Educational and Psychological
Measurement
1–18
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164420925885
journals.sagepub.com/home/epm
SAGE

Differential Item Functioning Effect Size From the Multigroup Confirmatory Factor Analysis for a Meta-Analysis: A Simulation Study

Sung Eun Park[1], Soyeon Ahn[1] and Cengiz Zopluoglu[1]

Abstract

This study presents a new approach to synthesizing differential item functioning (DIF) effect size: First, using correlation matrices from each study, we perform a multigroup confirmatory factor analysis (MGCFA) that examines measurement invariance of a test item between two subgroups (i.e., focal and reference groups). Then we synthesize, across the studies, the differences in the estimated factor loadings between the two subgroups, resulting in a meta-analytic summary of the MGCFA effect sizes (MGCFA-ES). The performance of this new approach was examined using a Monte Carlo simulation, where we created 108 conditions by four factors: (1) three levels of item difficulty, (2) four magnitudes of DIF, (3) three levels of sample size, and (4) three types of correlation matrix (tetrachoric, adjusted Pearson, and Pearson). Results indicate that when MGCFA is fitted to tetrachoric correlation matrices, the meta-analytic summary of the MGCFA-ES performed best in terms of bias and mean square error values, 95% confidence interval coverages, empirical standard errors, Type I error rates, and statistical power; and reasonably well with adjusted Pearson correlation matrices. In addition, when tetrachoric correlation matrices are used, a meta-analytic summary of the MGCFA-ES performed well, particularly, under the condition that a high difficulty item with a large DIF was administered to a large sample size. Our result offers an option for synthesizing the magnitude of DIF on a flagged item across studies in practice.

[1]University of Miami, Coral Gables, FL, USA

Corresponding Author:
Sung Eun Park, University of Miami, 5202 University of Drive, Coral Gables, FL 33124-2040, USA.
Email: sxp372@miami.edu

Keywords

differential item functioning (DIF), Pearson correlation, adjusted Pearson correlation, tetrachoric correlation, meta-analysis, multigroup confirmatory factor analysis (MGCFA)

Fairness is a critical component of high-stakes testing from both ethical and legal standpoints. The Educational Testing Service (2014) considers a test to be ''fair if any group differences in performance are derived from construct-relevant sources of variance. The existence of group differences in performance does not necessarily make a test unfair, because the groups may differ on the construct being measured'' (p. 57). However, if an examinee's performance on any test item is affected by construct-irrelevant characteristics (e.g., male vs. female, students with disability vs. those without disability), a test or an item is said to be biased, providing an unfair advantage for one group over others.

Differential item functioning (DIF) is a statistical character of an item that can help detect whether a test contains a systematic bias based on construct-irrelevant characteristics. DIF displays the extent to which the performance on an item systematically differs by subgroups (Osterlind & Everson, 2009). The DIF is said to be present when examinees in subgroups (e.g., race/ethnicity, gender, socioeconomic status) with the same level of latent traits have different probabilities of correctly responding to a given item. Although two types of DIF can be identified in practice—uniform DIF (constant across ability levels) and nonuniform DIF (varying across ability level), the current study particularly focuses on a uniform DIF for a dichotomous item.

In the literature, various parametric and nonparametric statistical techniques and the associated effect-size measures detecting uniform DIF have been well documented. These are the Mantel–Haenszel (MH) procedure (Camilli & Shepard, 1994; Hidalgo et al., 2014; Holland & Thayer, 1988; Zwick et al., 2012), logistic regression (LR) modeling (Gómez-Benito et al., 2009; Hidalgo et al., 2014), item response theory (IRT)–based method (Oshima et al., 2015; Raju, 1988; Steinberg & Thissen, 2006), structural equation modeling (SEM; Bauer, 2017; Steinmetz et al., 2009; Woods & Grimm, 2011), and variations of the aforementioned techniques (Chang et al., 1995; Penfield, 2007; Walker, 2011).

While research on DIF detection procedures and the associated effect-size measures has been proliferating in the field, literature regarding the synthesis of DIF indicators is limited. To our knowledge, of the many DIF indices discussed above, only the MH and the LR models have been examined as an effect size indicator for the meta-analyses of DIF detection on an item (i.e., Koo, 2012; Koo et al., 2014; Van de Water, 2014). Koo (2012) suggested using the MH DIF index in meta-analyses, and conducted a simulation study that examined its performance. In 2014, Van de Water conducted a simulation study that compared the Type I error rates and statistical power of using the LR and the MH DIF indices in meta-analyses. He further examined the differential effects of other study characteristics, such as sample size, test length, and magnitude of DIF, on Type I error rates and statistical power between LR and MH in meta-analyses.

Researchers have increasingly used SEM approaches in detecting an item or a test displaying DIF among subgroups. The two most commonly used SEM approaches are the multiple indicator and multiple cause (MIMIC) and the multigroup confirmatory factor analysis (MGCFA) models. Based on the well-known parametric equivalence between the MIMIC and the IRT models (Muthén et al., 1991), Jin et al. (2012) have proposed the effect size measure for MIMIC (MIMIC-ES) as given by

$$\text{MIMIC} - \text{ES} = \frac{\tau_i - \beta_i}{\lambda_i} - \frac{\tau_i}{\lambda_i} = -\frac{\beta_i}{\lambda_i}, \tag{1}$$

where τ_i is the threshold, β_i is direct effect of the grouping as a dummy variable on the latent factor, and λ_i is the factor loading for *i*th item.

Similarly, given that the parameters estimated by the MGCFA model are equivalent to item parameters estimated by the IRT model (Stark et al., 2006), the *b*-parameter on the *i*th item can be written as

$$b_i = \frac{\tau_i}{\lambda_i}, \tag{2}$$

where τ_i is the threshold for the *i*th item.

With the parametric equivalence of difficulty parameters between the MGCFA and IRT models, the following effect size (MGCFA-ES) can be used as an indicator that quantifies the magnitude and direction of a uniform DIF on an item between subgroups for the MGCFA model as given by

$$\text{MGCFA} - \text{ES} = b_i^F - b_i^R = \frac{\tau_i^F}{\lambda_i^F} - \frac{\tau_i^R}{\lambda_i^R}, \tag{3}$$

where *F* and *R* are the focal and reference groups, respectively.

The Current Study

Given that the SEM approaches (either MIMIC or MGCFA) have been increasingly utilized, it is practically important to evaluate whether an effect size estimator derived from the SEM approaches can be used in meta-analyses. We found that studies do not always provide sample responses for each item, as would be required in a MIMIC approach. More often, studies provide correlation matrices among sample responses on items, making MGCFA-ES a more suitable and practical approach for meta-analyses. In particular, the current study assumes that two separate correlation matrices for focal and reference groups were reported in each study. From these correlation matrices, MGCFA-ES and its associated standard error can be estimated and then synthesized across studies to estimate the DIF on an item.

Specifically, the current study aims to examine the performance of MGCFA-ES as an effect-size in meta-analyses and evaluate it using a Monte Carlo simulation. In the simulation, the bias and mean square error (*MSE*) values, empirical Type I error rates, empirical statistical powers, coverage rates of 95% confidence intervals, and empirical standard errors are all evaluated as outcomes in relation to the following factors: (1) the type of correlation matrices, (2) the magnitudes of a DIF, (3) the level of an item difficulty, and (4) the sample size.

Method

For the simulation employed in the current study, it is assumed that six items are used to measure the underlying ability on a dichotomous scale, with 1 being a correct answer and 0 being an incorrect answer. Of the six items, it is assumed that one item displays DIF between the focal and reference groups with different magnitudes of difficulty ($b_F - b_R$).

Data Generation

Using the *sim* (*"irtoys"*) function available in the R Version 3.5.3 (R Core Team, 2019), the response patterns on six items for a total $N_R + N_F$ observations (i.e., N_R for reference and N_F for focal groups, respectively) for 30 studies included in meta-analyses were generated based on test-takers' ability level, which is assumed to be normally distributed with a mean of 0 and a standard deviation of 1 under the 1-PL (one-parameter logistic) model, where only *b*-parameters for the biased item are manipulated. For each of 30 included studies, we extracted three different types of correlation matrices from the response patterns of a total $N_R + N_F$ observations on six items, separately for the reference and focal groups. In addition, the threshold (τ_j) for each item was obtained from the proportion of correct answers, which is used as a mean in the MGCFA.

For each individual study, the MGCFA model (as shown in Figure 1) was fitted to each type of correlation matrices using the *cfa* (*"lavaan"*) function available in R Version 3.5.3 (R Core Team, 2019). The model was specified to be 1-PL by constraining all loadings, residuals, and thresholds to be constant for two groups, except the thresholds of a flagged item. The latent factor for the reference group was fixed to have a mean of 0 with a variance of 1, while the mean and variance of latent factor for the focal group were freely estimated (Millsap, 2012; Stark et al., 2006). The model parameters (i.e., loadings and thresholds) of MGCFA were converted to the item difficulty parameters using Equation 2, and the MGCFA-ES was computed via Equation 3 for each of the 30 included studies.

Manipulating Factors in the Simulation

Item Parameters. The values of *a*-parameters were fixed to 1.17 for all six items. The values of *b*-parameters for five unbiased items (Items 1-5) were normally distributed

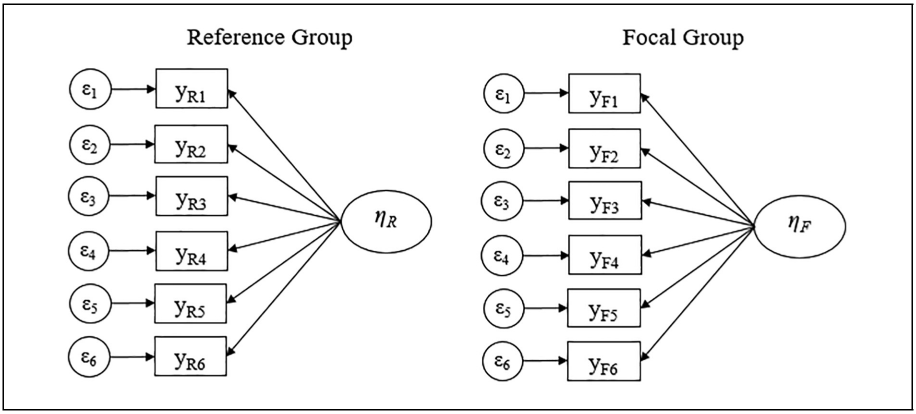


Figure 1. Specified multigroup confirmatory factor analysis (MGCFA) model for detecting differential item functioning (DIF) on one flagged item (Item 6).

with a mean of 0 and variance of 1, and three different conditions with low difficulty ($b = -1$), medium difficulty ($b = 0$), and high difficulty ($b = 1$) were manipulated for the biased item (Item 6), which was modified within the range obtained from previous studies (e.g., Jin et al., 2012).

DIF Magnitudes. Followed by a simulation study by Jin et al. (2012), DIF magnitude for the biased item was manipulated with four different conditions: no DIF ($b_F - b_R = 0$), small DIF ($b_F - b_R = .3$), medium DIF ($b_F - b_R = .5$), and large DIF ($b_F - b_R = .7$).

Sample Size. Three different sample size levels were generated, including small ($N_F = 200$, $N_R = 400$), medium ($N_F = 350$, $N_R = 700$), and large ($N_F = 500$, $N_R = 1{,}000$), which are reflective of real test settings by assigning the unbalanced sample sizes for the focal and reference groups (Jin et al., 2012).

Correlation Type. Three types of correlation matrices were extracted from item responses for focal and reference groups. In particular, the tetrachoric correlation has arisen as an alternative, since the Pearson product moment correlation is known to underestimate the true relationship between dichotomous items. Also, Fillmore et al. (1998) suggested transforming the Pearson correlation to a tetrachoric correlation by multiplying it by 3/2. Given that different correlation matrices can be used for MGCFA, the current study compared how the performance of MGCFA-ES for meta-analysis differs depending on the type of correlation (i.e., Pearson correlation, tetrachoric correlation, or adjusted Pearson correlation—Pearson correlation × 3/2 as suggested by Fillmore et al., 1998).

Summary. A total of 108 conditions were utilized in the current study, where MGCFA was fitted to three different correlation types generated from 36 item response patterns with 500 replications, totaling 54,000 data points (i.e., 108 × 500 replications).

Meta-Analytic Estimator of MGCFA

Figure 1 shows the MGCFA model for the current simulation study. Once MGCFA-ES and its associated standard errors are computed from each study, the population magnitude of DIF on the flagged item between focal and reference groups was estimated using the weighted average of MGCFA-ESs extracted from the individual studies, which can be computed as

$$\text{MGCFA}-\text{ES}_{\bullet} = \frac{\sum_{i=1}^{k} W_i[\text{MGCFA-ES}_i]}{\sum_{i=1}^{k} W_i}, \tag{4}$$

and

$$V_{\text{MGCFA}-\text{ES}_{\bullet}} = \frac{1}{\sum_{i=1}^{k} W_i}, \tag{5}$$

where k is the number of studies included in the meta-analysis and W_i is the inverse of the associated estimated variance of MGCFA-ES$_i$. Below, MGCFA-ES$_.$ is a meta-analytic estimator of all MGCFA effect sizes from individual studies (MGCFA-ES$_i$).

Evaluation of Meta-Analytic Estimator of MGCFA-ES

The performance of MGCFA-ES as the DIF index was evaluated using bias and *MSE* values, which are given by

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta, \tag{6}$$

and

$$MSE(\hat{\theta}) = [Bias(\hat{\theta})]^2 + \text{var}(\hat{\theta}), \tag{7}$$

where $\hat{\theta}$ is MGCFA-ES$_.$ across all replications for each condition and θ is the preset population value of DIF magnitude. Mean bias values of MGCFA-ES$_.$ less than $|\pm 0.05|$ were considered to be within an acceptable range (Hoogland & Boomsma, 1998). In addition, the coverage rate of 95% confidence intervals, empirical standard errors of the $\hat{\theta}$ ($\sqrt{var(\hat{\theta})}$), and empirical rejection rates of MGCFA-ES$_.$ were computed. In order to control for the overall type I error rate, Bonferroni's adjusted alpha level of .0083 was used (Kim & Oshima, 2012).

Results

Performance of Meta-Analytic Estimator of MGCFA-ES

The overall performances of MGCFA-ES$_{.}$ are summarized below in terms of (1) empirical Type I error rates and statistical power, (2) bias and *MSE* values, and (3) coverage rate of 95% confidence intervals and empirical standard errors.

Type I Error Rates and Statistical Power. Under the condition that DIF does not occur, the percentage rates of incorrectly rejecting the null hypothesis were 0.9%, which were all slightly above the preset nominal Type I error rate of .0083, regardless of correlation type. In addition, under the condition that DIF is set to occur, the mean percentages of correctly rejecting the null hypothesis were all equal to 100%. This result indicates that MGCFA-ES$_{.}$, as the DIF index, has sufficient statistical power for correctly detecting DIF on the biased item, regardless of the correlation type.

Bias and MSE values. Figure 2 depicts bias and *MSE* values of MGCFA-ES$_{.}$ by correlation type. When a tetrachoric correlation was used, mean bias and *MSE* values of MGCFA-ES$_{.}$ were found to be the smallest (i.e., less than |.05|), indicating that MGCFA-ES$_{.}$ extracted from a tetrachoric correlation yielded the most accurate estimate of the population magnitude of DIF. However, when a Pearson correlation was used, bias values of MGCFA-ES$_{.}$ were higher than |.05|. Similarly, the *MSE* value of MGCFA-ES$_{.}$ was the smallest when MGCFA was fitted to tetrachoric correlations, followed by an adjusted Pearson correlation. Regardless of correlation type, mean bias, and *MSE* values of MGCFA-ES$_{.}$ were the largest when a large DIF appeared on the flagged item, followed by medium and small DIFs.

Coverages of 95% Confidence Intervals and Empirical Standard Error. Figure 2 shows coverage of 95% confidence intervals and empirical standard error for MGCFA-ES$_{.}$ by correlation type and DIF magnitude. Regardless of DIF magnitude, coverage of 95% confidence intervals around MGCFA-ES$_{.}$ was the largest when a tetrachoric correlation was used, while it was lowest for the Pearson correlation. The empirical standard errors of MGCFA-ES$_{.}$ were all below .05 and were almost identical, though slightly less for the tetrachoric correlation.

Factors Affecting Meta-Analytic Estimator of MGCFA-ES

The differential effects of various factors on MGCFA-ES$_{.}$ are summarized below in terms of (1) bias and *MSE* values and (2) coverage rate of 95% confidence intervals and empirical standard errors.

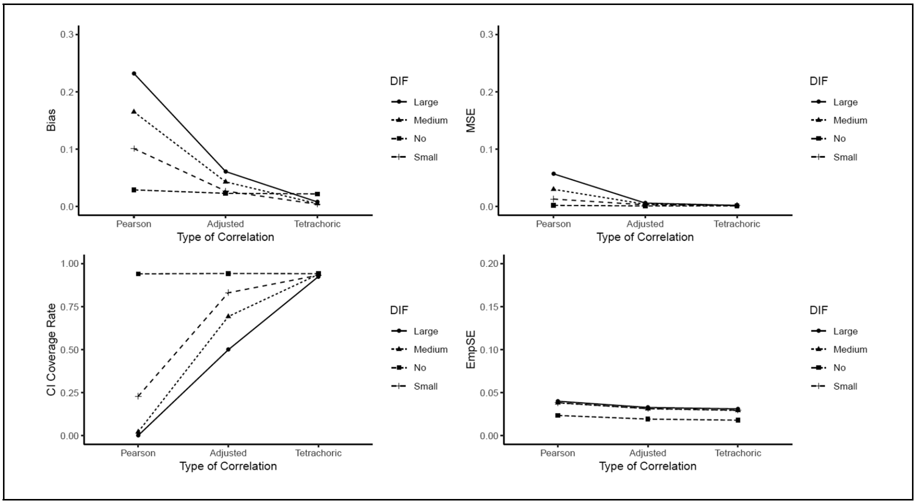


Figure 2. The bias, *MSE*, confidence interval (CI) coverage rate, and empirical standard error (EmpSE) of the meta-analytic estimator of MGCFA-ES, when different correlation matrices were used to fit MGCFA.

Note. DIF = differential item functioning; Pearson = Pearson correlation matrices were used; Adjusted = Pearson correlation matrices × 3/2 were used; tetrachoric = Tetrachoric correlation matrices were used; *MSE* = mean square error; MGCFA-ES = multigroup confirmatory factor analysis effect sizes.

Bias and MSE Values. As shown in Figure 3, the mean bias values of MGCFA-ES$_{.}$ were found to be greatest for identifying an item showing a large difference in item response between focal and reference groups, followed by medium, small, and no DIF items. An exception was found when a tetrachoric correlation was used, showing the greatest mean bias values with no DIF item. As shown in Figure 4, the *MSE* values of MGCFA-ES$_{.}$ were found to be greatest for identifying an item showing large DIF in responses between focal and reference groups, followed by medium, small, and no DIF items.

Coverages of 95% Confidence Intervals and Empirical Standard Error. As shown in Figure 5, coverage rates of MGCFA-ES$_{.}$ were found to be greatest for identifying an item showing no difference in item responses, followed by small, medium, and large DIF items for the Pearson and adjusted Pearson correlations. For a tetrachoric correlation, the coverage rate was the largest for no DIF item, followed by large, medium, and small DIFs. Such a pattern was consistent across all levels of item difficulty. One exception was when coverage rate of MGCFA-ES$_{.}$ was the largest for an item with small DIF when the sample size was set to be small. As shown in Figure 6, the empirical standard error of MGCFA-ES$_{.}$ was found to be largest for identifying a large DIF item, followed by small, medium, and no DIF items. The pattern was consistent regardless of the types of correlation matrices.

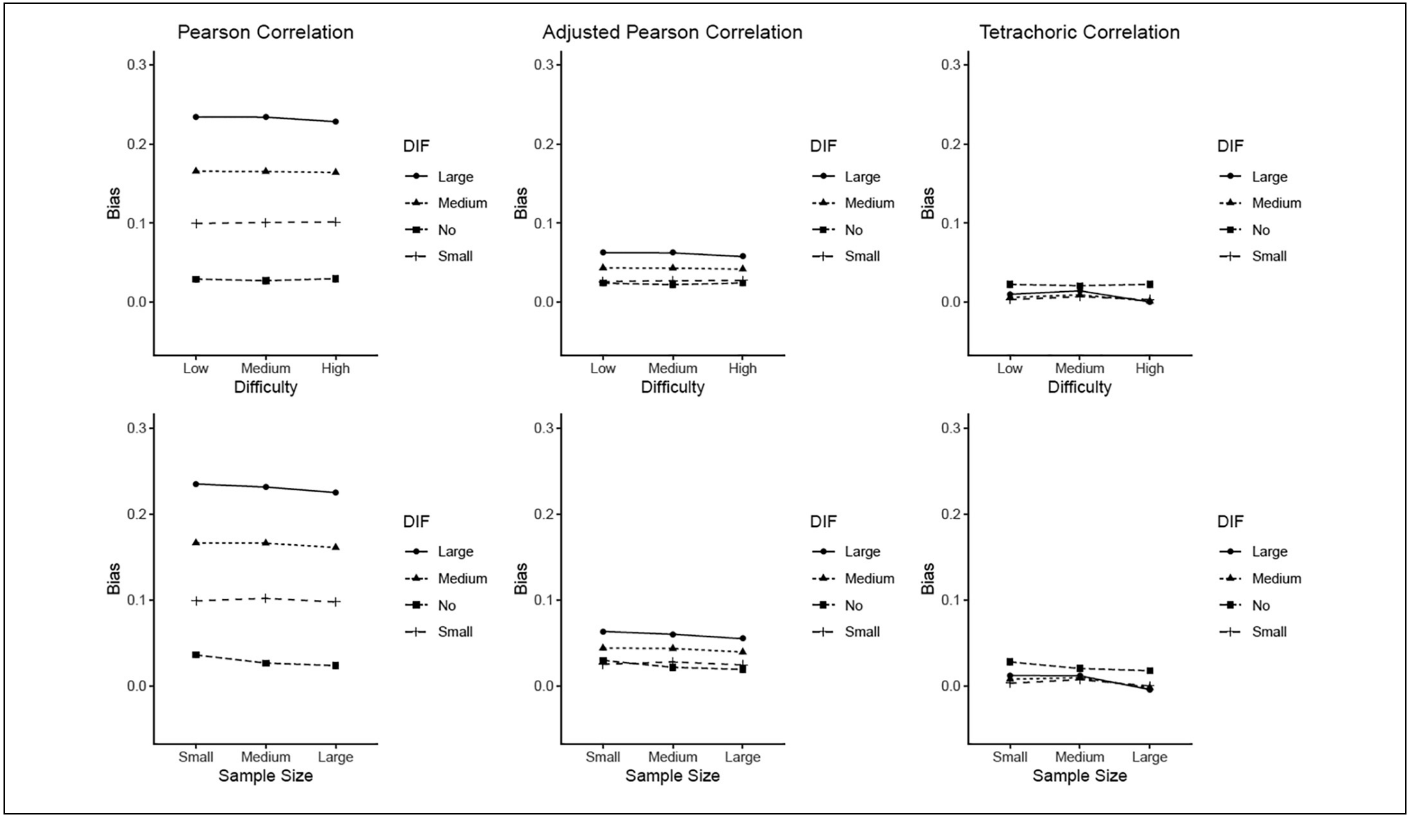


Figure 3. The bias of meta-analytic estimator of MGCFA-ES by difficulty and DIF, and sample size and DIF.

Note. DIF = differential item functioning condition; MGCFA-ES = multigroup confirmatory factor analysis effect sizes.

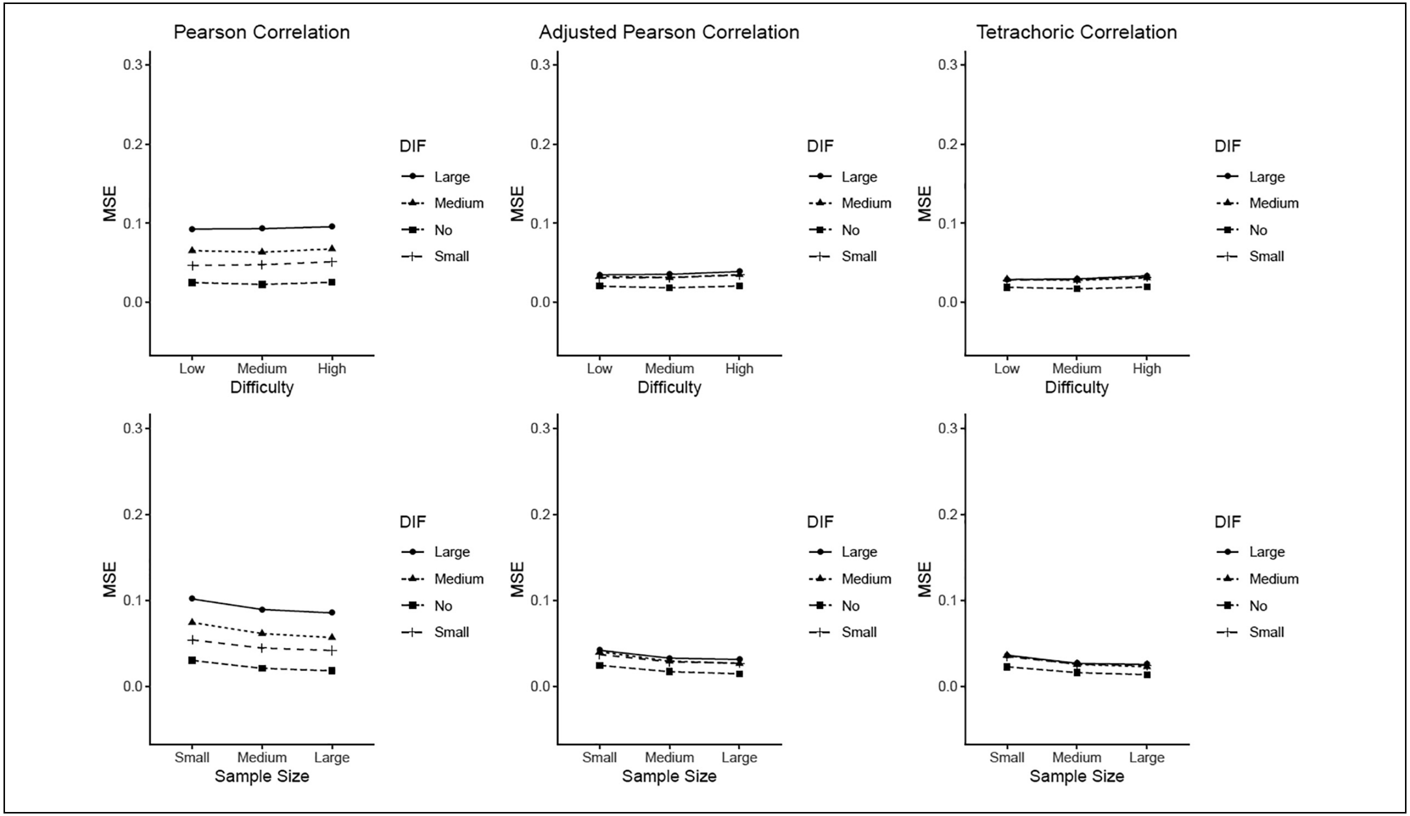


Figure 4. The *MSE* of meta-analytic estimator of MGCFA-ES by difficulty and DIF, and sample size and DIF.

Note. DIF = differential item functioning condition; *MSE* = mean square error; MGCFA-ES = multigroup confirmatory factor analysis effect sizes.

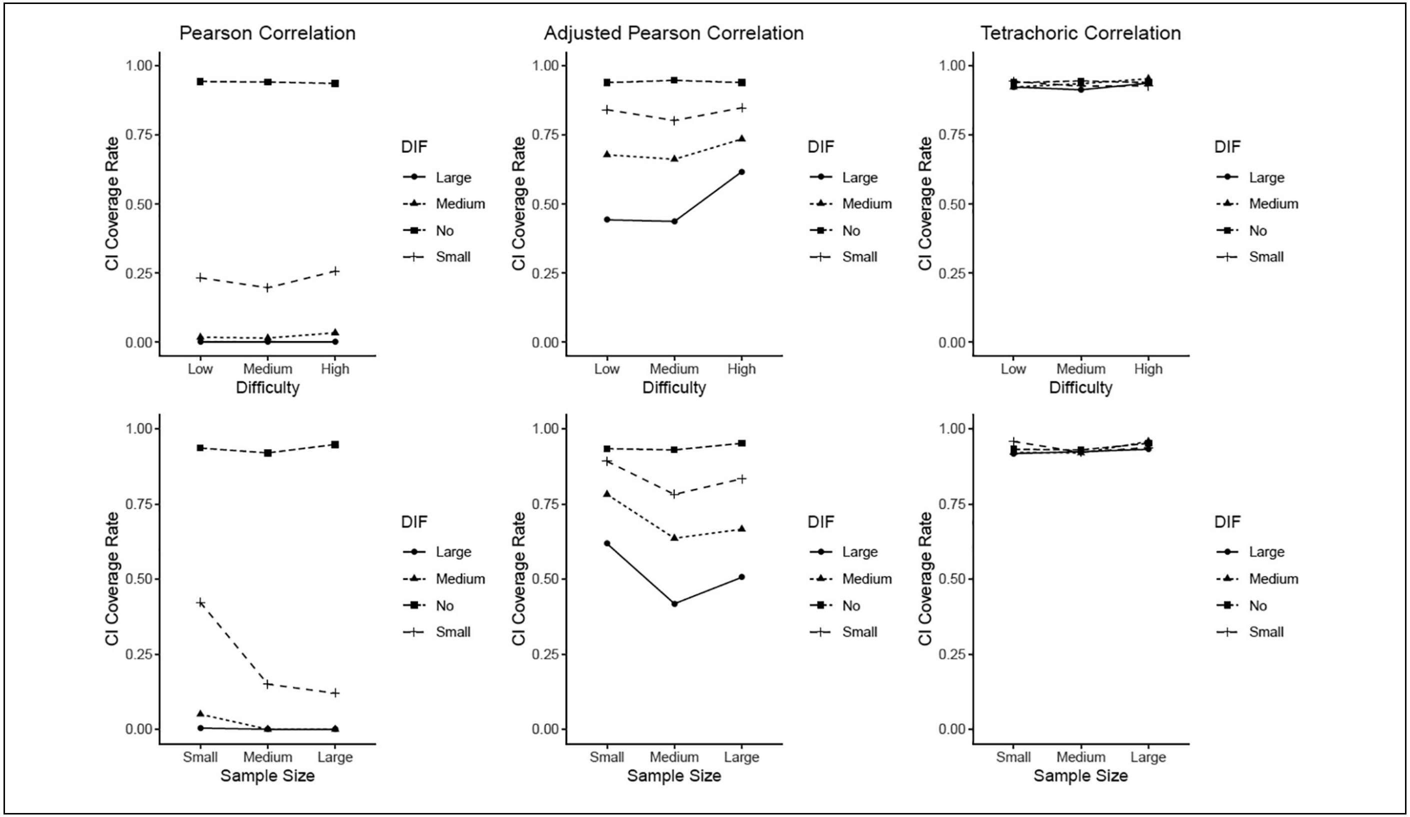


Figure 5. Confidence interval (CI) coverage rate of the meta-analytic estimator of MGCFA-ES by difficulty and DIF, and sample size and DIF.

Note. DIF = differential item functioning condition; MGCFA-ES = multigroup confirmatory factor analysis effect sizes.

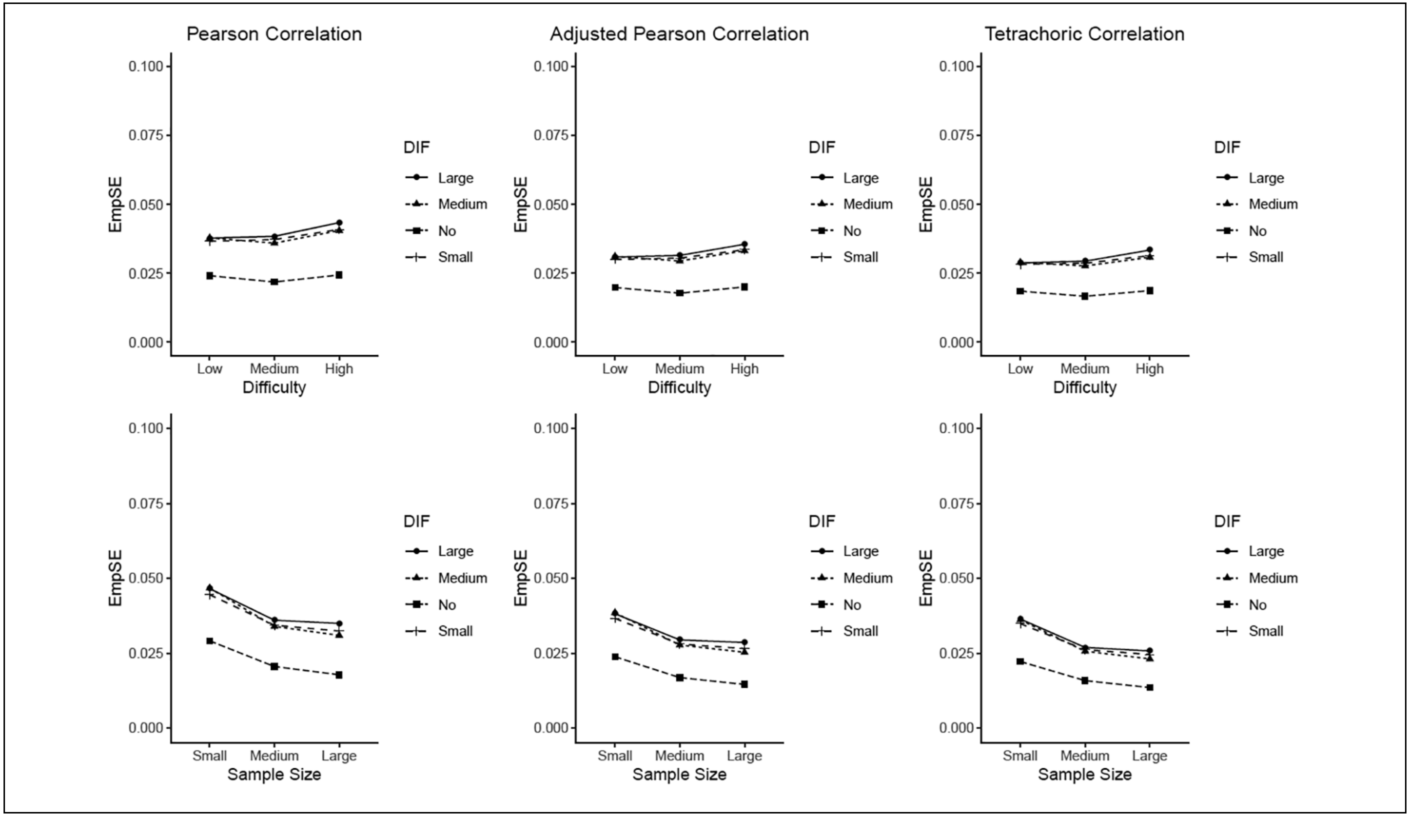
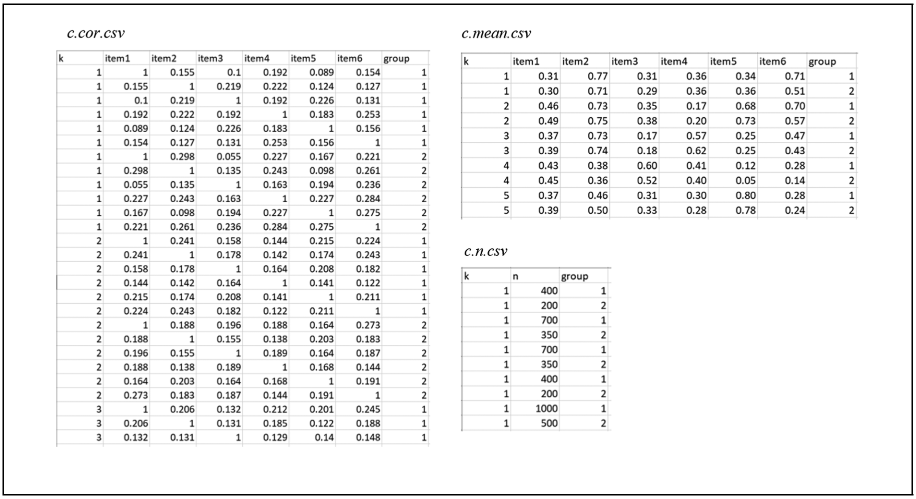


Figure 6. Empirical standard error (EmpSE) of the meta-analytic estimator of MGCFA-ES by difficulty and DIF, and sample size and DIF.

Note. DIF = differential item functioning condition; MGCFA-ES = multigroup confirmatory factor analysis effect sizes.



c.cor.csv

k	item1	item2	item3	item4	item5	item6	group
1	1	0.155	0.1	0.192	0.089	0.154	1
1	0.155	1	0.219	0.222	0.124	0.127	1
1	0.1	0.219	1	0.192	0.226	0.131	1
1	0.192	0.222	0.192	1	0.183	0.253	1
1	0.089	0.124	0.226	0.183	1	0.156	1
1	0.154	0.127	0.131	0.253	0.156	1	1
1	1	0.298	0.055	0.227	0.167	0.221	2
1	0.298	1	0.135	0.243	0.098	0.261	2
1	0.055	0.135	1	0.163	0.194	0.236	2
1	0.227	0.243	0.163	1	0.227	0.284	2
1	0.167	0.098	0.194	0.227	1	0.275	2
1	0.221	0.261	0.236	0.284	0.275	1	2
2	1	0.241	0.158	0.144	0.215	0.224	1
2	0.241	1	0.178	0.142	0.174	0.243	1
2	0.158	0.178	1	0.164	0.208	0.182	1
2	0.144	0.142	0.164	1	0.141	0.122	1
2	0.215	0.174	0.208	0.141	1	0.211	1
2	0.224	0.243	0.182	0.122	0.211	1	1
2	1	0.188	0.196	0.188	0.164	0.273	2
2	0.188	1	0.155	0.138	0.203	0.183	2
2	0.196	0.155	1	0.189	0.164	0.187	2
2	0.188	0.138	0.189	1	0.168	0.144	2
2	0.164	0.203	0.164	0.168	1	0.191	2
2	0.273	0.183	0.187	0.144	0.191	1	2
3	1	0.206	0.132	0.212	0.201	0.245	1
3	0.206	1	0.131	0.185	0.122	0.188	1
3	0.132	0.131	1	0.129	0.14	0.148	1

c.mean.csv

k	item1	item2	item3	item4	item5	item6	group
1	0.31	0.77	0.31	0.36	0.34	0.71	1
1	0.30	0.71	0.29	0.36	0.36	0.51	2
2	0.46	0.73	0.35	0.17	0.68	0.70	1
2	0.49	0.75	0.38	0.20	0.73	0.57	2
3	0.37	0.73	0.17	0.57	0.25	0.47	1
3	0.39	0.74	0.18	0.62	0.25	0.43	2
4	0.43	0.38	0.60	0.41	0.12	0.28	1
4	0.45	0.36	0.52	0.40	0.05	0.14	2
5	0.37	0.46	0.31	0.30	0.80	0.28	1
5	0.39	0.50	0.33	0.28	0.78	0.24	2

c.n.csv

k	n	group
1	400	1
1	200	2
1	700	1
1	350	2
1	700	1
1	350	2
1	400	1
1	200	2
1	1000	1
1	500	2

Figure 7. The data set for correlation matrices, the mean of correct answers on each item, and sample sizes.

Application to Empirical Data Using R

The current section describes how the proposed method can be utilized in practice. In this section, we assume five MGCFA-ESs between p subgroups (i.e., $p = 2$ in this demonstration) extracted from k independent studies ($k = 5$ in this demonstration) that provide (1) Pearson product moment correlation matrices and (2) the proportion of correct answers on each item. Below, the part of R codes (presented in italics) that are necessary for each step are presented.

Step 1: Creating a Data Set. As shown in Figure 7, three data sets are created: (1) the five sets of two correlation matrices among items stacked by rows (saved as "*c.cor*"), (2) the five sets of two proportions of correct answers on each item stacked by rows (saved as "*c.mean*"), and (3) the five sets of two separate sample sizes stacked by rows (saved as "*c.n*"). As shown in R code below, the five sets of two separate correlation matrices among items stacked by rows are corrected by multiplying by 3/2 in note 1. The five sets of two proportions of correct answers on each item are converted into a threshold value as shown in notes 2 and 3.

```
for(i in 1:k){
    c.cor.r=as.matrix(cor.r[(1+6*(i-1)):(6*i), 2:7])
    c.cor.r1<-(3/2)*(c.cor.r) # adjust Pearson correlation matrices (note #1).
    diag(c.cor.r1)<-1
    c.cor.f=as.matrix(cor.f[(1+6*(i-1)):(6*i), 2:7])
    c.cor.f1<-(3/2)*(c.cor.f)
    diag(c.cor.f1)<-1
    c.mean.r=as.matrix(mean.r[i,2:7])
    c.t.r<-c() # convert the proportion of correct answers on each item to its corresponding threshold for reference group (note #2)
    for(ii in 1:6){
    c.t.r[ii]<-qnorm(1-c.mean.r[, ii])
    }
    c.mean.f=as.matrix(mean.f[i,2:7])
    c.t.f<-c()# convert to threshold for focal (note #3)
    for(ii in 1:6){
    c.t.f[ii]<-qnorm(1-c.mean.f[, ii])
    }
    c.n.r=as.matrix(n.r[i, 2])
    c.n.f=as.matrix(n.f[i, 2])
```

Step 2: Fitting MGCFA to the Data Set. Using the *cfa (“lavaan”)* available in the R Version 3.5.3 (R Core Team, 2019), the MGCFA model is specified to fit the data extracted from each study. In particular, thresholds for all items ($x1$ to $x5$) except a flagged item ($x6$) are fixed to be the same for the two groups, and factor scores ($f1$) are fixed to have means of zero and variances of 1 for the reference group while they were estimated freely for the focal group as shown in notes 4 to 7. Then, MGCFA-ES (DIF = $b7-b6$) and its associated standard error are estimated as shown in note 8. The specified model (*model*) is fitted using three data sets (*c.cor, c.mean*, and *c.n*) in the *cfa (“*lavaan*”)* function as shown in note 9.

```
  model <-’
    group: 1
    f1 =~ l*x1 + l*x2 +l*x3 + l*x4 + l*x5 + l*x6
    x1 ~~ residual*x1
    x2 ~~ residual*x2
    x3 ~~ residual*x3
    x4 ~~ residual*x4
    x5 ~~ residual*x5
    x6 ~~ residual*x6
    x1 ~ i1*1
    x2 ~ i2*1
    x3 ~ i3*1
    x4 ~ i4*1
    x5 ~ i5*1
    x6 ~ i6*1
    f1 ~~ 1*f1 # factor scores (f1) have a variance of 1 (note #4)
    f1~0*f1 # factor scores (f1) has a mean of 0 (note #5)
    group: 2
    f1 =~ l*x1 + l*x2 + l*x3 + l*x4 + l*x5 + l*x6
```

```
x1 ~~ residual*x1
x2 ~~ residual*x2
x3 ~~ residual*x3
x4 ~~ residual*x4
x5 ~~ residual*x5
x6 ~~ residual*x6
x1 ~ i1*1
x2 ~ i2*1
x3 ~ i3*1
x4 ~ i4*1
x5 ~ i5*1
x6 ~ i7*1
f1 ~~ f1 # factor scores are estimated freely (note #6)
f1~f1 # factor scores are estimated freely (note #7)
a := 1.7*(l)/sqrt(residual)
b1 := (i1)/(l)
b2 := (i2)/(l) b3 := (i3)/(l)
b4 := (i4)/(l)
b5 := (i5)/(l)
b6 := (i6)/(l)
b7 := (i7)/(l)
DIF= b7-b6 ' # define DIF (note #8)

Require(lavaan) # load library called lavaan.

fit.cor = cfa(model, sample.cov = c.cor, sample.mean=c.mean, sample.nobs =
c.n, std.lv=TRUE) # run multiple group CFA using datasets (note #9)
```

Step 3: Meta-Analysis of MGCFA-ES. The MGCFA-ESs and their associated standard errors for each study as estimated in Step 2 are combined using the *rma*(''metafor'') available in the R Version 3.5.3 (R Core Team, 2019).

```
require(metafor)
RE_cor=rma(yi=DIF_cor$dif, sei=DIF_cor$se, data=DIF_cor,
measure="GEN", method="REML")
Summary(RE_cor)
```

Conclusion and Discussion

The current study (1) proposes a new meta-analytic index (MGCFA-ES) for synthesizing the magnitude of DIF across studies, when an MGCFA was used and (2) evaluates the performance of MGCFA-ES using a Monte Carlo Simulation technique, where a number of factors were manipulated. The meta-analytic summary of

MGCFA-ES was found to perform reasonably well in terms of the bias and *MSE* values, empirical Type I error rates, statistical power, coverage rates, and empirical standard error. Of the three types of correlations used to fit the MGCFA, tetrachoric correlation provided the most accurate and unbiased estimate, followed by the adjusted Pearson product moment correlation. In addition, the proposed meta-analytic approach for synthesizing DIF across studies was found to yield the most accurate estimator when a tetrachoric correlation was used to find the item, showing a large DIF. It was also found that the meta-analytic summary of MGCFA-ES was most efficient when a tetrachoric correlation based on a large sample size was used.

The current simulation study demonstrated that MGCFA-ES could be used to summarize the magnitude of DIF across studies included in meta-analyses, when studies provide correlation matrices and item thresholds summarizing responses for the known subgroups. The use of tetrachoric correlations yields the most accurate and efficient estimators of DIF when fitting MGCFA for measurement invariance. However, when Pearson correlation matrices were reported from the studies, adjusting them by multiplying by 3/2, as suggested by Fillmore et al. (1998), produced more accurate and unbiased results. As shown in the simulation results, the meta-analytic summary of MGCFA-ES extracted from the adjusted Pearson correlation performed almost the same as the meta-analytic summary of MGCFA-ES when tetrachoric correlations were fitted. Although the current study might be limited, as not all features of meta-analysis (i.e., the number of studies included was fixed to 30 and the fixed effect model was only investigated) were manipulated, we believe that our study clearly shows MGCFA-ES as a viable option for summarizing the magnitude of DIF in meta-analyses.

It is our belief that with more published studies using SEM for identifying items showing DIF in the future, MGCFA-ES shows great promise for wider use. We strongly recommend that researchers provide separate correlation matrices among items and the proportion of correct answers on each item. Further research is necessary to expand the current study so that the performance of MGCFA-ES can be evaluated under conditions that mirror different meta-analytic settings. For example, given that SEM studies are mostly based on larger sample sizes, we used a minimum sample size of 200 in our simulation. Yet the performance of the MGCFA-ES needs to be assessed when SEM studies with smaller sample sizes or more items are included. In addition, future studies should explore meta-analytic methods that can be used to detect a nonuniform DIF or DIF on polytomous items. Last, an empirical study that examines the comparability of different DIF effect size indicators is necessary to provide practical guidelines to meta-analysts.

Declaration of Conflicting Interests

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

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Sung Eun Park https://orcid.org/0000-0003-4227-6446
Cengiz Zopluoglu https://orcid.org/0000-0002-9397-0262

References

Bauer, D. J. (2017). A more general model for testing measurement invariance and differential item functioning. *Psychological Methods*, *22*(3), 507-526. https://doi.org/10.1037/met0000077

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Sage.

Chang, H., Mazzeo, J., & Roussos, L. (1995). *Detecting DIF for polytomously scored items: An adaptation of the SIBTEST procedure* (Research Report). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1995.tb01640.x

Educational Testing Service. (2014). *ETS standards for quality and fairness*. https://www.ets.org/s/about/pdf/standards.pdf

Fillmore, K. M., Golding, J. M., Graves, K. L., Kniep, S., Leino, E. V., Romelsjo, A., Shoemaker, C., Ager, C. R., Allebeck, P., & Ferrer, H. P. (1998). Alcohol consumption and mortality. I. Chracteristics of drinking of groups. *Addictions*, *93*(2), 183-203. https://doi.org/10.1046/j.1360-0443.1998.9321834.x

Gómez-Benito, J., Hidalgo, M. D., & Padilla, J. (2009). Efficacy of effect size measures in logistic regression: An application for detecting DIF. *Methodology*, *5*(1), 18-25. https://doi.org/10.1027/1614-2241.5.1.18

Hidalgo, M. D., Gómez-Benito, J., & Zumbo, B. D. (2014). Binary logistic regression analysis for detecting differential item functioning: Effectiveness of R2 and delta log odds ratio effect size measures. *Educational Psychological Measurement*, *74*(6), 927-949. https://doi.org/10.1177/0013164414523618

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Lawrence Erlbaum.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, *26*(3), 329-367. https://doi.org/10.1177/0049124198026003003

Jin, Y., Myers, N. D., Ahn, S., & Penfield, R. D. (2012). A comparison of uniform DIF effect size estimators under the MIMIC and Rasch models. *Educational and Psychological Measurement*, *73*(2), 339-358. https://doi.org/10.1177/0013164412462705

Kim, J., & Oshima, T. C. (2012). Effect of multiple testing adjustment in differential item functioning detection. *Educational Psychological Measurement*, *73*(3), 458-470. https: //doi.org/10.1177/0013164412467033

Koo, J. (2012). *Meta-analysis of the Mantel-Haenszel index for the detection of differential item functioning* [Unpublished doctoral dissertation]. Florida State University, Tallahassee.

Koo, J., Becker, B. J., & Kim, Y. S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing*, *31*(1), 89-109. https://doi.org/10.1177/0265532213496097

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Routledge/Taylor & Francis. https://doi.org/10.4324/9780203821961

Muthén, B. O., Kao, C., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*(1), 1-22. https://doi.org/10.1111/j.1745-3984.1991.tb00340.x

Oshima, T. C., Wright, K., & White, N. (2015). Multiple-group noncompensatory differential item functioning in Raju's differential functioning of items and tests. *International Journal of Testing*, *15*(3), 254-273. https://doi.org/10.1080/15305058.2015.1009980

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.) [Quantitative Applications in the Social Sciences: *Vol. 161*]. Sage. https://doi.org/10.4135/9781412993913

Penfield, R. D. (2007). Assessing differential step functioning in polytomous items using a common odds ratio estimator. *Journal of Educational Measurement*, *44*(3), 187-210. https://doi.org/10.1111/j.1745-3984.2007.00034.x

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

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

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*(6), 1292-1306. https://doi.org/10.1037/0021-9010.91.6.1292

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: Examples using item response theory to analyze differential item functioning. *Psychological Methods*, *11*(4), 402-415. https://doi.org/10.1037/1082-989X.11.4.402

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, *43*(4), Article 599. https://doi.org/10.1007/s11135-007-9143-x

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

Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376. https://doi.org/10.1177/0734282911406666

Woods, C. M., & Grimm, K. J. (2011). Testing for nonuniform differential item functioning with multiple indicator multiple cause models. *Applied Psychological Measurement*, *35*(5), 339-361. https://doi.org/10.1177/0146621611405984

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. https://doi.org/10.3102/1076998611431085