Article

A Monte Carlo Study of an Iterative Wald Test Procedure for DIF Analysis

Educational and Psychological Measurement
2016, Vol. 77(1) 104–118

DOI: 10.1177/0013164416637104
journals.sagepub.com/home/epm
SAGE

Mengyang Cao[1], Louis Tay[2], and Yaowu Liu[2]

Abstract

This study examined the performance of a proposed iterative Wald approach for detecting differential item functioning (DIF) between two groups when preknowledge of anchor items is absent. The iterative approach utilizes the Wald-2 approach to identify anchor items and then iteratively tests for DIF items with the Wald-1 approach. Monte Carlo simulation was conducted across several conditions including the number of response options, test length, sample size, percentage of DIF items, DIF effect size, and type of cumulative DIF. Results indicated that the iterative approach performed well for polytomous data in all conditions, with well-controlled Type I error rates and high power. For dichotomous data, the iterative approach also exhibited better control over Type I error rates than the Wald-2 approach without sacrificing the power in detecting DIF. However, inflated Type I error rates were found for the iterative approach in conditions with dichotomous data, noncompensatory DIF, large percentage of DIF items, and medium to large DIF effect sizes. Nevertheless, the Type I error rates were substantially less inflated in those conditions compared with the Wald-2 approach.

Keywords

differential item functioning, Monte Carlo simulation, item response theory, iterative Wald test

In educational and psychological research, it is important to establish the measurement equivalence (ME) of assessment tools across different groups. This is because

[1]University of Illinois at Urbana-Champaign, Champaign, IL, USA
[2]Purdue University, West Lafayette, IN, USA

Corresponding Author:
Mengyang Cao, Department of Psychology, University of Illinois at Urbana-Champaign, 603 E. Daniel Street, Champaign, IL 61820, USA.
Email: mcao3@illinois.edu

ME ensures that mean comparisons between groups reflect latent trait differences rather than measurement bias (Vandenberg & Lance, 2000). Within the item response theory (IRT) framework, ME is often evaluated by examining differential item functioning (DIF), which tests whether or not the relationship between latent trait and observed scores is identical between a reference group and a focal group (Drasgow, 1984). Because of the importance of the issue, numerous procedures have been proposed for assessing DIF, including parametric and nonparametric approaches (Raju, van der Linden, & Fleer, 1995; Sijtsma, 1998; Tay, Meade, & Cao, 2015), and researchers are continually seeking to improve on DIF statistics and procedures.

Based on one of the earliest proposed DIF statistic—the Lord's (1980) Wald χ^2 test—a recent implementation of the Wald χ^2 test has been shown to be a viable option for assessing DIF (Langer, 2008; Woods, Cai, & Wang, 2013). The implementation of the Wald χ^2 test has been shown to have Type I error rates close to nominal Type I error rates and substantial power. However, when anchor items are unknown, the procedure used for the Wald χ^2 test produces inflated Type I error rates (Woods et al., 2013). Given that anchor items are often unknown in practice (Woods, 2008, 2009), there is a critical need to ensure accurate DIF detection (i.e., low Type I error rates and high power). To address this issue, an iterative approach founded upon the prior approach (Cai, Thissen, & du Toit, 2011; Woods et al., 2013) was suggested and illustrated by Tay et al. (2015). However, the performance of the iterative approach has not been rigorously examined using Monte Carlo simulations. Therefore, in this article, we conducted Monte Carlo simulations comparing the iterative approach to the prior approach in both dichotomous and polytomous data.

Wald-1 and Wald-2 Approaches

According to Lord (1980), the Wald χ^2 statistic used for comparing the item parameters of two groups is computed as

$$\chi^2 = (v_{\mathrm{R}} - v_{\mathrm{F}})^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathrm{R}} + \boldsymbol{\Sigma}_{\mathrm{F}})^{-1}(v_{\mathrm{R}} - v_{\mathrm{F}}), \quad (1)$$

where v_R, v_F represent the vectors of the maximum likelihood item parameter estimators of the reference group and the focal group, and $\boldsymbol{\Sigma}_{\mathrm{R}}$, $\boldsymbol{\Sigma}_{\mathrm{F}}$ denote the asymptotic variance and covariance matrices for v_{R} and v_{F}, respectively. The test statistic is then compared with a critical value in a χ^2 distribution, with the degrees of freedom equal to the number of item parameters in the IRT model.

There are however several issues with using the Lord's Wald χ^2 statistic for DIF analysis. First, in order to use Equation (1) to generate the test statistic, one needs to first perform linking to place the item parameters separately estimated in the two groups on the same scale. An iterative linking approach has been proposed to increase the accuracy of the linking procedure (Stocking & Lord, 1983), but in practice, the implementation is cumbersome as different statistical programs are needed to scale and rescale the item parameters for testing DIF (e.g., Stark, 2002). Second, and more

important, the Lord's Wald χ^2 test does not have accurate standard error estimates, which results in inaccuracies in statistical testing (Langer, 2008; McLaughlin & Drasgow, 1987).

To address these issues, Langer (2008) improved on the Lord's Wald χ^2 test by employing the supplemented expectation maximization algorithm to obtain more accurate standard error estimates (Meng & Rubin, 1991). Langer (2008) also introduced a two-stage procedure to replace ad hoc linking with concurrent calibration (Kolen & Brennan, 2004). Specifically, the first step is to estimate the latent trait parameters of the focal group, while fixing the mean and *SD* of the reference group to be 0 and 1, respectively, and constraining the item parameters to be equal between the two groups. The second step is to fix the latent trait distribution of the focal group at the values obtained in the first step, and freely estimate the item parameters of the two groups. The item parameters found in the second step can then be used to compute the chi-square test statistic with Equation (1), as the item parameters are now placed on the same scale. This is known as the Wald-2 approach (Woods et al., 2013).

However, a potential problem associated with the Wald-2 approach is that the latent trait distribution estimated in the first stage occurs under the assumption that overall there is no DIF at the scale level. If this assumption does not hold, the latent trait estimation of the focal group will likely be biased, which may lead to inaccurate DIF detection (Tay et al., 2015). Such proposition was supported by a recent simulation study, which showed that the Wald-2 approach led to unacceptably high Type I error rates in almost all DIF detection conditions (Woods et al., 2013).

Based on their simulation results, Woods et al. (2013) recommended the Wald-1 approach for detecting DIF (Cai et al., 2011), as it demonstrated superior performance over the Wald-2 approach in terms of Type I error rate and power. Unlike the Wald-2 approach, the Wald-1 approach requires only one stage for testing DIF, assuming preknowledge of anchor items. Specifically, item parameters are scaled on the same metric by constraining the item parameters of anchor items to be equal between groups. All other items are freely estimated between groups and DIF for each item is tested using the Wald χ^2 statistic. Although the Wald-1 approach performs better, it requires preknowledge of anchor items. Therefore, in spite for the better performance, the Wald-1 approach is difficult to implement in cases where anchor items are unknown.

An Iterative Wald Approach

In order to accurately test for DIF without preknowledge of anchor items, an iterative approach was proposed by Tay et al. (2015). This approach is analogous to iterative linking (Stocking & Lord, 1983) in which non-DIF items are identified as anchor items and used to put the reference and focal groups on the same ''scale.'' The set of anchor items are further refined through an iterative procedure. Steps to implement this iterative approach are outlined below:

1. *The Wald-2 approach is used to identify anchor items*. In the presence of DIF, the Wald-2 approach has high Type I error rates, so that there is a high probability that non-DIF items are identified as having DIF. However, given that the Wald-2 approach has good power, when items do not have DIF, there is greater confidence that these items are likely to be non-DIF items, or anchor items. This is akin to a fully constrained baseline approach where all items are used for linking; and because of high Type I error rates, non-DIF items are subsequently suggested to be used as anchor items (Stark, Chernyshenko, & Drasgow, 2006).
2. *Using anchors, the Wald-1 approach is used to test for DIF in the remaining items*. The Wald-1 approach is known to have good Type I error rates and substantial power for detecting DIF when anchor items are known.
3. *Items that do not have significant DIF are added as anchor items*. Given that Type I error rates are well-controlled for the Wald-1 procedure, items that do not display DIF are assumed to be anchor items.
4. *If no new items are added in Step 3, the DIF procedure ends*. Otherwise, the procedure repeats from Step 2 onward. The procedure ends when all nonanchor items exhibit significant DIF based on the Wald χ^2 test.

The above iterative DIF detection approach utilizes the rationale of the iterative purification procedure (Candell & Drasgow, 1988; Lord, 1980), which involves iteratively relinking the metrics of the parameters and removing DIF items until the same DIF items are identified in two successive iterations. The iterative purification procedure has been adopted in many DIF detection methods, such as Wald test, Raju's area measure, and likelihood ratio test, and has demonstrated better performance than noniterative DIF detection methods (Candell & Drasgow, 1988; Cohen & Kim, 1993; Stark et al., 2006). Compared with the Lord's Wald test with iterative purification procedure, the iterative Wald approach outlined in this study is superior in that it concurrently calibrates the item parameters of the compared groups to avoid linking and relinking, and that it provides more desirable standard error estimates than the traditional Lord's Wald test (Langer, 2008).

In their review article on IRT DIF detection methods, Tay et al. (2015) provided illustrations on using the iterative Wald approach to detect DIF items in both dichotomous and polytomous data. They found that the iterative approach was successful in detecting the DIF items as simulated. However, the illustrations were only based on two simulated samples (i.e., dichotomous and polytomous responses), and thus were unable to provide information about how well the iterative Wald approach would perform in detecting DIF across a variety of conditions. It is important to rigorously examine this using Monte Carlo simulations.

The Current Study

Our interest is to examine the case where preknowledge of anchor items is absent; thus, the Wald-1 approach is not applicable. In this case, would the iterative approach

improve on the previously examined Wald-2 approach? We conducted a simulation study to determine whether the iterative approach would successfully reduce the Type I error rates compared with the Wald-2 approach, while maintaining a decent level of power in detecting DIF.

There are several factors we are interested in. First, different research design factors may affect the performance of the iterative approach. For example, Woods et al. (2013) found that sample size, ratio of DIF cases, and percentage of DIF items would influence the Type I error rate of the Wald-2 method. Second, DIF effect size could also be a potentially influential factor, as Langer (2008) simulated smaller DIF effect size than Woods et al. (2013) and did not find severely inflated Type I error rate of Wald-2. Third, the type of cumulative DIF is another factor that is often overlooked in simulation studies. Compensatory DIF describes that the direction and effect size of DIF items compensate with each other at the scale level, whereas noncompensatory DIF denotes that DIF does not cancel out at the scale level (Raju et al., 1995). Given that the first stage of both the Wald-2 and the iterative approach (given that it uses the Wald-2 as the first step) may lead to biased estimates of the focal group trait distribution when there is cumulative DIF at the scale level, it may be that compensatory DIF would reduce the problem of inflated Type I error rates. On the other hand, noncompensatory DIF may lead to high cumulative DIF at the scale level leading to more problems with correctly detecting DIF.

Method

Overview of Study Design

We conducted a Monte Carlo simulation study to compare the performance of the Wald-2 approach and the iterative Wald approach by simulating different conditions on the following factors:

1. Number of response categories (dichotomous, polytomous)
2. Test length (number of items = 15, 30)
3. Sample size in reference (R) and focal (F) groups (equal smaller [R = 500; F = 500], equal larger [R = 1,000; F = 1,000], unequal smaller [R = 750; F = 250], unequal larger [R = 1,500; F = 500])
4. Percentage of DIF items (20%, 40%)
5. DIF effect size (0.3, 0.5, 0.7)
6. Type of cumulative DIF (compensatory, noncompensatory)

Altogether, there were $2 \times 2 \times 4 \times 2 \times 3 \times 2 = 192$ conditions. A total of 500 replications were undertaken for each condition. We designed R scripts to automate the simulation process.

Data Generation

We followed the simulation procedure in Woods et al. (2013) to generate response data. Specifically, we drew the latent trait distribution of the reference group (i.e., θ_R) from $N(0, 1)$, and the latent trait distribution of the focal group (i.e., θ_F) from $N(-0.6, 1)$, with different sample sizes indicated in the section above.

Consistent with Woods et al. (2013), we used Samejima's graded response model (SGRM; Samejima, 1969) to generate polytomous response data. According to the SGRM, the probability of endorsing response option k on item i is given by

$$P[X_{ik} = 1|\theta_j] = \frac{1}{1 + \exp\left[-a_i\left(\theta_j - b_{ik}\right)\right]} - \frac{1}{1 + \exp\left[-a_i\left(\theta_j - b_{i(k+1)}\right)\right]}, \quad (2)$$

where θ_j denotes the latent trait of respondent j, a_i represent the discrimination parameter of item i, and b_{ik} refers to the threshold parameter for response option k on item i. In the current simulation, we drew the discrimination parameters of the reference group (i.e., $a_{i\mathrm{R}}$) from a $U[0.5 \times 1.7, 0.8 \times 1.7]$ distribution. The four threshold parameters of each item of the reference group (i.e., $b_{i1\mathrm{R}} - b_{i4\mathrm{R}}$) were drawn from $U[-2, -1]$, $U[-1, 0]$, $U[0, 1]$, and $U[1, 2]$, respectively.

Unlike Woods et al. (2013), which only examined polytomous responses, we also simulated dichotomous data with the two-parameter logistic model (2PLM), which has been commonly used as an IRT model for noncognitive individual difference variables (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001; O'Brien & LaHuis, 2011). The 2PLM is stated as

$$P[X_i = 1|\theta_j] = \frac{1}{1 + \exp\left[-a_i\left(\theta_j - b_i\right)\right]}, \quad (3)$$

where θ_j denotes the latent trait of respondent j, and a_i and b_i refer to the discrimination and difficulty parameters of item i. In the present simulation, we drew the discrimination parameters of the reference group (i.e., $a_{i\mathrm{R}}$) from a $N(1.7, 0.6)$ distribution, with truncation on the upper end at 4.0, and on the lower end at 0.8. The location parameters of the focal group were drawn from a $U[-2, 2]$ distribution.

DIF Simulation and Detection

To simulate noncompensatory DIF effects, we randomly selected 20% or 40% of the items as DIF items, and added a constant δ to all the parameters of each DIF item. The constant δ was set to equal 0.3, 0.5, or 0.7 for different effect size conditions. For conditions with compensatory DIF effects, we randomly selected 20% or 40% of the items as DIF items. For each DIF item, we either added or subtracted the same constant δ for each parameter (i.e., a_i and b_{ik}). The decision to add or subtract was made independently for each parameter with equal probability (i.e., P[addition] = P[subtraction] = .5). The size of the DIF effects and the choice of adding or subtracting a common constant to all the parameters of each DIF items followed the

procedure by Woods et al. (2013). The difference is that in the Woods et al. (2013) study, the different effect sizes are independently manipulated for the item slope (i.e., a_i) and the item locations (i.e., b_{ik}).

After the response data were generated, two methods were used to detect DIF. The Wald-2 method was implemented in IRTPRO (Cai et al., 2011). For the iterative DIF detection method, we followed the aforementioned four steps, using the non-DIF items found with the Wald-2 method as anchor items, and iteratively performed the Wald-1 procedure to detect DIF in the remaining items.

Criteria

For each simulation conditions, we evaluated the performance of the two DIF detection methods in terms of Type I error rate and statistical power. Type I error rate was computed as the number of non-DIF items incorrectly identified as DIF items divided by the total number of non-DIF items in the scale. Statistical power was calculated as the number of DIF items correctly detected by each method divided by the total number of DIF items in the scale. Type I error rates well-controlled at the nominal Type I error rate of .05 and higher statistical power indicate better performance of a DIF detection method.

Results

Type I Error Rate

Table 1 presents the Type I error rate for polytomous data. In general, the iterative method (on the left; $M = .04$) led to lower Type I error rates than did the Wald-2 method (on the right; $M = .08$). Specifically, the Type I error rates of the iterative method were close to the nominal rate of .05 for all conditions, whereas the Wald-2 method resulted in Type I error rates above .05 for most conditions, consistent with results in previous simulation studies (Woods et al., 2013). Moreover, the Wald-2 method appeared to generate more Type I errors when there were more DIF items in the scale, when DIF effect size was larger, or when noncompensatory DIF rather than compensatory DIF was simulated. In contrast, the Type I error rates of the iterative method were always well controlled regardless of those factors.

As shown in Table 2, the iterative method ($M = .07$) also produced substantially lower Type I error rates—and closer to the nominal Type I error rate—for dichotomous data than did the Wald-2 method ($M = .15$). Unlike the results for the polytomous case, the iterative method seemed to be influenced by a few factors in the dichotomous case, such that the Type I error rates were above the nominal rate of .05 in several conditions. Type I error rates slightly increased as sample size increased. Even so, the Type I error rates of the iterative method were below .10 on most occasions except when noncompensatory DIF was simulated, and when there were 40% DIF items with medium (i.e., 0.5) or large (i.e., 0.7) effect sizes. The Wald-2 method, however, led to Type I error rates over .05 in most conditions. Similar to the results

Table 1. Type I Error Rates for Polytomous Data.

Scale length	Iterative Wald approach								Wald-2 approach							
	15 items				30 items				15 items				30 items			
NR/NF	500/ 500	750/ 250	1,000/ 1,000	1,500/ 500	500/ 500	750/ 250	1,000/ 1,000	1,500/ 500	500/ 500	750/ 250	1,000/ 1,000	1,500/ 500	500/ 500	750/ 250	1,000/ 1,000	1,500/ 500
Compensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.04	.04	.03	.04	.04	.04	.03	.04	.04	.04	.05	.04	.04	.05	.04	.04
DIF size = 0.5	.04	.04	.04	.04	.03	.04	.03	.04	.05	.05	.07	.06	.05	.05	.06	.05
DIF size = 0.7	.04	.04	.03	.04	.03	.04	.03	.04	.07	.06	.12	.08	.05	.05	.07	.06
% DIF = 0.4																
DIF size = 0.3	.04	.04	.03	.04	.03	.04	.03	.04	.05	.05	.06	.05	.04	.05	.05	.05
DIF size = 0.5	.03	.04	.03	.04	.03	.04	.03	.04	.07	.06	.11	.08	.06	.05	.08	.07
DIF size = 0.7	.03	.04	.03	.03	.03	.03	.03	.03	.13	.08	.21	.14	.08	.06	.13	.09
Noncompensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.03	.04	.03	.04	.03	.04	.03	.03	.04	.05	.05	.04	.05	.05	.06	.05
DIF size = 0.5	.03	.04	.03	.04	.03	.04	.03	.04	.05	.05	.06	.06	.05	.06	.06	.07
DIF size = 0.7	.03	.03	.03	.04	.03	.04	.03	.04	.05	.05	.06	.06	.05	.06	.07	.07
% DIF = 0.4																
DIF size = 0.3	.03	.04	.03	.04	.03	.04	.03	.04	.06	.06	.11	.10	.07	.07	.12	.11
DIF size = 0.5	.03	.03	.03	.04	.03	.04	.03	.04	.11	.09	.18	.16	.11	.11	.21	.18
DIF size = 0.7	.03	.04	.03	.04	.03	.04	.03	.04	.09	.10	.18	.19	.12	.11	.21	.22

Note. DIF = differential item functioning; NR = number of respondents in reference group; NF = number of respondents in focal group; % DIF = percentage of DIF items.

Table 2. Type I Error Rates for Dichotomous Data.

Scale length	Iterative Wald approach								Wald-2 approach							
	15 items				30 items				15 items				30 items			
NR/NF	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500
Compensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.04	.04	.08	.07	.05	.05	.09	.08	.06	.05	.10	.08	.06	.06	.11	.09
DIF size = 0.5	.05	.05	.08	.07	.05	.05	.08	.07	.07	.07	.12	.10	.06	.07	.11	.09
DIF size = 0.7	.05	.04	.07	.07	.05	.05	.08	.07	.08	.07	.15	.13	.07	.07	.13	.11
% DIF = 0.4																
DIF size = 0.3	.05	.05	.07	.07	.05	.05	.08	.08	.07	.06	.10	.09	.06	.06	.14	.12
DIF size = 0.5	.05	.06	.07	.07	.05	.06	.05	.05	.09	.08	.15	.13	.08	.07	.10	.08
DIF size = 0.7	.05	.05	.08	.06	.04	.06	.07	.08	.12	.10	.24	.17	.09	.07	.18	.15
Noncompensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.02	.02	.02	.03	.02	.02	.02	.02	.03	.03	.06	.05	.03	.03	.03	.04
DIF size = 0.5	.02	.03	.02	.03	.02	.02	.02	.02	.07	.05	.17	.11	.05	.04	.09	.08
DIF size = 0.7	.02	.03	.02	.03	.02	.02	.02	.02	.15	.10	.33	.23	.10	.07	.21	.17
% DIF = 0.4																
DIF size = 0.3	.05	.05	.08	.09	.04	.05	.07	.07	.07	.06	.18	.13	.05	.06	.11	.10
DIF size = 0.5	.11	.11	.19	.17	.10	.10	.16	.14	.23	.17	.55	.40	.19	.15	.38	.32
DIF size = 0.7	.21	.21	.24	.27	.19	.15	.28	.21	.51	.34	.76	.66	.41	.30	.71	.59

Note. DIF = differential Item Functioning; NR = number of respondents in reference group; NF = number of respondents in focal group; % DIF = percentage of DIF items.

with polytomous data, the Type I error rates of the Wald-2 method increased as percentage of DIF items and DIF effect sizes increased. Compared with the iterative method, the Wald-2 method had substantially worse performance in controlling Type I error rates when examining noncompensatory DIF with medium or large DIF effect sizes.

Power

As presented in Table 3, the power of detecting DIF is close to 1.0 in most conditions with polytomous data, suggesting that both the iterative and the Wald-2 methods can successfully detect DIF, regardless of sample size, percentage of DIF items, and type of cumulative DIF. The only conditions that possessed nonperfect power were those with small DIF effect sizes (i.e., 0.3). There was no noticeable difference in power between the two DIF detection methods (both *M*s = .98).

Table 4 shows the DIF detection power for dichotomous data. In general, both the iterative method and the Wald-2 method exhibited low power in most conditions compared with results for polytomous data, especially when DIF effect sizes were small and when the sample sizes of reference and focal groups were unbalanced. Comparing the two DIF detection methods, the power of the iterative method (M = .62) was slightly lower than that of the Wald-2 method (M = .65). However, the differences in power were always below .05 and were negligible in most condition considering the relatively large values of power.

Discussion

An iterative approach has been proposed to improve the Wald-2 approach in detecting DIF when preknowledge of anchor items is absent (Tay et al., 2015). In the present study, we conducted Monte Carlo simulation to compare the performance of the proposed iterative approach with the Wald-2 approach (Langer, 2008). In general, the iterative Wald approach exhibited satisfactory performance in DIF detection. The Type I error rates of the iterative approach were well-controlled in most conditions, indicating a substantial improvement over the Wald-2 approach. Moreover, the reduction in Type I error rates was not at the expense of sacrificing the power in detecting DIF.

The performance of the iterative approach was also found to be influenced by several factors. Foremost, the number of response was an influential factor. For polytomous data, the iterative approach exhibited low Type I error rates (i.e., below the nominal rate of .05) and high power (i.e., close to unity) in all conditions, whereas for dichotomous data, inflated Type I error rates and low power were found in a few conditions. This shows that iterative approach works better in the polytomous condition than in the dichotomous condition. Additionally, as we expected, the iterative approach performs poorly when a noncompensatory DIF is present, as it leads to inaccurate identification of anchor items in the first stage. An important caveat is that

Table 3. Power for Polytomous Data.

Scale length	Iterative Wald approach								Wald-2 approach							
	15 items				30 items				15 items				30 items			
NR/NF	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500
Compensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.94	.84	1.00	.99	.95	.87	1.00	1.00	.94	.85	1.00	.99	.95	.88	1.00	1.00
DIF size = 0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DIF size = 0.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
% DIF = 0.4																
DIF size = 0.3	.93	.82	1.00	.99	.95	.87	1.00	1.00	.94	.83	1.00	.99	.96	.88	1.00	1.00
DIF size = 0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DIF size = 0.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Noncompensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.94	.88	1.00	1.00	.96	.90	1.00	1.00	.94	.89	1.00	1.00	.96	.90	1.00	1.00
DIF size = 0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DIF size = 0.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
% DIF = 0.4																
DIF size = 0.3	.92	.86	1.00	.99	.93	.88	1.00	.99	.92	.86	1.00	.99	.93	.88	1.00	.99
DIF size = 0.5	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DIF size = 0.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Note. DIF = differential item functioning; NR = number of respondents in reference group; NF = Number of respondents in focal group; % DIF = percentage of DIF items.

Table 4. Power for Dichotomous Data.

Scale length	Iterative Wald approach								Wald-2 approach							
	15 items				30 items				15 items				30 items			
NR/NF	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500	500/500	750/250	1,000/1,000	1,500/500
Compensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.31	.25	.54	.45	.35	.29	.62	.52	.34	.27	.57	.47	.38	.30	.64	.54
DIF size = 0.5	.64	.54	.81	.78	.67	.58	.86	.80	.67	.57	.82	.80	.69	.60	.87	.82
DIF size = 0.7	.81	.74	.91	.88	.85	.78	.94	.91	.83	.76	.92	.89	.86	.79	.95	.91
% DIF = 0.4																
DIF size = 0.3	.31	.24	.49	.46	.33	.27	.84	.78	.36	.28	.55	.49	.37	.29	.86	.80
DIF size = 0.5	.59	.51	.78	.73	.66	.57	.82	.77	.65	.55	.81	.76	.70	.60	.84	.78
DIF size = 0.7	.75	.68	.89	.85	.83	.57	.93	.90	.80	.73	.91	.88	.84	.60	.94	.91
Noncompensatory DIF																
% DIF = 0.2																
DIF size = 0.3	.37	.23	.68	.58	.32	.22	.57	.46	.40	.24	.71	.59	.34	.23	.59	.49
DIF size = 0.5	.72	.57	.90	.87	.68	.59	.87	.81	.76	.60	.93	.88	.70	.61	.88	.83
DIF size = 0.7	.86	.72	.96	.92	.84	.74	.95	.92	.89	.75	.98	.93	.86	.77	.96	.93
% DIF = 0.4																
DIF size = 0.3	.22	.15	.49	.37	.22	.15	.41	.32	.24	.16	.53	.39	.23	.16	.43	.34
DIF size = 0.5	.49	.35	.73	.66	.49	.39	.70	.63	.53	.38	.81	.71	.52	.41	.72	.67
DIF size = 0.7	.64	.47	.80	.74	.64	.57	.82	.78	.72	.54	.90	.83	.68	.60	.87	.83

Note. DIF = differential item functioning; NR = number of respondents in reference group; NF = number of respondents in focal group; %DIF = percentage of DIF items.

the performance of the iterative approach was found to be worse for noncompensatory DIF than for compensatory DIF only when a large percentage of DIF items was simulated.

In general, our simulation results suggests that the iterative approach performs better than the Wald-2 approach in absence of anchor items for both the polytomous and dichotomous conditions. Practically, the iterative approach can also be easily and efficiently implemented with modern IRT software, such as IRTPRO (Cai et al., 2011) and flexMIRT (Cai, 2012). Readers are referred to Tay et al. (2015) for several illustrations of the iterative approach.

Limitations and Future Research

There were several limitations to this study. First, we only focused on DIF analysis between two groups, which is the most frequently observed scenario in DIF studies (Tay et al., 2015). As shown in Woods et al. (2013), the Wald test can be easily extended to detect DIF in more than two groups by utilizing the generalization of Lord's statistics (Kim, Cohen, & Park, 1995). Future simulation studies can focus on investigating the performance of the new iterative approach in detecting DIF among three or more groups. Second, to keep the simulation study manageable, we fixed the mean difference in latent trait distribution at 0.6, and did not examine the magnitude of mean difference as a potential factor. This is because past research has shown that the Wald test and associated approaches seems to be robust to varying levels of latent mean differences. Third, we did not examine the iterative Wald approach with respect to other DIF procedures as we wanted to focus on whether the iterative Wald approach improves on the Wald-2 approach. Future research can examine whether the iterative Wald approach fares as well compared with other DIF procedures in which the pre-knowledge of anchor items is unknown.

Conclusion

To conclude, the simulation study shows that iterative Wald approach performs as well if not better than the Wald-2 approach. We encourage researchers to use this procedure as compared with the Wald-2 approach when anchor items are unknown. Future research can also build on iterative procedure to potentially improve on it.

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement*, *12*, 253-260.

Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. A. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*, 523-562.

Cohen, A. S., & Kim, S.-H. (1993). A comparison of Lord's chi-square and Raju's area measures in detection of DIF. *Applied Psychological Measurement*, *17*, 39-52.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables. *Psychological Bulletin*, *95*, 134-135.

Kim, S.-H., Cohen, A. S., & Park, T.-H. (1995). Detection of differential item functioning in multiple groups. *Journal of Educational Measurement*, *32*, 261-276.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York, NY: Springer.

Langer, M. (2008). *A reexamination of Lord's Wald test for differential item functioning using item response theory and modern error estimation* (Unpublished doctoral dissertation). University of North Carolina, Chapel Hill.

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

McLaughlin, M. E., & Drasgow, F. (1987). Lord's chi-square test of item bias with estimated and with known person parameters. *Applied Psychological Measurement*, *11*, 161-173.

Meng, X., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, *86*, 899-909.

O'Brien, E., & LaHuis, D. M. (2011). Do applicants and incumbents respond to personality items similarly? A comparison of dominance and ideal point response models. *International Journal of Selection and Assessment*, *19*, 109-118.

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*, 353-368.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. (*Psychometrika Monograph Supplement*, No. 17). Richmond, VA: Psychometric Society.

Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement*, *22*, 3-31.

Stark, S. (2002). ITERLINK: Iterative linking and pairwise DIF detection for the 3PL model using Lord's chi-square [Computer program]. Urbana-Champaign: Department of Psychology, University of Illinois at Urbana-Champaign.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292-1306. doi:10.1037/0021-9010.91.6.1292

Stocking, M., & Lord, F. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*, 201-210.

Tay, L., Meade, A. W., & Cao, M. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*, 3-46. doi: 10.1177/1094428114553062

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70. doi:10.1177/109442810031002

Woods, C. M. (2008). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57. doi:10.1177/0146621607314044

Woods, C. M. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*, 42-57.

Woods, C. M., Cai, L., & Wang, M. (2013). The Langer-improved Wald test for DIF testing with multiple groups: Evaluation and comparison to two-group IRT. *Educational and Psychological Measurement*, *73*, 532-547.