Liu, I. [et al.]. Generalized Mantel–Haenszel estimators for simultaneous differential item functioning tests. "Educational and psychological measurement", 15 Octubre 2022. DOI 10.1177/00131644221128341

Educational and Psychological Measurement

Generalized Mantel-Haenszel Estimators for Simultaneous Differential Item Functioning Tests

Journal:	*Educational and Psychological Measurement*
Manuscript ID	EPM-21-0285.R3
Manuscript Type:	Original Research Article
Keywords:	Differential Item Functioning, dually consistent, Mantel-Haenszel estimator, multiple items
Abstract:	The Mantel-Haenszel estimator is one of the most popular techniques for measuring Differential Item Functioning (DIF). A generalization of this estimator is applied to the context of DIF to compare items by taking the covariance of odds ratio estimators between dependent items into account. Unlike the Item Response Theory, the method does not rely on the local item independence assumption which is likely to be violated when one item provides clues about the answer of another item. Furthermore, we use these (co)variance estimators to construct a hypothesis test to assess DIF for multiple items simultaneously. A simulation study is presented to assess the performance of several tests. Finally, the use of these DIF tests is illustrated via application to two real data sets.

GENERALIZED MANTEL–HAENSZEL ESTIMATORS FOR SIMULTANEOUS DIFFERENTIAL ITEM FUNCTIONING TESTS

September 7, 2022

GENERALIZED MANTEL–HAENSZEL ESTIMATORS FOR SIMULTANEOUS DIFFERENTIAL ITEM FUNCTIONING TESTS

Abstract

The Mantel-Haenszel estimator is one of the most popular techniques for measuring Differential Item Functioning (DIF). A generalization of this estimator is applied to the context of DIF to compare items by taking the covariance of odds ratio estimators between dependent items into account. Unlike the Item Response Theory, the method does not rely on the local item independence assumption which is likely to be violated when one item provides clues about the answer of another item. Furthermore, we use these (co)variance estimators to construct a hypothesis test to assess DIF for multiple items simultaneously. A simulation study is presented to assess the performance of several tests. Finally, the use of these DIF tests is illustrated via application to two real data sets.

Key words:

Differential Item Functioning; dually consistent; Mantel–Haenszel estimator; multiple items.

1. Introduction

Differential Item Functioning (DIF) generally refers to the study of bias in testing (Osterlind & Everson, 2009, pp. 8–13). However, DIF is a necessary, but not sufficient, condition for item bias (Zumbo, 1999). Members of a group of interest, often a particular gender or ethnic group, may be dis/advantaged on a test in comparison to other groups even when conditioning for overall ability. If an item (or a question on a test in this context) has a different probability of being answered correctly for respondents from different groups who possess the same ability, the item is said to exhibit DIF. When comparing two groups, by convention, the group expected to be at a disadvantage is called the focal group, with the other group denoted as the reference group.

In the study of DIF, one might present data in stratified contingency tables, where the rows represent the reference (row 1) and focal (row 2) groups, the columns represent the correct (column 1) and incorrect (column 2) answers to a question and the strata represent different levels of the ability distribution. There have been numerous different techniques applied to modelling DIF, including Item Response Theory (IRT) (Osterlind & Everson, 2009, pp. 39–60), Mantel–Haenszel (MH) estimators (Holland & Thayer, 1988, pp. 129–145), Logistic Regression (Swaminathan & Rogers, 1990) and various non-parametric tests such as the Simultaneous Item Bias Test (Shealy & Stout, 1993).

The IRT-based approach (Millsap & Everson, 1993; Osterlind & Everson, 2009) uses IRT models that include examinees' abilities as latent variables. The inference of parameter estimators is obtained under the assumption of local independence that items are only be correlated through latent variables. This assumption of local item independence is likely to be violated when one item provides clues about the answer of another item or when a series of items share common material. If it is violated, inference would be misleading (Baghaei, 2008). In the literature, many authors have discussed and provided methods to evaluate local item independence (Ip, 2001; Baghaei, 2008; Lee, 2004; Christensen *et al.*, 2017; Edwards *et al.*, 2018). Alternatively, many testlet-based methods (Douglas *et al.*, 1996; Bradlow *et al.*, 1999; Li *et al.*, 2006; Wainer *et al.*, 2007; Rijmen, 2010) have been developed when there exists a common testlet to violate the local item independence.

The Mantel-Haenszel (MH) estimator (Mantel & Haenszel, 1959) allows for the calculation of a common odds ratio for the 2×2 tables across K strata. The use of the MH estimator in assessing DIF was introduced by Holland & Thayer (1988). It was found to be highly successful in this context, coming to prominence as one of the most popular measures of DIF (Camilli *et al.*, 1994; Dorans, 1989; Zwick *et al.*, 2012). Because the MH estimator is consistent even when the sample size per stratum is small, it can be useful in DIF studies even where there is a very fine partition of the ability distribution. To the best of our knowledge, the MH method has never been used to test a set of items for DIF.

When there is more than one item in a test to be analyzed, the MH procedure is applied to one item at a time by holding everything else constant to calculate the total raw scores as the ability level (Wang, 2004). The procedure is applicable when the percentages of DIF items are small (0% to 5%) (Rogers & Swaminathan, 1993). Otherwise, one might use purification approaches (Clauser *et al.*, 1993; French & Maller, 2007; Wang *et al.*, 2012; Socha *et al.*, 2015) to adjust the ability estimate. Selecting a set of anchor items that have no or only a trivial amount of DIF to analyze the other items for evidence of DIF (Wang, 2004; Kopf *et al.*, 2015) is also an alternative approach.

For a simultaneous test of DIF across several items, the procedure depends on whether the assumption of local item independence holds or not. If one assumes that the items are independent given the ability level, then a simultaneous test can be constructed in a straightforward fashion. For example, the sum of m independent chi-squared test statistics with 1 degree of freedom follows a chi-squared distribution with m degrees of freedom. Otherwise, without the local item independence assumption, a multiple testing adjustment such as the Bonferroni correction is needed (Kim & Oshima, 2013) to control the type I error. Many researchers (Rothman, 1990; Kim & Oshima, 2013; Magis *et al.*, 2015; Park *et al.*, 2021; Penfield, 2001; Sauder & DeMars, 2020; Thissen *et al.*, 2002) have discussed methods for simultaneous and multiple comparison tests.

The main contribution of this paper is the use of the generalized Mantel-Haenszel (GMH) estimators, derived by Greenland (1989) and extended by Liu & Suesse (2008), to test a set of items for DIF simultaneously without assuming local item independence, removing the need to make a multiple testing adjustment. One may also use this technique to compare DIF between two items. Similar to the ordinary MH estimator, the GMH estimators are dually consistent in a sparse data limiting model and in a large stratum limiting model. A sparse data limiting model assumes that the sample size per stratum is fixed and the number of strata → ∞. A large stratum limiting model assumes that the number of strata is fixed and the sample size → ∞. This provides flexibility for using the GMH estimators with different partitions of ability distributions. Finally, Liu & Suesse (2008) derived a method for comparing two items which we apply to pairwise comparisons in the context of DIF.

This paper defines and describes the GMH estimators and their (co)variance estimators in Section 2. Section 3 shows pairwise comparisons and simultaneous tests for DIF. Section 4 evaluates the performance of the proposed simultaneous test using a simulation study. Examples are provided in Section 5. We conclude with a discussion in Section 6.

2. Generalized Mantel-Haenszel (GMH) estimators

A three parameter logistic model frequently used in item response theory (IRT) has the following form

$$\frac{\pi(\theta)-c}{1-c}=\frac{1}{1+\exp(-a(\theta-b))} \tag{1}$$

with parameters a, b and c. The variable θ is the person ability and $\pi(\theta)$ approaches c when $\theta \to -\infty$. Often the parameter c is set to zero and then (1) reduces to a standard two parameter logistic regression model

$$\log\left(\frac{\pi(\theta)}{1-\pi(\theta)}\right)=\beta_0+\beta_1\theta, \tag{2}$$

where $\beta_0=-ab$ is the intercept and $\beta_1=a$ is the slope. By adding the group effect, the following model shows different probabilities of correctly responding for the focal and reference groups conditioning on the person ability:

$$\log\left(\frac{\pi(\theta)}{1-\pi(\theta)}\right)=\beta_0+\beta_1\theta+\gamma I_{(\text{group=ref})}, \tag{3}$$

where $I_{(\text{group=ref})}=1$ for the reference group and $I_{(\text{group=ref})}=0$ for the focal group.

Next, we describe the connection between the two parameter IRT model (2) and the model used in the paper. Adding the m items into (2), we have

$$\log\left(\frac{\pi_j(\theta)}{1-\pi_j(\theta)}\right) = \beta_{0j} + \beta_{1j}\theta, \tag{4}$$

where $j = 1, \ldots, m$. Model (5) treats the person ability effect $\{\beta_{1j}\theta\}$ as discrete using the idea from a Rasch model (Rasch, 2003):

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_j + u_{ij}, \tag{5}$$

where $i = 1, \ldots, n$ refers to person i and u_{ij} is the ability for person i on item j. Notice that the model (5) uses $(n-1) \times m$ parameters u_{jk} to describe person ability effects, whereas the model (4) uses m parameters.

Two popular methods to test for DIF include the MH method (Holland & Thayer, 1988) and the logistic regression method (Swaminathan & Rogers, 1990). This paper focuses on the former, by treating the person ability as discrete having K levels. Thus, the model (5) is simplified to

$$\log\left(\frac{\pi_{j|k}}{1-\pi_{j|k}}\right) = \beta_j + u_{jk}, \tag{6}$$

where $k = 1, \ldots, K$ and $K < n$. The model used in the paper has the following form:

$$\log\left(\frac{\pi_{j|gk}}{1-\pi_{j|gk}}\right) = \beta_{gj} + u_{jk}, \tag{7}$$

where g, j, and k refer to the group (reference or focal), item, and ability level, respectively. Model (7) adds the group effects into the model (6). The $\pi_{j|gk}$ is the probability of giving a correct answer for item j when a subject lies in group g with ability level k. For the scope of this paper, we focus on the situation of having two groups (reference and focal), i.e. $G = 2$. In general, it can be extended to cases with $G > 2$.

Model (7) assumes that there is no interaction between ability and group, that is, the odds of giving a correct answer between two groups remain the same across different ability levels. It implies that the common log odds ratio comparing the reference ($g = 1$) with the focal ($g = 2$) group is $\gamma_j = \beta_{1j} - \beta_{2j}$ for the jth item. When $\gamma_j \neq 0$, we say that item j exhibits DIF. If the common odds ratio assumption does not hold, the MH method cannot be used. This assumption limits the type of DIF one could test using the MH method. For example, adding the ability and group interactions to Model (3), we have

$$\log\left(\frac{\pi(\theta)}{1-\pi(\theta)}\right) = \beta_0 + \beta_1\theta + \gamma I_{(\text{group=ref})} + \tau\theta I_{(\text{group=ref})}. \tag{8}$$

Model (8) violates the common odds ratio assumption because the DIF is not uniform across the ability levels. That is, the odds of a correct answer between the reference and the focal groups could vary across different ability levels. Paek (2012) discussed various types of DIF tests associated with different logistic regression models.

In the DIF context, while being restricted to testing only for uniform DIF limits the use of the MH method, there are several advantages including the generalization of the model (e.g., Model (7)), and the flexibility on the sample size within strata. Because Model (7) treats the person ability effect as discrete instead of continuous, it allows the probability of a correct answer to have a non-monotonic trend association with ability. When the total sample size is large, the MH approach is appropriate even if there are few observations within strata. The test statistics based on the dually consist estimators (Greenland, 1989) behaves well in the two limiting cases.

Suppose there are m items to be assessed for DIF. We consider a $2 \times m$ table, with one row representing the reference group, the other the focal group, and where the columns represent items. The cell count represents the number of people who respond correctly for the corresponding row and column. There are K such tables for K ability levels. In the $2 \times m \times K$ table, the cell counts are not independent, because the same people contribute to multiple cell counts. Thus, the sum of cell counts is not equal to the total number of people. Therefore, the traditional Chi-squared test fails to test the independence between the row and column variables. This issue has been widely discussed in the context of multiple column responses (Loughin & Scherer, 1998; Agresti & Liu, 1999, 2001), where survey respondents are asked to tick all answers that apply. To analyze such data, we need complete information on the response profile which shows all patterns of correct answers on m items. Because each item has only two possible responses (correct and incorrect), the total number of patterns is 2^m. Therefore, the complete data can be expressed in a $2 \times 2^m \times K$ contingency table.

Let $X_{j|gk}$ denote the number of people who responded correctly to item j from group g

within stratum k and $\overline{X}_{j|gk}$ denote the complement i.e. the number of people who responded incorrectly to item j from group g within stratum k. Within stratum k, there are n_{gk} subjects in group g. Then the total number of respondents in the k^{th} stratum is $N_k = \sum_g n_{gk}$. The MH estimator (Mantel & Haenszel, 1959) of γ_j becomes:

$$\widehat{\gamma}_j = \log\left(\frac{\sum_{k=1}^{K} X_{j|1k}\overline{X}_{j|2k}/N_k}{\sum_{k=1}^{K} X_{j|2k}\overline{X}_{j|1k}/N_k}\right). \tag{9}$$

Note that $\widehat{\gamma}_j$ is the log of the ordinary MH estimator.

The estimator for the variance of the (log) odds ratio between different groups for item j derived from Greenland (1989) is:

$$\widehat{\text{Var}}(\widehat{\gamma}_j) = \frac{\sum_{k=1}^{K} c_{j|k}h_{j|k}}{2C_j^2} + \frac{\sum_{k=1}^{K} c_{j|k}\overline{h}_{j|k} + \overline{c}_{j|k}h_{j|k}}{2C_j\overline{C}_j} + \frac{\sum_{k=1}^{K} \overline{c}_{j|k}\overline{h}_{j|k}}{2\overline{C}_j^2}. \tag{10}$$

The estimator for the covariance between the (log) odds ratios between different groups for items j and ℓ derived by Liu & Suesse (2008) is:

$$\widehat{\text{Cov}}(\widehat{\gamma}_j, \widehat{\gamma}_\ell) = \frac{D^{11}}{C_jC_\ell} + \frac{D^{01}}{\overline{C}_jC_\ell} + \frac{D^{10}}{C_j\overline{C}_\ell} + \frac{D^{00}}{\overline{C}_j\overline{C}_\ell}, \tag{11}$$

where:

$$\begin{aligned}
D^{st} &= \sum_{k=1}^{K} d_k^{st} \\
d_k^{st} &= \frac{1}{N_k^2}(X^s_{j|1k}X^t_{\ell|1k}X^{\overline{s}\overline{t}}_{j\ell|2k} + X^{st}_{j\ell|1k}X^{\overline{s}}_{j|2k}X^{\overline{t}}_{\ell|2k} - X^{st}_{j\ell|1k}X^{\overline{s}\overline{t}}_{j\ell|2k}) \\
C_j &= \sum_{k=1}^{K} c_{j|k},\ \overline{C}_j = \sum_{k=1}^{K} \overline{c}_{j|k} \\
c_{j|k} &= X_{j|1k}\overline{X}_{j|2k}/N_k,\ \overline{c}_{j|k} = X_{j|2k}\overline{X}_{j|1k}/N_k, \\
h_{j|k} &= (X_{j|1k} + \overline{X}_{j|2k})/N_k, \quad \text{and } \overline{h}_{j|k} = (X_{j|2k} + \overline{X}_{j|1k})/N_k.
\end{aligned}$$

We extend the notation $X_{j|gk}$ for the counts by introducing $X^s_{j|gk}$ for those who responded positively or negatively indicated by superscript s, where s denotes either 0 (incorrect) or 1 (correct) to item j from group g in stratum k. Similarly when considering multiple items simultaneously $X^{st}_{j\ell|gk}$ refers to respondents from group g in stratum k who answered s to item j and t to item ℓ. Note that $\overline{s} = 1 - s$. So for example $X^{\overline{10}}_{j\ell|gk} = X^{01}_{j\ell|gk}$.

Using the above estimators for $\widehat{\gamma}_j$ and $\widehat{\text{Cov}}(\widehat{\gamma}_j, \widehat{\gamma}_\ell)$ we can estimate the covariance matrix $\mathbf{\Sigma}$ of $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \dots, \widehat{\gamma}_m)^\top$ by

$$\widehat{\mathbf{\Sigma}} = \begin{pmatrix} \widehat{\text{Var}}(\widehat{\gamma}_1) & \widehat{\text{Cov}}(\widehat{\gamma}_1, \widehat{\gamma}_2) & \dots & \widehat{\text{Cov}}(\widehat{\gamma}_1, \widehat{\gamma}_m) \\ \widehat{\text{Cov}}(\widehat{\gamma}_2, \widehat{\gamma}_1) & \widehat{\text{Var}}(\widehat{\gamma}_2) & \dots & \widehat{\text{Cov}}(\widehat{\gamma}_2, \widehat{\gamma}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}(\widehat{\gamma}_m, \widehat{\gamma}_1) & \widehat{\text{Cov}}(\widehat{\gamma}_m, \widehat{\gamma}_2) & \dots & \widehat{\text{Var}}(\widehat{\gamma}_m) \end{pmatrix}.$$

The notation in this section follows the work derived by Liu & Suesse (2008). Although the formulae are complex, the estimates in (9) – (11) have a closed form expression. The R code (on request) is available for calculation.

3. Tests for DIF

Single Tests for DIF

Although this paper focuses on the GMH estimators discussed in the previous section applied to DIF tests for multiple items, we first review a test when there is only one item of interest. Common practice for assessing the extent of DIF uses the Educational Testing Services classification scheme (Dorans & Holland, 1992). This classification scheme applies the MH Delta–DIF statistic given by the following steps:

$$\text{Let } \widehat{\Delta}_j = -2.35\widehat{\gamma}_j,$$

where $\widehat{\gamma}_j$ is as in equation (9). The level of DIF depends on the value of $|\widehat{\Delta}_j|$.

- Step 1: The extent of DIF is said to be negligible if $|\widehat{\Delta}_j| < 1.0$ or, if the statistic is not significantly different from 0.
- Step 2: The extent of DIF is said to be moderate to large if $|\widehat{\Delta}_j| \geq 1.5$ and the statistic is significantly larger than 1.0

- Step 3: Otherwise the extent of the DIF is said to be slight to moderate.

For Step 1, an analogue statistical test can be conducted for item j by testing

$$H_{0,j} : \gamma_j = 0 \ \ vs. \ \ H_{1,j} : \gamma_j \neq 0,$$

using a Wald type statistic

$$Z_j = \frac{\widehat{\gamma}_j}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\gamma}_j)}}. \tag{12}$$

The test statistic Z_j is asymptotically distributed standard normal under H_0. Due to the dual consistency, the asymptotic distribution applies to both limiting models, the sparse data and the large stratum limiting models.

For Step 2, if $\Delta = 1.0$, we have $\gamma = -0.426$ $(= -\frac{1.0}{2.35})$. Thus, a null interval hypothesis $H_0 : |\gamma_j| \leq 0.426$ $vs.$ $H_1 : |\gamma_j| > 0.426$ can be conducted to find whether $|\Delta_j|$ is significantly larger than 1.0. Paek & Holland (2015) showed the p-value based on $|\widehat{\Delta}_j|$ using Schervish's p-value formula for a null interval hypothesis (Schervish, 1996). Based on $|\widehat{\gamma}_j|$, the p-value has the following formula:

$$p - \text{value} = \Phi\left(\frac{-0.426 - |\widehat{\gamma}_j|}{s}\right) + \Phi\left(\frac{0.426 - |\widehat{\gamma}_j|}{s}\right),$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution and $s = \sqrt{\widehat{\mathrm{Var}}(\widehat{\gamma}_j)}$.

For Step 3, the DIF is said to be slight to moderate if the DIF is not negligible (from Step 1) nor moderate to large (from Step 2). Note that one can use the function `difMH` from the R package `difR` (Magis *et al.*, 2010, 2020) to conduct the above tests.

Pairwise Comparisons for DIF

Instead of analyzing each item separately one may also compare DIF between two items. For example if two items j and ℓ are measured to exhibit some level of DIF, we may wish to determine whether the extent of DIF for item j is significantly different from the extent of DIF in item ℓ. While it is true in general that test designers want to detect and remove all DIF items, pairwise comparisons can still be very useful in determining if one or both of the two items need to be modified or removed. Item revisions are essentially a balancing act informed by previous trials. During cycles of test validation, decisions are constantly being made as to which items are performing as expected in eliciting the target competencies and which are not. In this sense, the more evidence available, including pairwise DIF differences, the more informed decisions made about revising the test will be.

We can calculate the covariance of the log odds ratio estimators for the two items using the GMH method, which then allows us to calculate a standard 95% confidence interval for the difference in the log odds ratios as follows:

$$\begin{aligned} 95\%\ \mathrm{CI}: & \widehat{\gamma}_j - \widehat{\gamma}_\ell \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\widehat{\gamma}_j - \widehat{\gamma}_\ell)} \\ & \widehat{\gamma}_j - \widehat{\gamma}_\ell \pm 1.96\sqrt{\widehat{\mathrm{Var}}(\widehat{\gamma}_j) + \widehat{\mathrm{Var}}(\widehat{\gamma}_\ell) - 2\widehat{\mathrm{Cov}}(\widehat{\gamma}_j, \widehat{\gamma}_\ell)}, \end{aligned} \tag{13}$$

where the variance and covariance estimators are given in (10) and (11). If the confidence interval (13) excludes 0, we conclude there is a 5% significant difference in the extent of DIF in the two items.

Simultaneous Tests for DIF

Unlike the current use of MH tests which allow one item to be tested at a time, the inference based on the GMH estimators allows us to test for DIF in multiple items simultaneously. By calculating the covariance matrix for the log odds ratio estimators $\widehat{\boldsymbol{\gamma}} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_m)^\top$ over m items, we are able to perform a simultaneous test. By utilizing a single GMH test we can keep the type I error rate the same as the significance level. Doing so may be beneficial to test developers (Roussos & Stout, 1996). Our methods do not need to make any adjustments for multiple comparisons. The test statistic itself has already taken the dependency between items into account.

Define the global null hypothesis as $H_0 = H_{0,1} \cap H_{0,2} \cap \cdots \cap H_{0,m}$, where $H_{0,j} : \gamma_j = 0$ for all $j = 1, \ldots, m$, and the alternative is that any of the individual null hypothesis $H_{0,j}$ does not hold (or in other words, any of the individual alternative hypothesis $H_{1,j} : \gamma_j \neq 0$ holds) or in formula $H_1 = H_{1,1} \cup H_{1,2} \cup \cdots \cup H_{1,m}$.

Then a simultaneous test for H_0 can be conducted by applying Hotelling's T^2 test statistic

(Wellek, 2010, pp. 235–264) as follows:

$$\text{Under } H_0 : W = (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^\top \widehat{\boldsymbol{\Sigma}}^{-1} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \to_d \chi^2_{df=m}. \tag{14}$$

For example, to perform a simultaneous test of no DIF in m items we can use the global test statistic W as follows:

$$\text{Under } H_0 : W = \widehat{\boldsymbol{\gamma}}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\gamma}} \to_d \chi^2_{df=m}. \tag{15}$$

This allows us to use a single test statistic to detect evidence of DIF across all m items.

The test statistics (15) requires a positive definite covariance matrix $\boldsymbol{\Sigma}$ in order for W to be positive. Unfortunately for large m, the estimated $\boldsymbol{\Sigma}$ is often not positive definite. If it is not positive definite, we adjust this matrix by using the closest positive definite matrix using the `R` function `nearPD` of the package `Matrix` (Bates & Maechler, 2021). The variances of $\widehat{\gamma}_j$ remain fixed, but the co-variance estimators between $\widehat{\gamma}_j$ and $\widehat{\gamma}_\ell$ are adjusted. This adjustment makes the covariance matrix positive definite, however it can make the covariance estimators unreliable and the asymptotic distribution invalid. The impact of the adjustment and the type I error rates based on the theoretical asymptotic distribution are evaluated later in the simulation section.

Under the assumption of local item independence, the test statistic

$$W_{ind} = \sum_{j=1}^{m} Z_j^2$$

follows a $\chi^2_{df=m}$ distribution asymptotically under H_0. It is an alternative global test statistic for the simultaneous test. When the local item independence assumption does not hold, the W_{ind} fails to follow the $\chi^2_{df=m}$ asymptotic distribution. One could compute the p-value using the bootstrap resampling technique (Efron & Tibshirani, 1993). The next section evaluates the performance of various global test statistics.

Comparison to the Likelihood Ratio Test

As well as the global tests discussed, we can also use a maximum likelihood approach for a simultaneous test for no DIF in any item. Using the standard likelihood-ratio (LR) test to compare models (6) and (7) we obtain a test for with a null hypothesis of no group effects, i.e. no effect of DIF on any item. As discussed by Breslow (1981), the maximum likelihood estimator of odds ratios are not consistent in sparse cases. Andersen (1980) showed that when each stratum consists of a single matched pair, the maximum likelihood estimator of γ_j converges to double the true value. Therefore, the LR test described, unlike the global tests, is not consistent in sparse data cases. We present this as a benchmark to compare to the results of the global tests in the next section. As discussed in section 6.1 of Wang *et al.* (2014), for unordered categorical variables the global invariance test is the normal likelihood ratio test. The tests provided can therefore also be viewed as score-based test for measurement invariance.

4. Simulation Study

In this section we evaluate the performance for the proposed global test W, the global test W_{ind}, the item based tests and pairwise based comparisons. Since the proposed method does not rely on the local independence assumption, we introduce another set of odds ratios $\{\Gamma_{j\ell|gk}\}$ as measures of dependence across items conditional on the ability as follows:

$$\Gamma_{j\ell|gk} = \frac{\pi^{11}_{j\ell|gk}\pi^{00}_{j\ell|gk}}{\pi^{10}_{j\ell|gk}\pi^{01}_{j\ell|gk}}, \qquad (16)$$

where $\pi^{st}_{j\ell|gk}$ refers to the probability of responding s to item j and t to item ℓ for respondents from group g in stratum k. The probabilities $\{\pi^{00}_{j\ell|gk}, \pi^{01}_{j\ell|gk}, \pi^{10}_{j\ell|gk}, \pi^{11}_{j\ell|gk}\}$ form a joint distribution for items j and ℓ for respondents from group g in stratum k. The odds ratios $\{\Gamma_{j\ell|gk}\}$ have been used by various authors in simulation studies for multiple response data, for example Liu & Suesse (2008). Often items are positively correlated resulting in $\Gamma_{j\ell|gk} > 1$. The case $\Gamma_{j\ell|gk} = 1$ refers to independence and $\Gamma_{j\ell|gk} < 1$ to negatively correlated items. For simplicity we consider a constant Γ that is $\Gamma = \Gamma_{j\ell|gk}$ with values of $\Gamma = 1, 10$.

We first determine the marginal probability distribution for each item j using Model (3):

$$\log\left(\frac{\pi_j(\theta)}{1-\pi_j(\theta)}\right) = \beta_{0j} + \beta_1\theta + \gamma I_{(\text{group=ref})},$$

by setting $\beta_{0j} = -1 + e_j$ with $e_j \sim N(0,1)$, $\beta_1 = 0.5$ and by simulating K different ability levels $\theta_1, \ldots, \theta_K$ (one for each stratum) according to $\theta \sim N(\mu_g, 1)$. Consider the scenarios by varying the following settings:

- The impact factor: (1) $\mu_1 = \mu_2 = 0$; (2) $\mu_1 = 0$ for the reference group, $\mu_2 = 1$ for the focal group. Note: the first option shows that the reference and focal groups have the same ability distribution. The second option shows that the mean ability for the focal group is higher than the reference group.
- The level of DIF: (1) $\gamma = 0$ ($\Delta = 0$) no DIF; (2) $\gamma = 0.426$ ($|\Delta| = 1$) moderate DIF (in favor of the reference group).
- The number of strata K: (1) $K = 20$; (2) $K = 100$. We stratified observations based on their ability into K tables.
- The number of observations in the reference group (N_1) and the focal group (N_2): (1) $N_1 = 1000$, $N_2 = 200$; (2) $N_1 = 500$, $N_2 = 500$. By default we let $N_{gk} = N_g/K$ for $g = 1, 2$ and $k = 1, \ldots, K$ leading to a balanced scenario. However, this balanced case gives always $\mu_1 = 0$ and $\mu_2 = 0$. For the case $\mu_2 = 1$ we need to change the N_{2k} values in order to obtain $\mu_2 = 1$, because each stratum has a different ability level θ. This means strata with a higher ability level need values $N_{2k} \geq N_g/K$ and strata with a lower ability level need generally values $N_{2k} \leq N_g/K$. The values of N_{2k} were determined by simulated annealing until the condition $\mu_2 = 1$ was met. There is usually no unique solutions. The final solution that meets $\mu_2 = 1$ was randomly selected. This procedure provides unbalanced sample sizes of N_{2k} across K strata.
- The number of items m for the global test: (1) $m = 2, 3, 4, 5, 7, 10, 50$.
- The number of items with DIF (called false hypotheses FH): (1) $FH = 0$; (2) $FH = 1$; (3) $FH = 5$; (4) $FH = 10$; (5) $FH = 20$.
- The dependency between items given strata: (1) $\Gamma = 1$ (independence); (2) $\Gamma = 10$ (dependence).

Given the values of Γ and π_j for a given stratum and group, the joint distribution of the m items is determined via the `R` package `mipfp` (Barthélemy & Suesse, 2018) using the IPF algorithm. This joint distribution is characterised by 2^m probabilities. Then a random binary sequence of length m is obtained. For large m this algorithm is infeasible, as 2^m is too large. To circumvent this problem, we created independent blocks of 10 items each for $m = 50$.

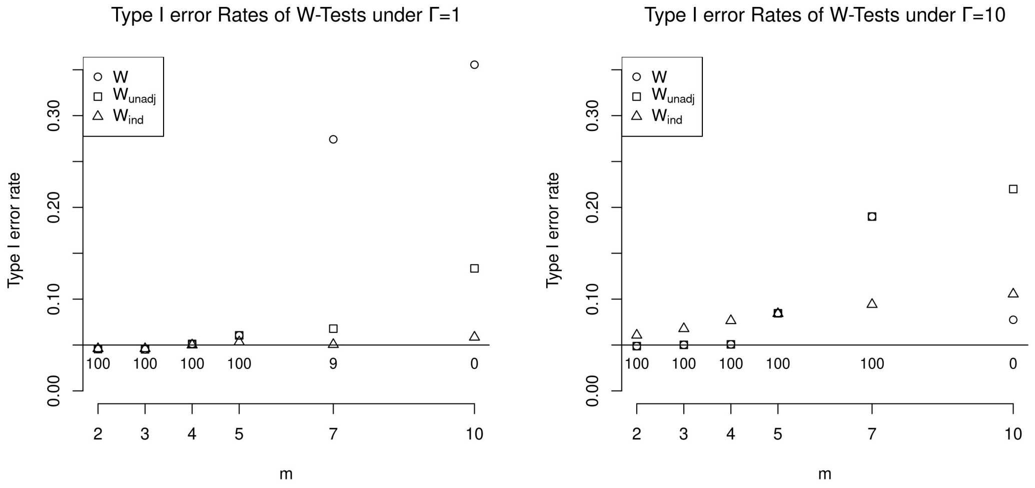

FIGURE 1.

Type I error rates ($FH = 0$) of W tests for various m, $K = 20$, $N_1 = 500$ and $N_2 = 500$ based on 10,000 simulated data sets. Here W is the adjusted statistic and W_{unadj} is the unadjusted statistic. Below the vertical line at 5%, the proportion when W and W_{unadj} are the same is shown.

We first evaluate the performance of the global test statistic W under the adjustment of $\widehat{\Sigma}$ to ensure a positive definite covariance matrix. Figure 1 shows the type I error rates at a 5% significance level based on 10,000 simulated data sets for the adjusted test statistic W, the unadjusted test statistic W_{unadj}, and the statistic W_{ind}. When the $\widehat{\Sigma}$ is positive definite without the adjustment, then $W = W_{unadj}$. The proportion of times when $W = W_{unadj}$ is given below the horizontal line at the 5% level. Our proposed test has its desired asymptotic distribution up to $m = 4$, but we cannot rely anymore on its asymptotic distribution for $m \geq 5$, especially when the adjustment is required. We recommend using the non-parametric bootstrap method when $m \geq 5$. We also notice from the results that the asymptotic distribution of W_{ind} is as expected under $\Gamma = 1$ (independence), but the type I error rate is exceeding the nominal level under $\Gamma = 10$ (dependence of items).

Figure 2 shows the power under $FH = 1$ based on the asymptotic distribution. Since W_{unadj} and W are identical and keep the nominal level only for small m (< 4), we focus on the

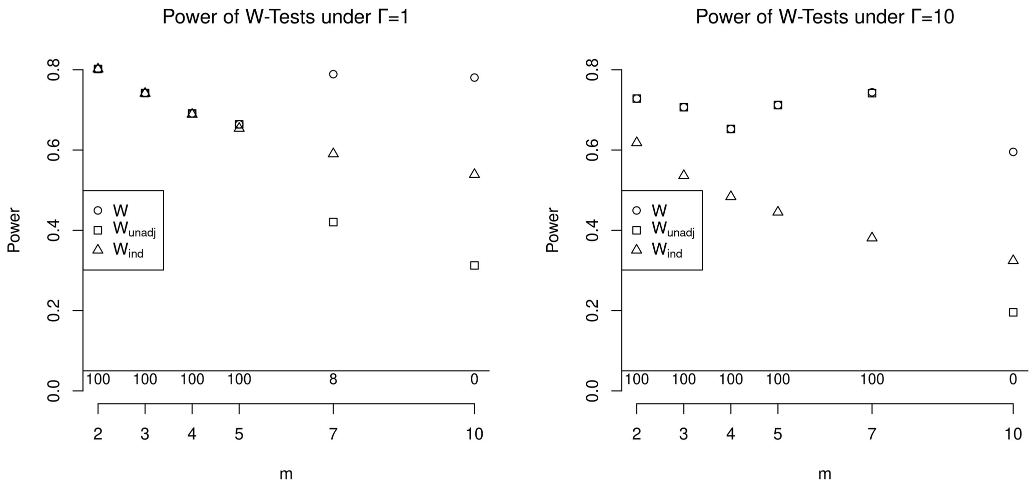

FIGURE 2.

Power of W tests for various m and $FH = 1$, $K = 20$, $N_1 = 500$ and $N_2 = 500$ based on 10,000 simulated data sets. Here W is the adjusted statistic and W_{unadj} is the unadjusted statistic. Below the vertical line at 5%, the proportion when W and W_{unadj} are the same is shown.

performance of power under small m. The proposed test W is more powerful than W_{ind}, as expected under $\Gamma = 10$ (dependence), because W uses the covariance estimators of $\widehat{\gamma}$, whereas W_{ind} does not.

Figures 3 and 4 show the rejection rates of the global tests W and W_{ind} for H_0 vs. H_1 using the asymptotic distribution (Asym) and using the non-parametric bootstrap method (Boot) at a 5% significance level, over $2,000$ simulated datasets. Similarly, we show results for testing $H_{0,j}$ vs. $H_{1,j}$ using the coverage of the 95% confidence intervals of γ_j. Because there is little difference across items, the averaged values across all items are presented to give a better summary. We also considered testing the pairwise hypothesis $H_{0,j\ell} : \gamma_j = \gamma_\ell$ vs. $H_{1,j\ell} : \gamma_j \neq \gamma_\ell$ by showing the averaged coverage of the 95% confidence interval of $\gamma_j - \gamma_\ell$ for all pairs. Tables 7 and 8 in Supplementary Document give the exact values. When $FH = 0$, the proportion of times that the null hypothesis was rejected equals 1−coverage. If a test is good, the proportion should be close to the value of 0.05 at a 5% significance level. For these figures and tables, we consider various numbers of strata and observations within strata, such that the scenarios cover the sparse data case (e. g., $K = 100$, $N_1 = N_2 = 500$) and the large strata case (e. g., $K = 20$, $N_1 = 1000$, $N_2 = 200$), and we only show the cases that $m = 50$.

The results show that the coverage of $\widehat{\gamma}_j$ and $\widehat{\gamma}_j - \widehat{\gamma}_{j'}$ is near 95% as expected, irrespective of whether the asymptotic confidence intervals are used or whether the bootstrap confidence intervals are used. It appears that the asymptotic tests are often slightly better for the sparse data case (e. g., $K = 100$, $N_1 = N_2 = 500$).

For the global test W, under H_0 the rejection rate exceeds 5% when the asymptotic method is used. This is not surprising, because when m is large, the performance of W is affected by the adjustment of $\widehat{\Sigma}$ shown in Figure 1. For the global test W_{ind}, the type I error rates based on the asymptotic distribution are around 5% when $\Gamma = 1$ (independence of items), whereas under dependence $\Gamma = 10$ the type I error rate exceeds 5%. In contrast the type I error rates of both test statistics using the bootstrap method are around the 5% level. This means that for medium and large m (e.g., $m \geq 5$) the bootstrap method has to be used in order to make valid statistical inference. We also see that W_{ind} has mostly larger power than W when using the bootstrap method. It seems surprising as W uses more information, i.e. the covariance matrix, than W_{ind}. However, the adjustment of $\widehat{\Sigma}$ to ensure positive definiteness seems to reduce power due to the uncertainty in many estimated covariances. The results are very similar for the scenario $\mu_1 = 0$, $\mu_2 = 1$ (Figure 4) and $\mu_1 = \mu_2 = 0$ (Figure 3).

We also compare the results of the global tests with the results of a maximum likelihood estimation approach. In Supplementary Document, Table 9 presents the result from the LR test comparing models (6) and (7). The table presents a number of different scenarios illustrating the behaviour of the estimator under a variety of different conditions. In the non-sparse case where $K = 5$ and $N_1 = N_2 = 200$ (80 observations per stratum) we observe the LR test gives a type I error rate of 5.2%, shown in the cell where $FH = 0$ i.e. H_0 is true. We see this the type I error rate increase as the data becomes increasingly sparse (case $K = 20, N_1 = N_2 = 60$ has 6 observations per stratum and a type I error rate of 29.6%). Moreover, for $\Gamma = 10$ we see the type I error rate is much higher than 5% even in the non-sparse case, $K = 20, N_1 = N_2 = 500$, where we observe a type I error rate of 14.2%. These cases demonstrate several scenarios where the LR

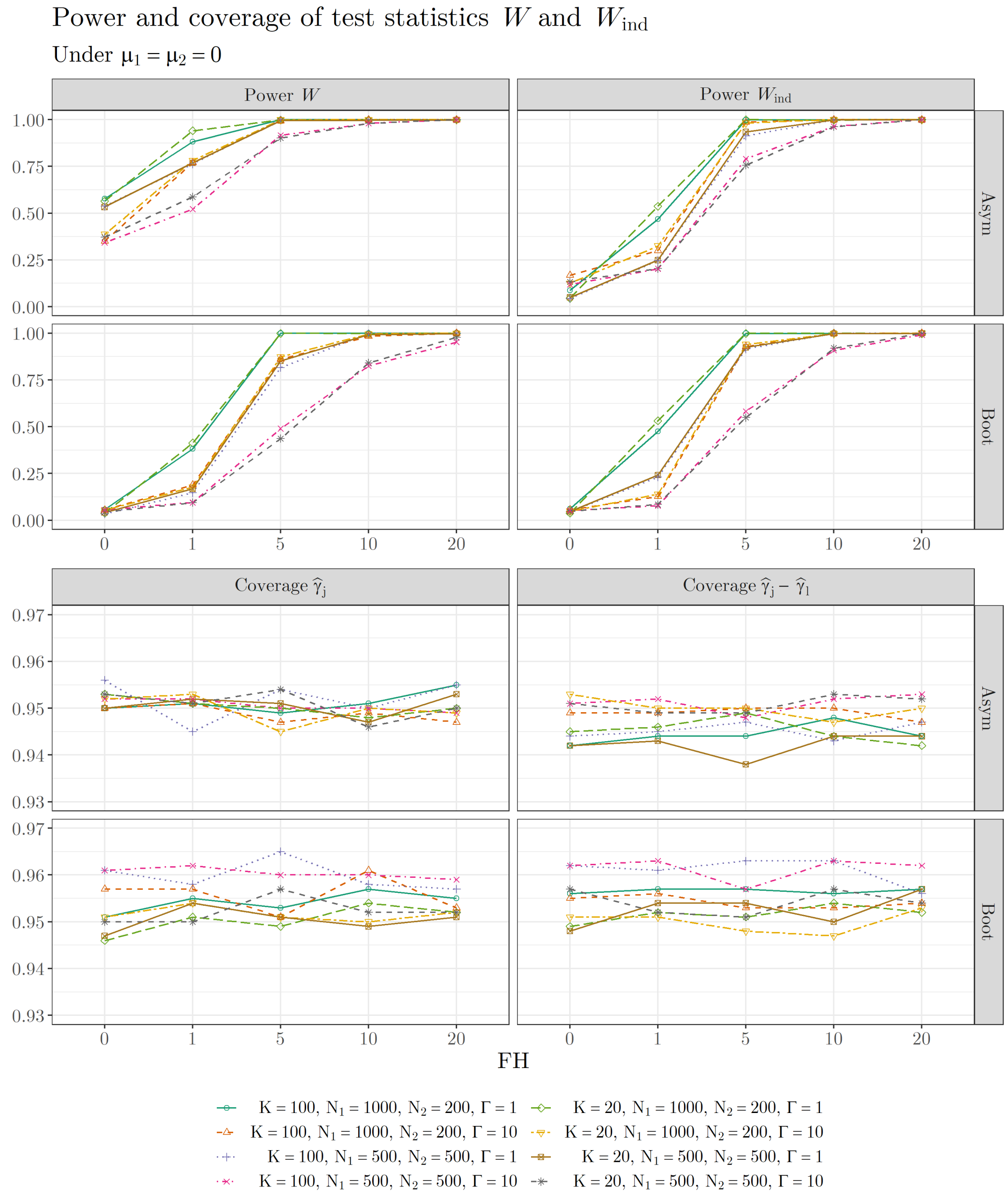

FIGURE 3.

Under $\mu_1 = \mu_2 = 0$, the power W and W_{ind} show the rejection rates of the global tests at a 5% level. The coverage $\widehat{\gamma}_j$ shows the averaged coverage of the 95% confidence intervals of γ_j over all j. The coverage $\widehat{\gamma}_j - \widehat{\gamma}_\ell$ shows the averaged coverage of the 95% confidence intervals of $\gamma_j - \gamma_\ell$ for all pairs. Both the asymptotic distribution (Asym) and the non-parametric bootstrap method (Boot) were used.

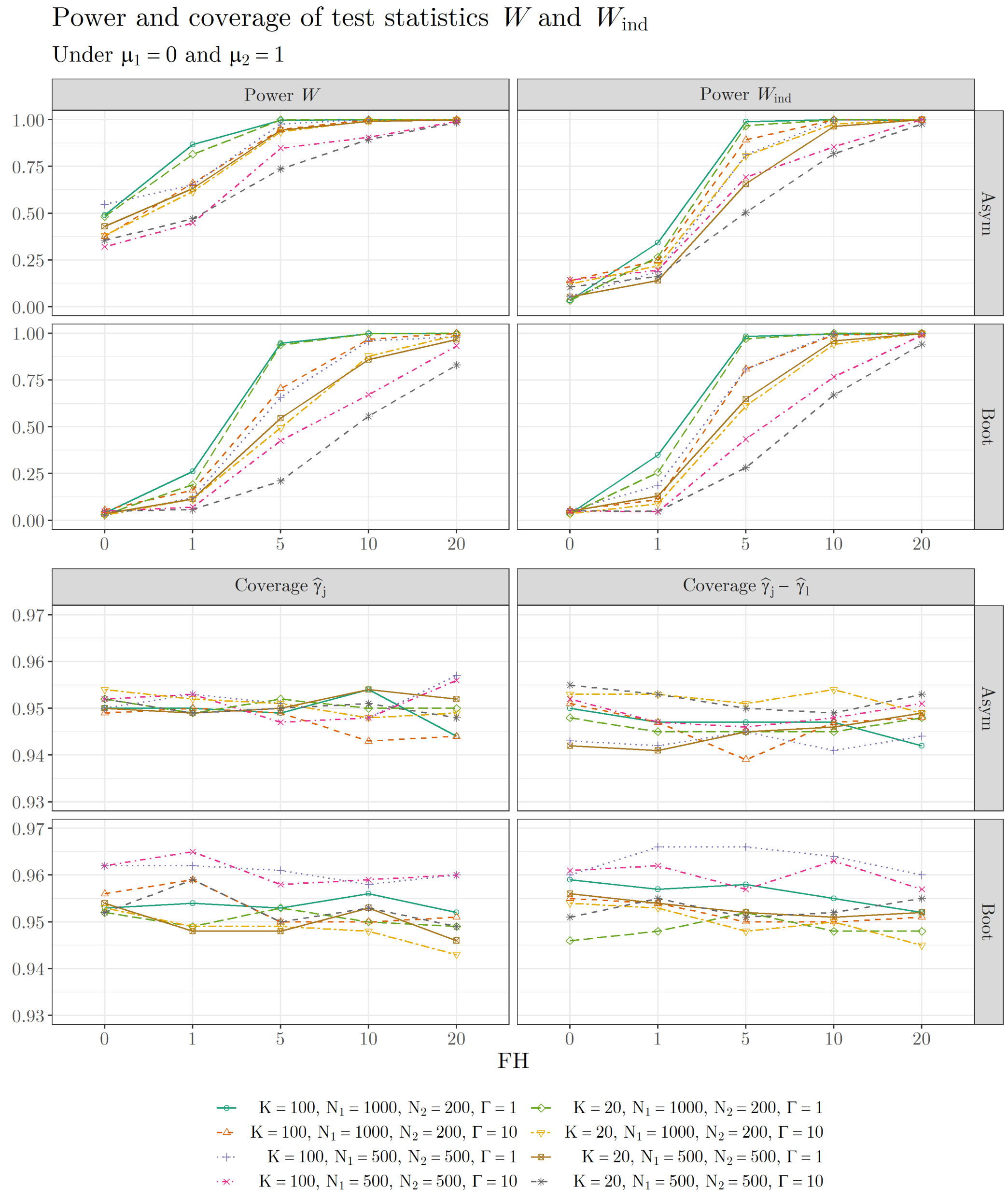

FIGURE 4.

Under $\mu_1 = 0$ and $\mu_2 = 1$, the power W and W_{ind} show the rejection rates of the global tests at a 5% level. The coverage $\widehat{\gamma}_j$ shows the averaged coverage of the 95% confidence intervals of γ_j over all j. The coverage $\widehat{\gamma}_j - \widehat{\gamma}_\ell$ shows the averaged coverage of the 95% confidence intervals of $\gamma_j - \gamma_\ell$ for all pairs. Both the asymptotic distribution (Asym) and the non-parametric bootstrap method (Boot) were used.

test has far worse performance.

From the simulation study, we conclude that of the proposed global tests the bootstrap method is recommended for most cases and W_{ind} is also generally preferred over W. For a small m (≤ 5), the asymptotic distribution of W behaves well and we recommend W over W_{ind}. The LR test approach is not recommended in sparse data cases (for sparse data or large stratum limiting models), or in cases with highly correlated items.

5. Examples

To illustrate the use of these DIF tests, we present two examples. The first example is from the 2012 Program for International Student Assessment (PISA) database (Organisation for Economic Co-operation and Development, 2012a) based on the responses provided by 15-year-old students in the subject area of science literacy. The second is from the existing test suite of the English for Academic Purposes (EAP) program at the first author's university.

5.1. PISA 2012

The Programme for International Student Assessment (PISA) is a well-known set of surveys which periodically measure the global academic performance of 15 year old students. PISA is commonly used in the DIF literature, with many papers analysing the occurrence of DIF in PISA data, for example Le (2009); Khorramdel *et al.* (2020); Chen & Jiao (2014). For the purposes of this study we consider data from the 2012 PISA paper-based assessment. In particular, we have considered the response of Australian students to the science section.

The raw data from the study, including student-level responses, was published by the OECD Organisation for Economic Co-operation and Development (2012a). We have processed this data and made it available on Kaggle for users who do not have access to SAS or SPSS (Organisation for Economic Co-operation and Development, 2021). Students participating in PISA sit one exam booklet, these booklets vary over students due to the balanced incomplete block test design of the study (Organisation for Economic Co-operation and Development, 2012b, pp. 30). We selected all Australian students who sat one specific booklet, providing a subset of students who all took a test containing the same questions. Responses were coded so that fully or partially correct answers were treated as correct, and all other responses were coded as incorrect. This provides a dataset with 1132 students responding to 35 different science questions.

We then consider the item purification approach as described by Lord (1980). All items are tested individually for DIF first and each item displaying non-negligible evidence of DIF is removed from the calculation of the ability distribution used for stratifying students. The test of each item in the purification process was assessed using the Education Testing Services classification scheme as in section 3. This left us with 25 questions after removing 10 questions which displayed non-negligible DIF. We then stratify students into $K = 5$ strata based on their total number of correct responses on the anchor items: $[0, 5) < [5, 10) < [10, 15) < [15, 20) < [20, 25)$, before assessing the test for DIF with regard to gender.

Table 1 displays the estimated common log odds ratios of each item ($\widehat{\gamma}_j$) and two-tailed p-values for individual questions by treating female students as the reference group and male students as the focal group. In 10 of the 35 questions individually assessed, DIF was detected at a 5% significance level. We also compare the extent of DIF among items through the pairwise comparison test, Table 2 shows pairwise comparisons of items which had DIF detected with 5% significance by the single item tests. We can see that of the 45 pairs of items, 25 have a 5% significant difference between them in the extent of DIF detected.

This example also shows the utility of the global test W for cases where an assumption of local independence between items does not hold. As discussed in Section 1, if this assumption does hold then W_{ind} can be used as an alternative test statistic. However, where local independence does not hold, we can still apply the global test W. This PISA dataset contains several subsets of questions which are related to each other. For example, there is a sequence of questions (PS326Q01 - PS326Q04), labelled as SCIE - P2003 Milk Q1-4. As in (Baghaei, 2008), an assumption of local item independence is likely to be violated when items share common material. To test this, we consider the odds ratios $\{\Gamma_{j\ell|gk}\}$ (16) of different items within the test set. If we examine the odds ratio pairs corresponding to these items, we find that a MH type common odds ratio estimate of $\Gamma_{j\ell} = \Gamma_{j\ell|gk}$ between questions PS326Q01 and PS326Q02 has a value of 5.801. To find p-values we used the bootstrap method based on 10,000 bootstrap samples

Question ID	PS131Q02D	PS131Q04D	PS256Q01	PS326Q01	PS326Q02
$\widehat{\gamma}_j$	0.0528	-0.0355	0.626	-0.0753	0.4415
se $\widehat{\gamma}_j$	0.1478	0.1506	0.2018	0.1441	0.1502
p_j	0.721	0.8137	0.0019	0.6012	0.0033
Question ID	PS326Q03	PS326Q04T	PS413Q04T	PS413Q05	PS413Q06
$\widehat{\gamma}_j$	0.093	-0.2287	-0.0893	0.0126	-0.5931
se $\widehat{\gamma}_j$	0.147	0.1543	0.1377	0.139	0.1594
p_j	0.5271	0.1382	0.5168	0.9276	0.0002
Question ID	PS415Q02	PS415Q07T	PS415Q08T	PS425Q02	PS425Q03
$\widehat{\gamma}_j$	0.149	0.3903	0.0443	-0.1707	-0.1911
se $\widehat{\gamma}_j$	0.1748	0.1721	0.1352	0.1427	0.1385
p_j	0.3938	0.0233	0.7433	0.2316	0.1677
Question ID	PS425Q04	PS425Q05	PS428Q01	PS428Q03	PS428Q05
$\widehat{\gamma}_j$	0.0516	0.1613	-0.3278	-0.3421	0.0377
se $\widehat{\gamma}_j$	0.1684	0.143	0.1335	0.1902	0.1518
p_j	0.7591	0.2594	0.0141	0.0721	0.8039
Question ID	PS438Q01T	PS438Q02	PS438Q03D	PS465Q01	PS465Q02
$\widehat{\gamma}_j$	-0.072	-0.0584	0.428	0.2134	0.137
se $\widehat{\gamma}_j$	0.1638	0.1718	0.1373	0.1472	0.1433
p_j	0.6605	0.734	0.0018	0.147	0.3391
Question ID	PS465Q04	PS478Q01	PS478Q02T	PS478Q03T	PS498Q02T
$\widehat{\gamma}_j$	0.0001	0.1848	0.0838	0.3308	-0.2152
se $\widehat{\gamma}_j$	0.127	0.1374	0.145	0.1379	0.129
p_j	0.9995	0.1788	0.5634	0.0164	0.0953
Question ID	PS498Q03	PS498Q04	PS514Q02	PS514Q03	PS514Q04
$\widehat{\gamma}_j$	-0.0337	0.4094	0.1616	-0.4692	-0.0617
se $\widehat{\gamma}_j$	0.131	0.1558	0.2185	0.132	0.1566
p_j	0.7971	0.0086	0.4597	0.0004	0.6938

Table 1.

MH Log odds ratio estimator $\widehat{\gamma}$ after item purification and their standard error (s.e.$(\widehat{\gamma})$), two-sided p-values p_j for identified PISA 2012 science questions.

Question ID	PS256Q01	PS326Q02	PS413Q06	PS415Q07T	PS428Q01	PS438Q03D	PS478Q03T	PS498Q02T	PS498Q04	PS514Q03
PS256Q01	-	0.6038	1.7565	0.5948	1.4156	0.732	0.7413	1.3623	0.7322	1.5567
PS326Q02	-0.2347	-	1.4872	0.4186	1.1493	0.4385	0.5173	1.0974	0.4724	1.3047
PS413Q06	0.6818	0.582	-	-0.4989	0.1101	-0.6134	-0.4956	0.0091	-0.5247	0.2431
PS415Q07T	-0.1234	-0.3163	-1.4679	-	1.1164	0.4357	0.4646	1.0852	0.4474	1.2777
PS428Q01	0.4921	0.3893	-0.6406	0.3199	-	-0.3728	-0.2875	0.2697	-0.3385	0.4983
PS438Q03D	-0.3359	-0.4115	-1.4288	-0.511	-1.1388	-	0.4647	0.9946	0.4231	1.281
PS478Q03T	-0.1508	-0.296	-1.3522	-0.3456	-1.0298	-0.2704	-	0.9198	0.3191	1.1883
PS498Q02T	0.3201	0.2158	-0.765	0.1258	-0.495	0.2917	0.1722	-	-0.2253	0.6074
PS498Q04	-0.2988	-0.4082	-1.4802	-0.4855	-1.1359	-0.3859	-0.4763	-1.0238	-	1.306
PS514Q03	0.6338	0.5166	-0.491	0.4414	-0.2156	0.5133	0.4117	-0.0993	0.4512	-

Table 2.

Pairwise comparisons: lower (lower triangular matrix/half) and upper (upper triangular matrix/half) endpoint of 95% confidence intervals for each pairwise comparison of items with DIF detected at 5% significance by the individual tests of PISA questions in Table 1 (using the stratification after item purification) under the local independence assumption, i.e. item responses were obtained independently for each item. We find that for questions PS326Q01 and PS326Q02 the bootstrap p-value is exactly 1/10,000=0.0001, i.e. the bootstrap sampled odds ratio did not exceed the observed value in any sample. We conclude that these two items are not locally independent. Examining all odds ratios where we find a p-value of less than 0.05, we find that 9 of the 21 pairs (43% where this occurs are pairs of items belonging to the same sequence of questions. In total 37 of 595 pairs (6.2% of these pairs) of items belong to the same sequence. The over-representation of items which are part of the same parent question among items with significantly different odds ratios agrees with the theory that local item independence is less likely to hold when items share common material.

Finally, we compare the values for the global test W and the test W_{ind} in Table 3. While we have already established that the local independence assumption of the W_{ind} test statistic following a χ^2_m distribution has been violated, and have $m \geq 5$ meaning we cannot rely on the asymptotically distribution, we consider the comparison of results to be informative. We see that with either method of estimating the global test statistic W we reject the null hypothesis of no DIF in any of the test items. We would have reached the same conclusion with the naive test W_{ind} if we were able to apply it and reach the same value without violation of its assumptions.

	Value	p-value (Asym)	p-value (Boot)
W	115.56	1.533×10^{-10}	0.0001
W_{ind}	99.61	4.086×10^{-8}	0.0001

Table 3.

The global test W based on asymptotic distribution using the adjustment of $\widehat{\Sigma}$ (Asym) and the bootstrap method (Boot); the global test W_{ind} based on the sum of independent χ^2 statistics (Asym) and the bootstrap method (Boot).

5.2. *EAP Reading Tests*

The reading comprehension test with 15 multiple choice items was designed for an English for Academic Purposes program. Seventy-seven students took the reading test that determines whether students meet the English language proficiency criteria for university entry. The high-stakes nature of the in-house tests means that new versions are being designed, and existing versions are being revised and validated on an on-going basis based on the psychometric properties of the tests. We argue that DIF information is particularly relevant in the situations where the stakes are high and where the test designers, often teachers themselves, are closely involved in the test validation process.

It is of interest to investigate whether the test exhibits DIF with regard to gender. For example, DIF could be caused by questions which are favorable for female students. This example treats female students as the reference group and male students as the focal group. For the global DIF test, we considered $m = 15$ questions. Note that unlike in Subsection 5.1, assessing each question individually does not show significant evidence of DIF in any question (see Table 4). Therefore, there is no item purification applied.

We created $K = 14$ strata based on student's ability using the final score calculated by adding up 0 (incorrect answer) or 1 (correct answer) for all 39 questions. The stratum range is as follows: < 10.5, $10.5 - 12.5$, $12.5 - 14.5$, $\ldots$, > 34.5. We estimated the odd ratio in (16) to check the local item independence assumption. To test this, we consider the odds ratios $\{\Gamma_{j\ell|gk}\}$ (16) of different questions for Passage A. When the local item independence assumption holds, $\Gamma_{j\ell|gk} = 1$ for all $j < \ell = 2, \ldots, 15$, $g = 1, 2$, and $k = 1, \ldots, K$. For simplicity, we consider a common odds ratio $\Gamma_{j\ell}$ for all $g = 1, 2$ and $k = 1, \ldots, K$ using a MH type common odds ratio estimate of $\Gamma_{j\ell}$. To find p-values we used the bootstrap method based on 10,000 bootstrap samples under the local independence assumption, i.e. item responses were obtained independently for each question. Among $m \times (m - 1)/2 = 105$ pairs, three $\Gamma_{j\ell}$'s are significantly different from 1 with a maximum value of 7.75 and corresponding bootstrap p-value of 0.027 (unadjusted for multiple testing). The local item independence assumption is only slightly violated for this example.

Table 4 shows the γ estimates, their standard errors and the p-values based on the Z statistic (12) using $\gamma_0 = 0$ for individual items. When $\widehat{\gamma}$ value is large, the question shows a disadvantage for the focal (male) group based on the definition of $\widehat{\gamma}_j$ in (9). The smallest p-value is 0.03 (Q4), followed by 0.06 (Q7) and 0.07 (Q12).

Testing all $m = 15$ items jointly does not result in rejecting H_0. For example, using the Bonferroni method, a p-value of a particular item needs to be smaller than $\alpha/m = 0.05/15 = 0.003$ to be rejected, which is clearly not the case as the smallest p-value is only

0.03. We also applied a range of other multiple testing procedures, such as Holm (1979), Benjamini & Hochberg (1995) and Benjamini *et al.* (2006). See Hemmelmann *et al.* (2005) for an overview of common multiple testing methods. None of these standard methods provided significant results. Table 5 shows the global tests W and W_{ind} based on both the theoretical asymptotic distribution and the non-parametric bootstrap method. None of these tests found DIF at the 5% significance level. For this example, the bootstrap method is recommended due to the value of $m \geq 5$ as discussed in Section 4.

Table 6 shows the 95% confidence intervals comparing any two of these items, where the upper right corner shows the upper limit and the lower left corner shows the lower limit of the confidence interval of $\gamma_j - \gamma_\ell$. The pairs (Q2, Q4), (Q7, Q8), (Q7, Q12), (Q7, Q13) are significantly different at a 5% level.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
$\widehat{\gamma}_j$	−0.09	0.17	−0.12	−1.32	−0.44	−0.23	1.06	−0.57	0.05	−0.15	−0.48	−1.20	−0.83	−0.16	−0.17
s.e.($\widehat{\gamma}_j$)	0.59	0.53	0.61	0.62	0.69	0.55	0.56	0.51	0.51	0.56	0.52	0.66	0.58	0.55	0.57
p-value	0.88	0.75	0.85	0.03	0.52	0.67	0.06	0.26	0.92	0.79	0.36	0.07	0.15	0.77	0.76

Table 4.

MH Log odds ratio estimator $\widehat{\gamma}$ and their standard error (s.e.($\widehat{\gamma}$)), two-sided p-values based on Z statistic.

	Value	p-value (Asym)	p-value (Boot)
W	23.04	0.08	0.43
W_{ind}	16.70	0.34	0.29
LR	11.77	0.70	–

Table 5.

The global test W based on asymptotic distribution using the adjustment of $\widehat{\Sigma}$ (Asym) and the bootstrap method (Boot); the global test W_{ind} based on the sum of independent χ^2 statistics (Asym) and the bootstrap method (Boot); LR corresponds to the likelihood ratio test.

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15
Q1	–	1.4	1.5	2.9	2.2	1.7	0.3	2.0	1.5	1.7	1.9	3.0	2.4	1.7	1.5
Q2	−1.9	–	2.0	2.7*	2.0	1.8	0.7	2.3	1.7	2.0	2.1	3.1	2.7	1.8	1.9
Q3	−1.4	−1.4	–	3.0	2.2	1.8	0.3	1.8	1.3	1.6	2.0	3.3	2.5	1.7	1.5
Q4	−0.4	0.2*	−0.6	–	0.8	0.6	−0.7	0.8	0.4	0.5	0.7	1.7	1.1	0.5	0.6
Q5	−1.5	−0.8	−1.5	−2.6	–	1.3	0.1	2.0	1.2	1.4	1.6	2.6	1.9	1.6	1.7
Q6	−1.4	−1.0	−1.6	−2.8	−1.7	–	0.1	1.8	1.5	1.5	1.7	2.6	2.2	1.5	1.6
Q7	−2.6	−2.4	−2.6	−4.0	−3.1	−2.7	–	3.3*	2.5	2.9	3.3	4.1*	3.5*	2.7	2.8
Q8	−1.0	−0.8	−0.9	−2.3	−1.8	−1.1	0.0*	–	0.9	1.1	1.4	2.3	1.8	0.9	1.1
Q9	−1.8	−1.4	−1.6	−3.1	−2.2	−2.1	−0.5	−2.2	–	1.7	2.1	2.9	2.3	1.7	1.7
Q10	−1.6	−1.3	−1.5	−2.9	−2.0	−1.6	−0.4	−2.0	−1.3	–	1.8	2.7	2.1	1.4	1.3
Q11	−1.1	−0.8	−1.2	−2.3	−1.5	−1.2	−0.2	−1.6	−1.0	−1.1	–	2.3	1.9	1.3	1.2
Q12	−0.8	−0.3	−1.2	−2.0	−1.1	−0.6	0.4*	−1.1	−0.4	−0.6	−0.9	–	1.0	0.4	0.9
Q13	−0.9	−0.7	−1.1	−2.1	−1.1	−1.0	0.3*	−1.2	−0.5	−0.8	−1.2	−1.7	–	0.8	0.9
Q14	−1.6	−1.1	−1.6	−2.8	−2.1	−1.6	−0.3	−1.7	−1.3	−1.3	−1.9	−2.5	−2.2	–	1.5

Table 6.

Pairwise comparisons: lower (lower triangular matrix/half) and upper (upper triangular matrix/half) endpoint of 95% confidence intervals for each pairwise comparison. The significantly different pairs at a 5% level are indicated by *.

6. Conclusion

By using the GMH estimators described, we are able to estimate the covariance of the common log odds ratio estimator across multiple dependent questions. This allows for simultaneous statistical tests for DIF when the local item independence assumption does not hold. This paper compares two global test statistics W and W_{ind} based on the theoretical asymptotic distribution and the non-parametric bootstrap method. When the simultaneous test involves less than 5 items, the proposed global test statistics W shows the best performance. If the number of items is large, the global test statistic W_{ind} is recommended in conjunction with the bootstrap method used to find the p-value of the test. The comparison to the standard LR test procedure illustrates the global test statistics are generally more suitable for cases with correlated items or highly sparse data.

We illustrate different tests using two examples. While this paper only compares two groups (reference and focal), the GMH estimators can be applied to compare more than two groups. Though there is still work to be done in determining a suitable multivariate analogue to the MH Delta–DIF scale.

It would be of interest to further investigate the performance of these estimators in different situations, such as testing these GMH estimation techniques against other simultaneous tests in highly sparse data cases. While the dual consistency of the GMH estimators provides a motivation for their use in sparse data cases, we are unaware of how they may perform compared to other simultaneous tests (e.g. SIBTEST) in these situations.

Further generalization of these GMH techniques to ordinal response data should also be possible. The Liu–Agresti estimator (Liu & Agresti, 1996) provides a method for estimating a common cumulative odds ratio across strata for ordinal data in a $2 \times c \times K$ table. In the context of DIF this allows for the detection of DIF in polytomous items i.e. questions with ordinal responses (Penfield & Algina, 2003). A combination of the Liu–Agresti estimator and the GMH techniques would allow for estimation of the covariance of the cumulative common odds ratio estimator between answers to different polytomous items.

Finally, the R code to calculate the GMH estimators and their (co)variance estimators is available on Github. Note: Due to the blinded manuscript requirement, the link is not provided for this version.

Acknowledgements

We would like to thank the reviewers for highly valuable comments.

References

Agresti, A. & Liu, I. (2001). Strategies for modeling a categorical variable allowing multiple category choices. *Sociological Methods & Research*, 29(4), 403–434.

Agresti, A. & Liu, I.-M. (1999). Modeling a categorical variable allowing arbitrarily many category choices. *Biometrics*, 55(3), 936–943.

Andersen, E. B. (1980). *Discrete statistical models with social science applications*. Amsterdam: North-Holland Publishing Company.

Baghaei, P. (2008). Local dependency and rasch measures. *Rasch measurement transactions*, 21(3), 1105–1106.

Barthélemy, J. & Suesse, T. (2018). mipfp: An R package for multidimensional array fitting and simulating multivariate bernoulli distributions. *Journal of Statistical Software*, 86, 1–20.

Bates, D. & Maechler, M. (2021). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.3-2. `https://CRAN.R-project.org/package=Matrix`.

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 (Methodological)*, 57(1), 289–300.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3), 491-507.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A bayesian random effects model for testlets. *Psychometrika*, 64(2), 153–168.

Breslow, N. (1981). Odds ratio estimators when the data are sparse. *Biometrika*, 68(1), 73–84.

Camilli, G., Shepard, L. A., & Shepard, L. (1994). *Methods for identifying biased test items*, volume 4. Thousand Oaks, California: Sage.

Chen, Y.-F. & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 reading assessment. *Educational Assessment*, 19(2), 77–96.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yens Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, 41(3), 178–194.

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.

Dorans, N. J. (1989). Two new approaches to assessing differential item functioning: Standardization and the Mantel–Haenszel method. *Applied Measurement in Education*, 2(3), 217–233.

Dorans, N. J. & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and Standardization. *ETS Research Report Series*, 1992(1), 1–40.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33(4), 465–484.

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, 23(1), 138-149.

Efron, B. & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Boca Raton, FL: Chapman & Hall.

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.

Greenland, S. (1989). Generalized Mantel-Haenszel estimators for K $2 \times J$ tables. *Biometrics*, 45(1), 183–191.

Hemmelmann, C., Horn, M., Süsse, T., Vollandt, R., & Weiss, S. (2005). New concepts of multiple tests and their use for evaluating high-dimensional EEG data. *Journal of Neuroscience Methods*, 142(2), 209–217.

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). Hillsdale, NJ: Erlbaum.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.

Ip, E. H.-s. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, 66(1), 109–132.

Khorramdel, L., Pokropek, A., Joo, S.-H., Kirsch, I., & Halderman, L. (2020). Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach. *Psychological Test and Assessment Modeling*, 62(2), 179–231.

Kim, J. & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, 73, 458–470.

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

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing*, 9(2), 122–133.

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing*, 21(1), 74–100.

Li, Y., Bolt, D. M., & Fu, J. (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement*, 30(1), 3–21.

Liu, I. & Agresti, A. (1996). Mantel-Haenszel-type inference for cumulative odds ratios with a stratified ordinal response. *Biometrics*, 52(4), 1223–1234.

Liu, I. & Suesse, T. (2008). The analysis of stratified multiple responses. *Biometrical Journal*, 50(1), 135–149.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. New York, NY: Routledge, Taylor and Francis Group, 1st edition.

Loughin, T. M. & Scherer, P. N. (1998). Testing for association in contingency tables with multiple column responses. *Biometrics*, 54(2), 630–637.

Magis, D., Beland, S., Raiche, G., & Magis, M. D. (2020). Package difR. `http://freebsd.yz.yamagata-u.ac.jp/pub/cran/web/packages/difR/difR.pdf`.

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.

Magis, D., Tuerlinckx, F., & Boeck, P. D. (2015). Detection of differential item functioning using the Lasso approach. *Journal of Educational and Behavioral Statistics*, 40(2), 111–135.

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.

Millsap, R. E. & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17(4), 297–334.

Organisation for Economic Co-operation and Development (OECD) (2012a). Data Base-PISA 2012.

Organisation for Economic Co-operation and Development (OECD) (2012b). Pisa 2012 Technical Report. `https://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm`.

Organisation for Economic Co-operation and Development (OECD) (2021). PISA 2012. DOI: http://dx.doi.org/10.34740/KAGGLE/DSV/2329039.

Osterlind, S. J. & Everson, H. T. (2009). *Differential Item Functioning*. Thousand Oaks, California: SAGE Publications, Inc, 2nd edition.

Paek, I. (2012). A note on three statistical tests in the logistic regression DIF procedure. *Journal of Educational Measurement*, 49(2), 121–126.

Paek, I. & Holland, P. (2015). A note on statistical hypothesis testing based on log transformation of the Mantel-Haenszel common odds ratio for differential item functioning classification. *Psychometrika*, 80(2), 406–411.

Park, S. E., Ahn, S., & Zopluoglu, C. (2021). Differential item functioning effect size from the multigroup confirmatory factor analysis for a meta-analysis: A simulation study. *Educational and Psychological Measurement*, 81(1), 182–199.

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel-Haenszel procedures.*Applied Measurement in Education*, 14, 235–259.

Penfield, R. D. & Algina, J. (2003). Applying the Liu-Agresti estimator of the cumulative common odds ratio to DIF detection in polytomous items. *Journal of Educational Measurement*, 40(4), 353–370.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 4 (pp. 321–333).

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional IRT model. *Journal of Educational Measurement*, 47(3), 361–372.

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

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 1(1), 43–46.

Roussos, L. A. & Stout, W. F. (1996). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement*, 33(2), 215–230.

Sauder, D. & DeMars, C. (2020). Applying a multiple comparison control to IRT item-fit testing. *Applied Measurement in Education*, 33(4), 362–377.

Schervish, M. J. (1996). P values: what they are and what they are not. *The American Statistician*, 50(3), 203–206.

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.

Socha, A., DeMars, C. E., Zilberberg, A., & Phan, H. (2015). Differential item functioning detection with the Mantel-Haenszel procedure: The effects of matching types and other factors. *International Journal of Testing*, 15(3), 193–215.

Swaminathan, H. & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27(4), 361–370.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics*, 27, 77–83.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: Cambridge University Press.

Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Frontiers in Psychology*, 5, 438.

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

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

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Boca Raton, FL: CRC Press, 2nd edition.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Ye, L., & Isham, S. (2012). Improving Mantel-Haenszel DIF estimation through Bayesian updating. *Journal of Educational and Behavioral Statistics*, 37(5), 601–629.