining multiple comparison adjustments and item purification were even more successful in this reduction.

However, the reduction of rejection and type I error rates achieved by the multiple comparison adjustments were also accompanied by a decrease in power, which was especially visible in small sample sizes. In some scenarios and especially for the logistic regression DIF detection method, the algorithms combining multiple comparison adjustments and item purification led to higher power rates than multiple comparison adjustments alone.

The proposed iterative combination of item purification and the multiple comparison adjustments required a smaller mean number of iterations than item purification alone. This was partly due to multiple comparison adjustments lowering the number of detected DIF items, thus increasing the number of cases when no DIF item was identified in the initial run.

Real data example

In the CLoSE reading dataset, items 1, 3, 4, 8, 10, 11, 13, and 14 were flagged as DIF with respect to gender of the respondents by at least one DIF detection method (Table 3). Based upon results of the logistic regression method, boys were favored in items 1 and 4 where guessing was possible (Figure A2a), while girls were favored in items 3, 10, 13, and 14, where students were supposed to demonstrate reasoning or to describe their feelings. Items 8 and 11 displayed a non-uniform DIF. Item 8 favored boys with lower levels of ability and girls with higher levels of ability, contrarily, item 11 favored girls with lower levels of ability and boys with higher levels of ability (Figure A2b).

While items 1, 4, and 11 were flagged by all DIF detection methods with any or no controlling procedure applied, many were flagged under only some of the scenarios: For example, item 14 was only flagged when no controlling procedure was applied and any DIF detection method was used, or by the Mantel-Haenszel test with BH adjustment.

For all three DIF detection methods, the proportion of items flagged as DIF was the highest when no controlling procedure was applied, followed by BH adjustment, and then by item purification only. In the case of this dataset and for all three DIF detection methods, item purification yielded very similar results to a case where no controlling

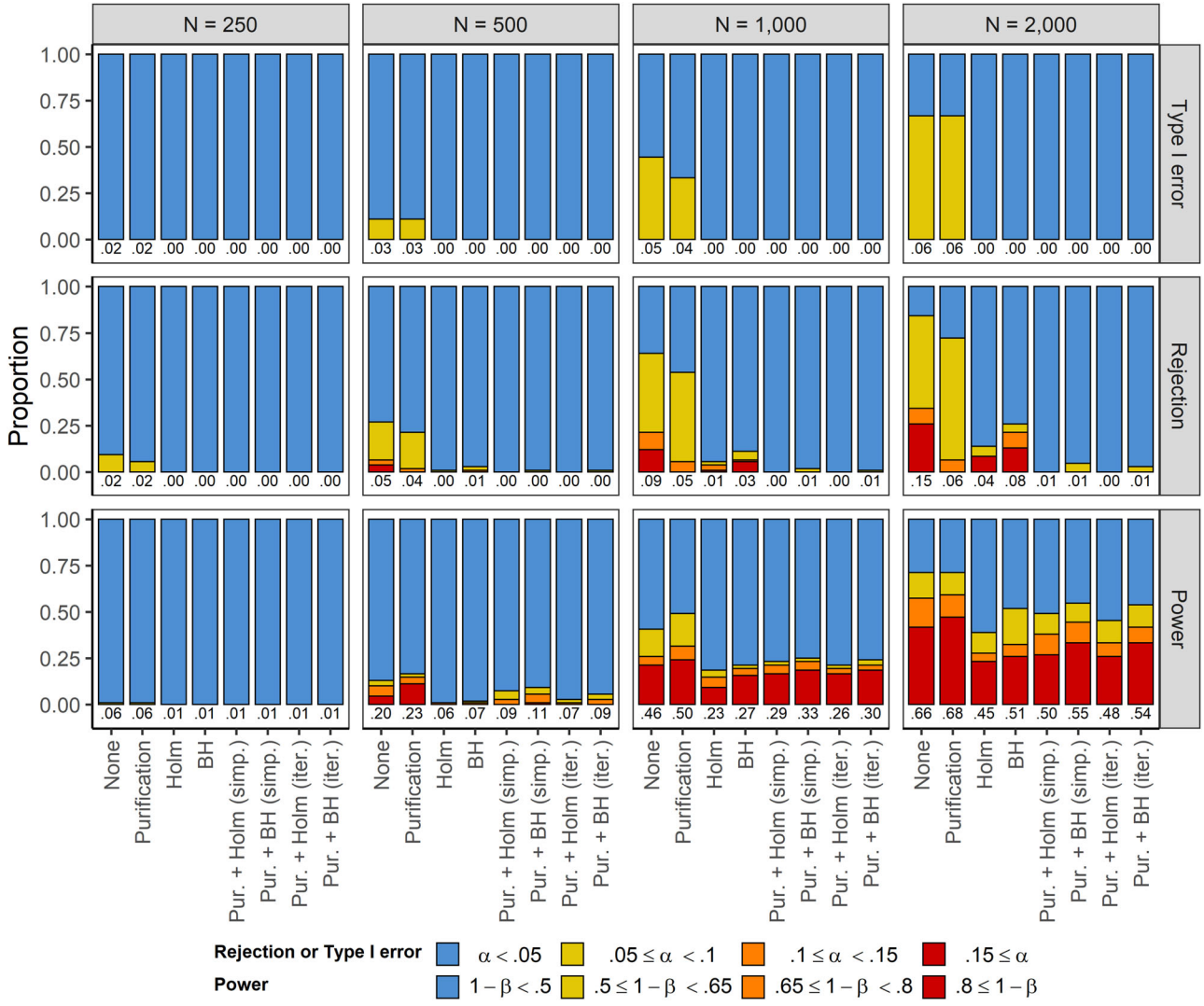

Figure 6. Empirical type I error rates, empirical rejection rates, and empirical power rates for the SIBTEST method. Plot shows proportions (among 1,000 data sets) of 4 levels of type I error, rejection, or power rates within given controlling procedure and sample size. Values below the bars indicate mean rates.

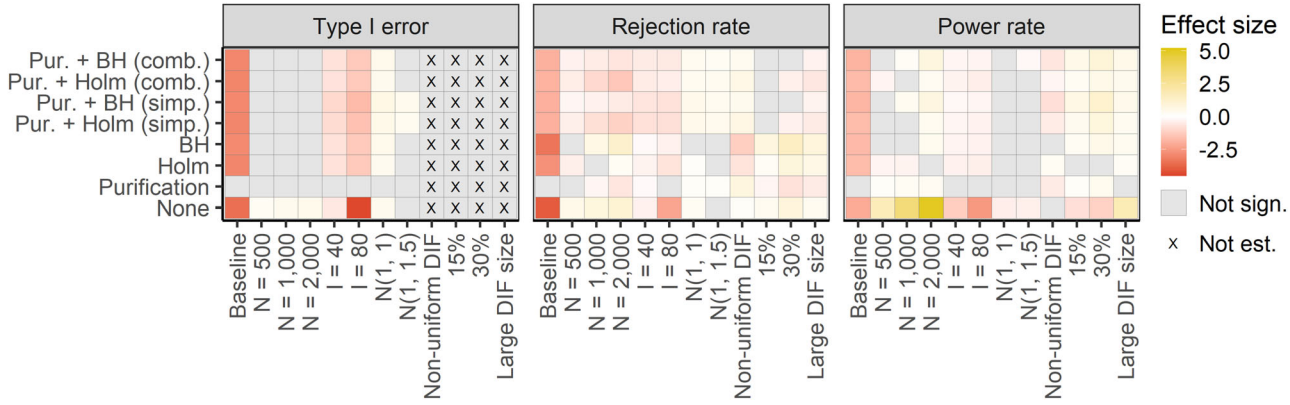

Figure 7. Beta regression coefficients for controlling procedures (in rows) and their interaction with other factors (in columns) on type I error, rejection, and power rates for the SIBTEST.

procedure was applied; the only difference could be seen in item 14, whereby DIF was not detected with item purification.

The effect of some choices for controlling procedures (none, item purification, and the simple combination of item purification and Holm's adjustment)

was consistent across all DIF detection methods, however, this was not the case when other choices for controlling procedures were considered. Especially, when using Holm's adjustment and either the simple or iterative combination of item purification and BH adjustment led to different items being flagged as DIF when applying different DIF detection methods. For example, with Holm's adjustment, items 8 and 14 were not flagged by the Mantel-Haenszel test, and the SIBTEST method did not flag item 10. On the other hand, the logistic regression method flagged item 3, but not item 10.

As expected, the real data example demonstrates that a choice in controlling procedures has an impact on which items are detected as DIF items. Some trends observed in the real data example are consistent with the results provided by the simulation study, however, we are reminded that in a case using real data examples, we can only speculate which controlling procedure provides the most precise DIF detection, since the real item parameters are not known.

Discussion

This work studied controlling procedures for DIF detection in cases where DIF is being analyzed item by item, namely item purification and adjustments for multiple comparisons. We proposed an iterative procedure in which adjustment for multiple comparisons was applied after each step of item purification. With empirical examples, we confirmed that, depending upon the controlling procedures applied, DIF detection may provide differing results. We performed a complex simulation study designed to evaluate the impact of controlling procedures on three DIF detection procedures (the Mantel-Haenszel test, the logistic regression method, and the SIBTEST method), specifically for their type I error, rejection, and power rates. To evaluate results of the simulation study, we used summary statistics and beta regression models.

In general, the results suggested that all three DIF detection methods, when applied item by item and without any adjustment for multiple comparisons, lead to rejection rates which somewhat exceeded the nominal significance level in some scenarios, especially those with large sample sizes. This issue has already been noted by many authors including DeMars (2009), Güler and Penfield (2009), and Herrera and Gómez (2008). Adjustments helped to reduce both the rejection rates and the type I error, however, this reduction was also accompanied by a decrease in power. Kim and Oshima (2013) previously noted that these adjustments caused a decrease in power to some extent, but this study established that in some scenarios the power rates are no longer sufficient. This means, in general, fewer items are detected as DIF, and some potentially unfair items may remain undetected.

The effect of item purification has been formerly researched by many authors including Candell and Drasgow (1988), Navas-Ara and Gómez-Benito (2002), and Wang and Su (2004). In our study, we confirmed some improvement in DIF detection when using item purification, primarily in the Mantel-Haenszel test and the SIBTEST method. Generally, we observed an improvement in DIF detection when a larger proportion of DIF items were present which was also shown for example by French and Maller (2007). However, item purification yielded increased type I error and rejection rates when applied within the logistic regression method, especially when a large sample size was considered, meaning that more items are detected as DIF and need to be assessed by content experts, which may give a false impression of suspicious test items behavior.

Benefits of the iterative algorithm combining item purification and multiple comparison adjustments were considerable in several scenarios and aspects. In most scenarios, and for all DIF detection methods, both the simple and iterative combination of item purification and multiple adjustments improved control of type I error and rejection rates, often to a greater extent than when applying the adjustment methods alone. The newly proposed iterative algorithm was superior to the simple combination of item purification and adjustment methods found in the logistic regression DIF detection method. This was especially visible in large sample sizes. Multiple comparisons adjustments generally decreased the power of all studied DIF detection methods, however, this appeared to be to a lesser amount when applied in combination with item purification. Moreover, the proposed iterative combination algorithm considerably lowered a number of iteration steps when compared to item purification alone. In spite of the demonstrated benefit of using the iterative algorithm, when creating an anchor item set, a greater concern might be a type II error (identifying an item as non-DIF when it is actually a DIF item; see, e.g., Edelen et al., 2006). Therefore, it might feel more comfortable being on the safe side by using the simple combination, since it uses an appropriate anchor item set to calculate the matching criterion.

By using a real data example from the CLoSE study, possible differences in DIF detection when applying different controlling procedures were demonstrated. This real data set analysis seems to confirm results of the simulation study by detecting the highest percentage of DIF items when no controlling procedure is applied, followed by item purification, and BH adjustment. These three choices of controlling procedures also yielded the largest rejection rate and the first two also provided the largest power rate in almost all scenarios. However, the precision of the different controlling procedures cannot be inferred from a real data example, since the true item parameters remain unknown.

There are some limitations to this simulation study which need to be considered. First, only a limited number of DIF detection methods were used. We decided to include classical methods in which DIF detection is traditionally done item by item, a method which we believed would benefit most from suggested controlling procedures. A simulation study showed that various choices for controlling procedures have different effects on DIF detection methods, thus conclusions need to be made with respect to only those used in this work. It might be interesting to extend this study by incorporating IRT-based methods such as Lord's chi-square test Lord (1980), likelihood ratio test Thissen et al. (1988), or differential functioning of items and tests framework Raju et al. (1995) and see if there is any significant impact upon controlling procedures in these alternate frameworks. Further studies are needed to explore the effect of item purification, multiple correction adjustments, and their combinations in those mentioned above and with other DIF detection approaches.

Second, in our simulation study we determined the significance of simulation factors *via* beta regression models, where only double interactions were considered. However, increasing the complexity of the study design goes hand in hand with an increased complexity of the results. Thus any further extension to the study design may complicate interpretability and thus lower readability of the results.

Third, we considered all of the DIF items to be harder, or more discriminating for the focal group, which resulted in the worst-case scenario where the DIF was extremely unbalanced. While this was done to illustrate benefits of the controlling procedures, such a scenario would probably not be realistic in real data examples. A natural extension of our present research is to measure and compare the efficiency of our approach in the case of balanced DIF, i.e., with some items being easier and other items being more difficult for the reference group.

Our study covers the current gap in DIF literature as it allows for a joint evaluation of the properties of item purification and adjustments for multiple comparisons. While some of the simulation settings were inspired by previous studies to allow for comparing the results, our study is more complex and its design goes beyond previous studies, including Kim and Oshima (2013), by also incorporating non-uniform DIF, a larger variety of sample sizes, and various distributions of ability levels for the focal group. Moreover, we newly proposed and implemented the iterative combination of item purification and adjustments for multiple comparisons which, to our best knowledge, have not yet been explored in literature, but appears to be a promising tool. As such, this study offers an innovative algorithm, a detailed assessment of controlling procedures in DIF detection, and a deeper insight which may be helpful to researchers and practitioners when testing for DIF.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: Adéla Hladká and Patrícia Martinková were funded by grant number 21-03658S from the Czech Science Foundation and by the institutional support RVO 67985807 from the Institute of Computer Science of the Czech Academy of Sciences. David Magis was funded by the Fonds de la Recherche Scientifique – FNRS.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

References

Agresti, A. (2002). *Categorial data analysis* (2nd ed.). John Wiley & Sons.

Belzak, W. C. M. (2020). Testing differential item functioning in small samples. *Multivariate Behavioral Research*, *55*(5), 722–747. https://doi.org/10.1080/00273171.2019.1671162

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Berger, M., & Tutz, G. (2016). Detection of uniform and nonuniform differential item functioning by item-focused trees. *Journal of Educational and Behavioral Statistics*, *41*(6), 559–592. https://doi.org/10.3102/1076998616659371

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

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. https://doi.org/10.1177/0146621617726786

Cho, S.-J., Suh, Y., & Lee, W.-Y. (2016). An NCME instructional module on latent DIF analysis using mixture item response models. *Educational Measurement: Issues and Practice*, *35*(1), 48–61. https://doi.org/10.1111/emip.12093

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. https://doi.org/10.1207/s15324818ame0604_2

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in R. *Journal of Statistical Software*, *34*(2), 1–24. https://doi.org/10.18637/jss.v034.i02

DeMars, C. E. (2009). Modification of the Mantel-Haenszel and logistic regression DIF procedures to incorporate the SIBTEST regression correction. *Journal of Educational and Behavioral Statistics*, *34*(2), 149–170. https://doi.org/10.3102/1076998607313923

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-IRT approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. https://doi.org/10.1111/jedm.12158

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of differential item functioning using item response theory and the likelihood-based model comparison approach: Application to the Mini-Mental State Examination. *Medical Care*, *44*(11 Suppl 3), S134–S142. https://doi.org/10.1097/01.mlr.0000245251.83359.8c

Fikis, D. R., & Oshima, T. (2017). Effect of purification procedures on DIF analysis in IRTPRO. *Educational and Psychological Measurement*, *77*(3), 415–428. https://doi.org/10.1177/0013164416645844

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

Greger, D., Straková, J., & Martinková, P. (2022). Extending the ILSA study design to a longitudinal design. TIMSS & PIRLS extension in the Czech Republic: CLoSE study. In T. Nilsen, A. Stancel-Piatak, & J.-E. Gustafsson (Eds.), *Springer international handbooks of education. International handbook of comparative large-scale studies in education: Perspectives, methods and findings* (pp. 1–24). Springer. https://doi.org/10.1007/978-3-030-38298-8_31-1

Güler, N., & Penfield, R. D. (2009). A comparison of the logistic regression and contingency table methods for simultaneous detection of uniform and nonuniform DIF. *Journal of Educational Measurement*, *46*(3), 314–329. https://doi.org/10.1111/j.1745-3984.2009.00083.x

Herrera, A.-N., & Gómez, J. (2008). Influence of equal or unequal comparison group sample sizes on the detection of differential item functioning using the Mantel-Haenszel and logistic regression techniques. *Quality & Quantity*, *42*(6), 739–755. https://doi.org/10.1007/s11135-006-9065-z

Hladká, A., & Martinková, P. (2020). difNLR: Generalized logistic regression models for DIF and DDF detection. *The R Journal*, *12*(1), 300–323. https://doi.org/10.32614/RJ-2020-014

Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin*, *104*(1), 145–149. https://doi.org/10.1037/0033-2909.104.1.145

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, Inc. https://doi.org/10.1002/j.2330-8516.1986.tb00186.x

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

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

Khalid, M. N., & Glas, C. A. (2014). A scale purification procedure for evaluation of differential item functioning. *Measurement*, *50*, 186–197. https://doi.org/10.1016/j.measurement.2013.12.019

Kim, J., & Oshima, T. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458–470. https://doi.org/10.1177/0013164412467033

Kingston, N., Leary, L., & Wightman, L. (1985). An exploratory study of the applicability of item response theory methods to the Graduate Management Admission Test. *ETS Research Report Series*, *1985*(2), i–56. https://doi.org/10.1002/j.2330-8516.1985.tb00119.x

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. https://doi.org/10.1177/0013164414529792

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. https://doi.org/10.1177/0146621614544195

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

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

Magis, D., & Facon, B. (2013). Item purification does not always improve DIF detection: A counter-example with Angoff's delta plot. *Educational and Psychological Measurement*, *73*(2), 293–311. https://doi.org/10.1177/0013164412451903

Maij-de Meij, A. M., Kelderman, H., & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research*, *45*(6), 975–999. https://doi.org/10.1080/00273171.2010.533047

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

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*(2), 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Martinková, P., & Drabinová, A. (2018). ShinyItemAnalysis for teaching psychometrics and to enforce routine analysis of educational tests. *The R Journal*, *10*(2), 503–515. https://doi.org/10.32614/RJ-2018-074

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307

Martinková, P., Hladká, A., & Potužníková, E. (2020). Is academic tracking related to gains in learning competence? Using propensity score matching and differential item change functioning analysis for better understanding of tracking implications. *Learning and Instruction*, *66*, 101286. https://doi.org/10.1016/j.learninstruc.2019.101286

Martinková, P., & Hladká, A. (2023). *Computational aspects of psychometric methods: With R*. Chapman and Hall/CRC.

Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied Psychological Measurement*, *20*(3), 257–274. https://doi.org/10.1177/014662169602000306

Navas-Ara, M. J., & Gómez-Benito, J. (2002). Effects of ability scale purification on the identification of DIF. *European Journal of Psychological Assessment*, *18*(1), 9–15. https://doi.org/10.1027//1015-5759.18.1.9

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). SAGE Publications, Inc.

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics : Vol. 26. Psychometrics* (1st ed., pp. 125–167). Elsevier. https://doi.org/10.1016/S0169-7161(06)26005-X

Penfield, R. D., Gattamorta, K., & Childs, R. A. (2009). An NCME instructional module on using differential step functioning to refine the analysis of DIF in polytomous items. *Educational Measurement: Issues and Practice*, *28*(1), 38–49. https://doi.org/10.1111/j.1745-3992.2009.01135.x

R Core Team. (2019). *R: A language and environment for statistical computing [Computer software manual]*. https://www.R-project.org/

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Raju, N. S., van der Linden, W. J., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement*, *19*(4), 353–368. https://doi.org/10.1177/014662169501900405

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2021). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*(5), 2101–2113. https://doi.org/10.3758/s13428-021-01689-0

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, *46*(1), 561–584. https://doi.org/10.1146/annurev.ps.46.020195.003021

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika*, *58*(2), 159–194. https://doi.org/10.1007/BF02294572

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

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147–169). Lawrence Erlbaum Associates.

van de Water, E. (2014). *A meta-analysis of type I error rates for detecting differential item functioning with logistic regression and Mantel-Haenszel in Monte Carlo studies* [Unpublished doctoral dissertation]. Georgia State University.

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

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42–57. https://doi.org/10.1177/0146621607314044